Paper 244-25

Test Development: Ten Steps to a Valid and Reliable Certification Exam

Linda A. Althouse, Ph.D., SAS Institute Inc., Cary, NC

ABSTRACT

The intent of a certification program is to evaluate the knowledge and skills of practitioners seeking a credential to ensure a desired level of competence. For certification examinations to serve their intended goal, the examination must assess the performance that the scores are intended to represent. Subsequently, each item on the examination must serve to assess the performance or knowledge being measured by the examination. The purpose of this paper is to describe the steps necessary to develop a valid and reliable certification examination which would fulfill the goal of a certification program.

INTRODUCTION

The world of statistical programming, application development, data management, and providing business solutions is competitive and continues to grow in complexity as new technology emerges. As a result, the demand for qualified, knowledgeable professionals also increases. As the number of SAS users and consultants continues to grow, the need to distinguish between those who have mastered a specified level of competence in their use of SAS products and/or solutions versus those who have not becomes increasingly important.

While certification is typically a voluntary process, it is becoming more and more necessary for IT professionals to obtain certification to remain competitive in the job market. Employers perceive employees with certification as being more competent and productive. In addition, employees view certification as contributing to their professional credibility. (Network World, 1998)

Within the past ten years, an influx of IT certification programs have emerged. In 1999, SAS Institute launched a global certification program. In addition to providing users with recognition of mastery at a specified level, the certification program provides an incentive to users to continue seeking the proper training, expand their skills, and strengthen their competitive edge in the job market.

With the growing number of certification organizations, much attention has been placed on the essential components of global certification programs and the standards established to govern them. Leading publications which provide guidelines to assist organizations in ensuring their examinations meet the standards necessary for a valid, reliable, and legally defensible examination include, but are not limited to, the following:

1. *Joint Technical Standards for Educational and Psychological Tests* (1999) jointly developed by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education.
2. *Standards for Accreditation of Certifying Agencies* (1995) developed by the National Commission for Certifying Agencies.
3. *Principles of Fairness: An Examining Guide for Certification Boards* (1993) jointly developed by the Council on Licensure, Enforcement, and Regulation and the National Organization for Competency Assurance.
4. *Uniform Guidelines on Employee Selection Procedures* (1978) developed by the Equal Opportunity Commission.

THE TEST DEVELOPMENT PROCESS

The development of a certification examination is a lengthy and involved process. The process needs to be followed to ensure that the examination is valid and reliable. Validity is the ability of the test to measure that which it is intended to measure. For an examination to have content validity, it must demonstrate at least two qualities. First, the content of the examination must be job-related. Second, the examination should cover areas where lack of knowledge would result in inability to perform the job.

Another critical element of the quality of a certification examination concerns how reliable, or consistent, the examination is in measuring candidates' ability levels. Reliability is the index of how accurately the examination measures the candidate's skills and is a necessary condition to achieve exam validity.

An examination must be both valid and reliable to be considered a well-developed and defensible examination. By following the rigid standards of the test development process, these two qualities are likely to be met. The test development process can be summarized into ten steps with each step instrumental in ensuring the validity, reliability, defensibility, and security of the examination:

1. Conducting the Job Task Analysis
2. Developing the Test Blueprint
3. Developing Items
4. Reviewing and Validating Items
5. Assembling and Delivering Beta Exams
6. Analyzing Beta Exam Results
7. Constructing Equivalent Exam Forms
8. Establishing the Passing Score
9. Administering/Scoring Operational Exams
10. Providing Ongoing Test Maintenance

STEP 1: CONDUCTING THE JOB TASK ANALYSIS

The content of the examination should be related to the job or role the individual is seeking to practice. The most widely used and accepted way of establishing job-relatedness is to conduct a job task analysis. This step of the test development process is the most complex and lengthy step. However, this step serves as the foundation for the examination. As with a home, a solid foundation will result in a solid structure.

The job task analysis is a systematic method of collecting data regarding the responsibilities, knowledge, and skills associated with acceptable performance within a profession. These data are then used to develop the blueprint for the examination. The job task analysis typically consists of two phases (Henderson, 1996):

1. Obtaining and describing the job information
2. Validating the job description

Phase 1 - Obtaining and describing the Job Information:
Prior to obtaining and describing the job information, the target audience for each certification examination needs to be defined, documented, and provided to the participants of the job analysis as reference material. Multiple methods and information-gathering approaches can and should be used in developing the job description. Training material, handbooks, product

specifications, and work-related procedure manuals can be used. In cases where the jobs or duties are changing so rapidly, individuals recognized as experts in the field (e.g. trainers) become good sources. Another common method is to consult advanced individuals currently performing or who have recently performed the role for which the certification is intended. Use of these subject matter experts (SMEs) increases the validity of the certification examination as data are obtained directly from individuals who are the most knowledgeable about a job. (Flaherty and Hogan, 1998)

While much information can be gathered through the review of literature, the validity of the job analysis may be greatly enhanced when information can be obtained directly from those who are most knowledgeable about the job for which the certification is intended. As a result, the selection and involvement of SMEs is a critical component in completing a thorough job analysis.

Selecting Subject Matter Experts: It is important to ensure that the group of SMEs selected is representative of the population for which the exam is intended. For example, having only novices participate may result in an incomplete job analysis as some necessary tasks may be omitted. Having only experts with many years of experience may result in a listing of tasks that are not reasonable for a minimally qualified candidate. As a result, the job analysis should not only include those who are experts and highly recognized within the profession, but it should include "entry-level" professionals to ensure that the tasks identified reflect work situations which are commonly encountered by those achieving the certification (CLEAR, 1998).

In addition to skill level, other factors to consider in forming a representative group are:

- Geographic region
- Ethnicity
- Gender
- Age
- Education Level
- Work Setting (e.g., university, corporation)
- Years of experience
- Specialty area

When establishing a committee of SMEs to participate in describing the job tasks, it is desirable to include SMEs who will also be involved in other phases of test development as some overlap between the different stages helps ensure a smoother transition throughout the process. The number of SMEs can range from 8-15. However, this number is only a guideline and the true number will depend heavily on the certification.

Organization of the Job Description: Typically, the description begins as an objective catalog of tasks that people do on the job. These descriptions usually take the form of a list of tasks. Once this exhaustive list is developed, the list is reviewed to ensure that each task listing is independent of the other tasks, the task list is comprehensive, and that no tasks have been overlooked. The tasks are categorized into major domains, categories, or technology area. These domains provide the organizational framework for the exam and can serve as the framework for a training program.

After the domains are defined, the list of tasks is categorized into the domains. A task should be placed in only one domain. To develop the test objectives, the job tasks are translated into specific, measurable knowledge, skills, and/or abilities. The list of objectives is then used as the framework for the certification examination and curriculum material.

Phase 2 - Validating the Job Description: Once the full listing of objectives is determined, it is necessary to have this listing validated with a representative sample of individuals currently performing the job for which the certification is intended. The validation effort is typically involves surveying a random sample of job incumbents. By using a survey design, the amount of empirical evidence to support the content validity of the examination is increased. (Henderson, 1996)

Where job incumbents do not exist, as with rapidly changing IT professions, a random sample of individuals familiar with the job being certified is selected. This group should include all development areas, including but not limited to, research design, development, quality assurance, and training.

Developing the Survey Instrument: Proper development of the survey is critical to the collection of useful information for examination development. Some of the main issues that must be considered include providing clear, meaningful directions to the respondents, collecting demographic information, and developing the questions and rating scales. (Flaherty and Hogan, 1998)

At a minimum, the directions provided to the survey respondents should include the statement of purpose, how the data will be used, projected amount of time needed, definitions of any scales used, the date for returning the survey, information on returning the survey, and a contact person or email address in the event that respondents have questions.

The demographic questions included on the survey should cover the same background information (e.g., geographic region, ethnicity, gender, age, education level, work setting, years of experience, specialty area) collected when selecting the SMEs in Phase 1 of the job analysis. The survey should also ask the amount of time respondents spend in their jobs performing the tasks associated with the certification. If the amount of time performing the specified job function is not adequate, then that respondent's data should not be included. The certifying organization may also ask questions regarding willingness to assist in future test development efforts.

The questions on the survey are typically in the form of a Likert rating scale with each item corresponding to a domain or objective. A variety of rating scales can be used and depends on the type of certification. For example, if public protection were the main purpose of the certification program then the consequence of error would be asked for each objective. If the development of the training curriculum is the main purpose, then the frequency with which a respondent performs an objective may be asked. If distinguishing between knowledgeable and unknowledgeable candidates for purposes of certification is the goal, then the importance of each objective may be asked. A combination of ratings such as criticality, frequency, and importance can also be used. In addition to the ratings for the objectives, respondents should be given the opportunity to identify the relative importance of each domain.

SAS Institute's certification exams currently use the importance rating scale as the key component in determining the weightings on the examination. A four point Likert scale is used with 1 indicating that knowledge of this task is not essential to the job performance of a certified professional and 4 indicating that knowledge of this objective is essential.

STEP 2: DEVELOPING THE TEST BLUEPRINT

The purpose of the test blueprint is to define the attributes of the examination. This blueprint is then used to ensure that the assembled test forms are consistent from form to form in content. That is, if one candidate receives Form A and another receives Form B, they will be taking equivalent exams. Statistical analysis discussed later in this paper will explain how the exams will be

made statistically equivalent as well. As a minimum, the test blueprint should include:

- Purpose of the exam
- Description of the target audience
- Total number of items on the exam
- Number of items per domain/objective
- Content outline
- Exam format and item types

The first three items in the above list are usually determined in Phase 1 of the job analysis or by the certifying organization. How the items are dispersed across the exam is determined using the empirical data from the job analysis survey. With these data, the objectives can be prioritized and weighted. For example, objectives receiving higher importance ratings will have more items allocated to them than objectives receiving lower importance ratings. Once the percentage of items has been determined, the test content outline can be finalized.

The type of exam and items to be used must also be determined at this point. Assessment of content knowledge and readiness for acceptable job performance can be conducted in multiple ways. The most common method is to use a multiple-choice examination. Performance assessments that involve the actual demonstration of job-related behavior, simulations, case studies, and short answer questions are additional methods. Each method has advantages and disadvantages (Dungan, 1996; Haladyna, 1999). An analysis of the tasks would need to be conducted to determine the best method for assessing knowledge of that task. The tradeoffs for each type of examination and budgetary constraints would also be considered in selecting the assessment type. For SAS Institute's certification examinations, the decision was made to use the multiple-choice examination. Guidelines and principles for developing multiple-choice items are well researched and contained in a number of sources (Osterlind, 1997; and Haladyna, 1999).

STEP 3: DEVELOPING ITEMS

Once the test blueprint is finalized, a pool of items is developed to measure each of the objectives. Each item is linked through a classification system to the test blueprint. The number of items and the type of items that are needed will have been determined in the test blueprint stage. To ensure that enough items survive the item review and beta test process, at least three times as many items need to be written. This section will focus on the multiple-choice item. However, many of the guidelines can be directly applied to other item types, such as short answer or essays.

SMEs should write the exam items. Therefore, the first step in item development is to assemble a group of SMEs to develop items.

Selecting Item Writers: Items developed for use on the SAS certification examinations must meet rigorous requirements designed to ensure fairness and exam validity to all candidates. Therefore, SMEs are used to generate, review, and validate all exam items prior to their use on the beta examination. As with the job task analysis, the SMEs should represent a diverse demographic group similar to the population being tested. The SMEs must possess both content knowledge as well as a clear understanding of the job for which the examination is certifying. Ideally, the item writers would consist of, but would not be limited to, trainers, consultants, quality assurance testers, developers, designers, quality partners, and recognized professionals within the SAS® software user community, both internally and externally. Item writers should minimally be certified at the previous, equivalent examination level. Item writers should be required to sign an agreement protecting the confidentiality of the items they submit. In addition, the agreement would establish the copyright ownership of the items as those of the certifying organization.

Item Writing Training: All involved SMEs must complete item writing training prior to writing items to ensure familiarity with psychometric processes and with the exam requirements. While there are many methods of accomplishing item writing, the most intensive training and writing method involves bringing the selected group of SMEs to an item development workshop lasting a minimum of two days (Dungan, 1996). During this workshop, item writers receive formal training in item writing and generate items in small groups.

When developing items, writers must ensure that the items generated are:

- Significant (i.e., important to measure)
- discriminate between knowledgeable and unknowledgeable candidates
- match the intended objective
- do not provide any unintentional source of difficulty or answer cues.

Item writers should also be aware of the cognitive level of the items they are writing. A well-known approach to classifying objectives by cognitive levels is the Bloom taxonomy (Bloom, Engelhart, Furst, Hill & Kratwohl, 1956). Bloom's taxonomy consists of six levels:

- Knowledge - identify, state, recall, define, list, specify
- Comprehension - distinguish, provide examples
- Application - calculate, apply, solve
- Analyze - compare and contrast, detect errors
- Synthesis - design, formulate, integrate
- Evaluation - assess, decide, appraise

While there is debate over the usefulness of Bloom's taxonomy in item writing, this taxonomy has served as a foundation for many modified cognitive classification schemes (Dungan, 1996). Item writers should minimize the use of simple knowledge-level items and strive to develop items that measure higher levels of cognitive understanding, as these items will better discriminate between knowledgeable and unknowledgeable candidates.

Initial Item Review: Once the first draft of the item is written, the SMEs should review the item and validate its importance/appropriateness and objective match. Items should also be reviewed to ensure that they are technically accurate, not misleading or tricky, unbiased toward any population subgroup or culture, and are clearly worded. Editorial changes can also be made at this point.

STEP 4: REVIEWING AND VALIDATING ITEMS

Once the items have passed the initial review, a psychometric/editing team should review the items to ensure they meet standard, accepted psychometric properties. Items should also be reviewed to ensure they meet any specific standards of the certifying organization. Minor edits can be made to the items at this point. Once this review is completed, the item should be in its "final" form.

The "final" items are then reviewed by a set of SMEs. As with the item writers, the composition of this team is critical to the overall test validity. This team should include representation from qualified practitioners currently performing the job for which the certification is intended, as well as trainers, consultants, and others who work closely with performing the position in question. Signed confidential agreements should be used, as these

reviewers will be exposed to a significant number of test questions.

During this stage, each item is thoroughly reviewed for technical accuracy, relevance, and clarity. All responses are reviewed to ensure that the incorrect choices are plausible but unquestionably incorrect. The correct answer is reviewed to ensure that there is only one appropriate answer. Final consensus on all technical issues and whether this item belongs in the item pool is reached. Final approval of the item as it is to appear in its beta tested format is also reached.

The item writing and review process can be conducted in many different ways. For example, to save on cost, this SME review can be combined with the item writing process. After psychometric/editing changes are made, a small team of internal technical reviewers can verify the accuracy of the items prior to field-testing. Regardless of the method used, the items should be written, reviewed, and approved by SMEs, followed by a psychometric/editing review with a final check on the item accuracy. Once an item has been field/beta tested, it cannot be changed. As a result, it is critical to ensure that the items agreed upon at this step are complete and correct.

STEP 5: ASSEMBLING AND DELIVERING BETA EXAMS

Once reviewed, edited, and approved, items are placed in the item pool. The item pool or item bank is a depository of all items that are viable candidates for the examination. Beta examinations are conducted during a limited time period. The time period is dependent upon how many candidates are anticipated to test within a given period. For example, a shorter time period can be used if beta testing is held in conjunction with a popular candidate attended conference as these events typically result in a high number of participants.

A main goal of the beta exam is to field test the entire item bank. While this may require multiple beta exam forms, it provides the certifying organization as many items as possible to use when developing the operational forms. The disadvantage to multiple forms is that more beta candidates will be needed so that there is meaningful data on each item on each form. Ideally, at least 75-100 candidates will take each beta examination form.

Data collected from the beta exams allow the certifying organization to assess how each item performs and provides a chance for unforeseen problems to be resolved prior to the development of the operational examination.

The data also provides preliminary information critical for the development of pre-equated test forms that can be operationally scored immediately following the test administration. However, in order for the pre-equating to be meaningful, large representative candidate samples are needed to ensure the stability of the data. Without an adequate sample, some form of post-exam equating should be considered. A description of various equating techniques is provided by Kolen (1995).

Longer examinations are given at the beta level to account for the items that do not perform well, as these items will not be used to determine a candidate's score. However, to provide a reliable and valid exam score, candidates need to answer a sufficient number of items. By making the examinations longer, the items that survive for scoring provide a means for providing the candidate with their test score. The longer forms should be proportional to the operational examination blueprint.

Examinees taking a beta exam cannot receive their scores immediately. Instead they must wait until all analysis of the beta exams is completed, the items which will comprise their examination version are selected, and the passing score has been determined. The benefit for the candidate is that they have the opportunity to achieve certification earlier than if they wait for the production exam. In addition, beta examinations are typically offered at a reduced cost.

The beta examinations should be administered using the same method as the operational examination. The administration procedures, directions, security, and amount of time per item should match the operational examination.

STEP 6: ANALYZING BETA EXAM RESULTS

As mentioned earlier, the main purpose of beta testing is to field test items. Item data are reviewed to determine if the items performed as intended. As a minimum, the following item statistics are to be considered by the certifying organization:

- item difficulty
- item discrimination.

Item statistics should be reviewed from a psychometric perspective. Content experts should flag any potentially flawed items for additional review before using the item on the operational examination or in the scoring of the beta exam. In reviewing the items, the content experts should be provided with the number of candidates selecting each option and the mean score achieved by each group selecting each option, in addition to the above mentioned statistics.

Item Difficulty: The item difficulty (p-value) of an item is defined as the proportion of candidates who answer the item correctly. In general, the correct response option for an item should be chosen more frequently than the incorrect options. Difficult items will have a lower p-value. For standard one answer, four option multiple-choice items, p-values less than .30 (a value slightly higher than the chance for guessing it correct) should be flagged for review, as these items may be too difficult. Many times low p-values are used to find items whose wording are not clear. Similarly, items having a p-value greater than .95 may be too easy. Since 95% of the candidates are answering this item correct, this item cannot provide distinguishing information between candidates who are knowledgeable on the content versus those who are not.

Item Discrimination: Regardless of the difficulty level, an item must also be able to distinguish between low scoring candidates and high scoring candidates. If low scoring candidates are getting a particular item correct, while the high scoring candidates are missing the item, there may be a problem with the item. For example, perhaps the wording of the item results in higher scoring candidates misinterpreting the item and selecting the incorrect response option, while lower performing candidates answered the item correctly. This case is called negative discrimination. In some cases, low and high scoring candidates may perform the same on the item. This situation is called no or zero discrimination. The goal is to have positive discrimination. With positive discrimination, higher performing candidates answer the item correctly while lower performing candidates miss the item. As a result, this item has predictive ability of total exam performance. Item discrimination can be thought of as the correlation of scores on the item with examinees' total scores. This correlation is known as the point-biserial and is referred to as the discrimination index. If the discrimination index is less than .25, then the item should be flagged for review.

The numeric values provided above for item difficulty and discrimination are only guidelines. The criteria may vary depending on the purpose of the testing program. In addition, there will be cases where items will be flagged, but they will still be retained. As an example, consider an item with a p-value of .94. Since the majority of candidates are getting the item correct, there is not much room for discrimination so the point-biserial may be lower than .25. Each flagged item may have a unique situation and should be reviewed prior to eliminating the item.

In addition, an item with excellent statistics may still not be a good item. For example, an item may have a p-value of .60 and a discrimination index of .80. Statistically, it appears that this item is performing well. However, upon further review, one might

discover that candidates were only selecting options A and B, and that none or very few candidates chose options C and D. In this case, the single answer, four choice multiple choice item has in a sense become a true-false item with a 50% chance of getting the item correct. In this case, the test developer may select to flag the item for further revision.

In addition to the p-value and discrimination index, advanced item analysis uses item response theory to provide the test developer with additional information on how examinees at different ability levels perform on an item. Item response theory would be necessary if an operational test was delivered as a computer adaptive examination rather than a sequential computer based examination. With computer adaptive testing, the individual test taker's ability is considered in determining what item the candidate receives next. The item response statistics are used to help make this determination so that each candidate receives a tailor-made examination. When a candidate answers an item correctly, the candidate is presented with an equal or more difficult item. An algorithm determines when the candidate has answered enough items correctly at a given level to determine with confidence the candidate's score.

Currently, SAS Institute's certification examinations are based on the classical test theory model (e.g., p-value, discrimination index). As a result, the particular item statistics for item response theory will not be discussed.

Storing/Maintaining Acceptable Items: Throughout the process, the test developer should be maintaining a record of the items written and beta tested. However, before beginning to assemble the examination, all items accepted should be denoted as potential exam items. All items not accepted should be flagged so that they are not used in any exam forms. While some may choose to have one database with a field designating whether the item is usable or not, two separate databases can be maintained.

An organized and maintained item bank can facilitate and enhance the test construction process. The item bank provides the history of the item and should allow for sorting features to assist in test construction. As a minimum, the item bank should include the following information for each item upon completion of the beta examination.

- Unique item identifier
- Objective number from test blueprint
- Beta form
- Date of beta administration
- Sequence number of item on the exam
- Number of candidates who attempted the item
- Number of candidates selected each option
- Number of candidates who omitted the item
- Discrimination index
- p-value
- Average time to answer the item
- Author of item
- Reference for answer verification
- Cognitive level (based on chosen taxonomy)
- Type of item (e.g. single answer multiple-choice)
- Equivalent items (i.e. similar items that should not appear on the same form)
- Graphic link, if graphic is part of item
- Item status (e.g. new, experimental, secure, non-secure)
- Comments

While many of these initial fields can be completed at the end of beta testing, operational exam results should be added once obtained through ongoing test maintenance. In addition, several fields are completed pre-beta during the development of items.

Items included on an examination can be selected through an automated item bank by randomly selecting items to meet prespecified parameters. For example, a 100 item exam with an overall p-value of .73, a discrimination index of at least .50, and no more than 20% of the items at the lower, knowledge cognitive level could be specified. The larger the item pool the more flexibility and capability the test developer would have in constructing the examination.

STEP 7: CONSTRUCTING EQUIVALENT EXAM FORMS

While designing the exam to meet the specifications of the test blueprint, test developers should also strive to maximize the reliability of the examination. This characteristic assures that the same results could be replicated if the same candidates were tested again under similar circumstances. A commonly used index to measure reliability of certification examinations is the KR (Kuder-Richardson) 20 coefficient. This value ranges from 0 to 1. The goal is to obtain the highest reliability estimate possible.

The value of the KR20 coefficient is directly related to the number of items on the exam. The more items on the exam, the higher the reliability of the examination. For exams with 150 or more items, reliability indices may be in the high .80s or low .90s. However, shorter exams with 50-100 items should have minimum coefficient values in the low .80s or high .70s. (Dungan, 1996)

For many reasons (e.g. test security, repeat test takers), it is desirable to have multiple forms of an exam. If multiple examinations are constructed, it is critical that the examinations are operationally equivalent from a content and statistical perspective, as well as being reliable measures.

The first step in establishing equivalence is to ensure that the examinations align with the test blueprint. This level of equivalence provides content validity to the examination and ensures equivalence at the content level. No matter which form a candidate receives, the candidate will have the same number of items on a particular topic as a candidate receiving a different form.

The second step is to ensure statistical equivalence. Candidates should not be penalized for taking harder versions of an examination, nor rewarded for taking an easier version. Test developers can control for statistical equivalence when constructing the exam by careful selection of the items. By using the beta test results, items can be selected so that pre-equated tests forms are generated. In developing pre-equated forms, the items selected should yield, at a minimum, equivalent average p-values. It is also desirable to have equivalent point biserials, time required to complete the items, mean scores, standard deviations, reliability, skewness, kurtosis, and standard error of measurement for all forms.

Unfortunately, without an adequate number of items, achieving equivalence at the p-value may be challenging. In addition, even though equivalence is obtained, if the beta sample is not representative of the population sample, than the equivalence may not hold for the operational forms.

To ensure that forms are statistically equivalent, a process called equating is used. The method discussed above is one preliminary method. However, it can only provide a small guarantee for equivalence. One common technique for equating certification examinations is to administer different groups of candidates a common set of items. While Form A and Form B differ, they share a common set of items, called the anchor set. Generally, about 20% of the total number of items on the test or 20 items, whichever is greater, should be used as the anchor set (Angoff, 1984). However, for the anchor set to perform as the equating set, special consideration must be taken in selecting items as part of this set. As a minimum, these items should have high discrimination power and be representative of the overall content of the examination.

Candidate performance on the anchor set can then be compared with performance on the unique items. If average candidate performance on the anchor set is higher for Form A than Form B,

but performance on the unique items is lower, than the unique items on Form A may be more difficult than those on Form B. As a result, a statistical adjustment is made to candidates receiving Form A so that they are not penalized for taking a harder set of unique items. This adjustment is made prior to comparing the candidate's score with the cut score.

This equating design allows for the regular revision of current examinations and the introduction of new examinations into the testing cycle. By using the common anchor set across all operational forms, the process ensures that all versions of the exam are at the same difficulty level as Form A, the base version.

Many other equating techniques exist using both classical and item response theory. Readers interested in a description of these methods should consult Kolen (1995).

STEP 8: ESTABLISHING THE PASSING SCORE

After the examinations are constructed, the passing score for the exam must be determined. In accordance with testing guidelines, pass/fail standards must be established in a manner that is generally accepted as being fair and reasonable, rather than arbitrarily set.

There are two broad categories of standard stetting: normative and absolute. Normative standards make pass/fail decisions based on how a candidate performs relative to the other candidates. The percentage of candidates that will pass is determined prior to the test administration. Candidates pass based on where their score is in comparison to the other candidates. An example of normative standards is an examination used for scholarship purposes. The sponsoring agency may know they can only provide scholarships to the top 10% of applicants.

Absolute standards, also called criterion-referenced standards, establish a specific level of performance which must be attained. Pass/fail decisions are made based on whether this level is met, regardless of the number of candidates passing. Certification examinations typically use absolute standards, as their purpose is to ensure that a specified level of competency has been met.

The most commonly used and widely accepted method for establishing the passing score on certification examinations is the Angoff (1971) method (Sireci & Biskin, 1992). While most certification programs developed within the past two decades are using Angoff, another popular technique in the IT industry is contrasting groups. This discussion will focus on the Angoff method as this method was used for SAS Institute's V6 certification examinations. Readers interested in learning more about the other techniques are referred to Crocker and Algina (1986), NOCA (1996), and Impara (1995).

The Angoff Method: The first step in the Angoff method is establishing the committee of SMEs, called judges. As with the job analysis, item writing, and item review, the group established should be representative of the profession and familiar with the level of knowledge for which the certification is intended. It is critical to include individuals at the level for which you are certifying in the standard setting process. For example, if the certification is intended for entry into the profession, than entry-level professionals should serve on the committee. These 'non-experts' can provide useful discussion material into the characteristics that certified professionals should possess. The size of the standard setting group is not as important as the composition of the members. However, general practice recommends no less than five judges should be used.

The judges must first agree upon the definition of the minimally qualified candidate. The judges are then asked to think of a group of minimally qualified candidates and, for each item, independently determine "what is the probability that a minimally qualified candidate *will* get this question correct?" For each item, the judges determine the average Angoff rating. The average of the averages across all items is the Angoff passing score. The Angoff rating for each item should be recorded in the item bank.

If only 5 judges are used, all ratings should be used. As the number of judges increases, the highest and lowest ratings can be deleted if these ratings are outlier values (i.e., 20% away from their closest neighboring rating).

The difficulty of the Angoff approach is in conceptualizing the definition of the 'minimally qualified' candidate. As a result, participants typically need to review the definition repeatedly during the process. Discussions among the judges, after their independent ratings, can be helpful to judges forming their final item ratings. In addition, review of the beta item statistics can also be helpful to the judges in conceptualizing the 'minimally qualified' candidate and providing a crosscheck of their ratings. Judges, however, must remember that the beta exams reflect all candidates, not just the 'minimally qualified' candidates and must use these data with caution.

STEP 9: ADMINISTERING/SCORING OPERATIONAL EXAMS

Once the passing point is established for the exams, the exam is ready for administration. Traditionally, examinations were given via paper-and-pencil in a group setting at a particular time. However, in the past several years, certification examinations have begun to shift to computerized administrations, particularly in the IT industry.

The importance of standardized testing administrations is directly addressed in the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999). The testing environment should be reasonably comfortable and have minimal distractors. The test administrators must follow standardized procedures. The exams must be securely stored and the administration must be proctored to ensure the validity of the test scores.

One major advantage to computerized testing over paper-and-pencil testing is the availability of immediate scores. With traditional paper-and-pencil testing, scores could not be calculated until all answer documents were returned to the central scoring location. The documents had to be scanned in order to process the scores. With computerized testing, the scoring issue becomes how much information should be provided to the candidate. Rosen (1996) suggests that candidates be given as much information as possible. The raw score should be provided along with section scores and the required passing score. Any equating or raw-to-scale score conversion would be done at the end of the examination. Regardless of what is reported, clearly written explanatory material should be provided so that scores are not misinterpreted.

Another advantage to computer administration is that greater standardization can exist than with paper-pencil examinations across a global setting. In addition, exams are available throughout the year for increased testing flexibility, rather than specific test dates.

Another issue regarding exam administration is repeat testing. It can be expected that not all candidates who take a certification examination will pass. Some may not pass due to lack of knowledge or readiness. Others may not pass due to situational reasons such as temporary illness or high-test anxiety. While examinees deserve the chance to be retested, some guidelines should be established. Whenever an examinee takes an exam, they have 'practiced' taking the exam. The more 'practices' a candidate has, the better chance for an increased test score. This increase in test scores is called a practice effect. Certification examinations are designed to ensure that those achieving the credential possess the appropriate level of knowledge. The validity of the candidate's score will be compromised if the practice effect is high. As a result, guidelines should be established as to how often an examinee can repeat

the examination within a given time period. SAS Institute's Certification Program requires a minimum of two months between testing and a maximum of three testing opportunities within a twelve-month period.

STEP 10: PROVIDING ONGOING TEST MAINTENANCE

At defined intervals throughout the testing cycle, item level and test form statistics should be reviewed. The operational data should be compared to the beta item statistics. In addition, periodic review of the statistics ensures that the keys are accurate and that the items are performing as intended. Similar patterns between the beta and operational examinations provide another measure of the exam's content validity. The final operational item statistics should be recorded in the item bank as part of the permanent item history.

The following data should be obtained and recorded at the end of each defined interval:

- group mean
- standard deviation
- standard error of measurement
- highest and lowest scores obtained
- the percent of candidates passing
- group mean on the anchor set of items, if using anchor sets, and the unique set
- exam reliability

Collection of these data allows the certifying organization to monitor the consistency of test form statistics, candidate characteristics, and the passing rate over time. For example, suppose the passing rate increases from 60% to 90% during one quarter. This unreasonably high jump should raise a flag to the certifying organization. Perhaps a new training course was developed and is responsible for the increase in scores, or perhaps the security of the examinations has been compromised. As another example, suppose the passing rate decreases substantially, then the certifying organization may want to ensure that there is not an error in the answer key or scoring program.

CONCLUSION

Candidates seeking a particular credential deserve the opportunity to take an examination covering material that is appropriate for the performance required for the credential. In addition, this examination should provide a reliable measure of the candidate's knowledge level. The examination should also have the ability to distinguish between those who deserve the credential and those who do not.

There are a variety of strategies and methods that can be used to develop a certification examination. The above steps outline the activities that should take place to ensure a valid, reliable, and defensible examination. Depending on the particular needs or situation of a certifying organization, specific activities may need to be added, altered, or rearranged. However, even minimal deviation from these steps can create a threat to the credibility or integrity of the examination.

While these steps provide assurance for developing a valid and reliable certification examination, a certifying organization should validate their efforts by conducting reliability and validity studies. As an example, the certifying organization can research whether candidates perceived as being highly qualified, given their number of years of experience, are scoring significantly higher on the exams than candidates perceived as minimally qualified.

The model described in this paper is fairly standard regardless of the testing program. However, many items could only be mentioned briefly although they are critical components. In addition, there are many options available within each step. Regardless of the options selected, adherence to the basic steps in the model should fulfill the ultimate intent of the certification program to evaluate the knowledge and skills of practitioners seeking a credential in a reliable and valid manner.

REFERENCES

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Bloom, B. S., Englethart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain.* New York: David McKay.

Browning, A.H., Bugbee, A.C., Jr., & M.A. Mullins (Eds.) (1996), *Certification: A NOCA handbook.* Washington, DC: National Organization for Competency Assurance.

Council on Licensure, Enforcement, and Regulation & National Organization for Competency Assurance. (1993). *Principles of Fairness: An examining guide for credentialing boards.* Lexington, KY: Author.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Orlando, FL: Holt, Rinehart and Winston, Inc.

Dungan, L. (1996). Examination Development. In A.H. Browning, A.C. Bugbee, Jr., & M.A. Mullins (Eds.), *Certification: A NOCA handbook.* Washington, DC: National Organization for Competency Assurance.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. (1978, August 25). Uniform Guidelines on Employee Selection Procedure. *Federal Register,* 43 (166), pp. 38290-38315.

Flaherty, V. L. & Hogan, J. B. (1998). Job analysis for high-stakes credentialing examinations. *CLEAR Exam Review,* IX (3), 23-28.

Haladyna, T. M. (1999). *Developing and Validating Multiple-Choice Test Items* (2nd edition), Mahway, NJ: Lawrence Erlbaum Associates.

Henderson, J. P. (1996). Job analysis. In A.H. Browning, A.C. Bugbee, Jr., & M.A. Mullins (Eds.), *Certification: A NOCA handbook.* Washington, DC: National Organization for Competency Assurance.

Impara, J. C. (Ed.) (1995). *Licensure Testing: Purposes, Procedures, and Practices.* Lincoln, NE: Buros Institute of Mental Measurement.

Kearns, Dave (1998). Certified, but qualified? *Network World, 52(1).*

Kolen, M. J., & Brennan, R. L. (1995). *Test Equating: Methods and Practices.* New York: Springer-Verlag.

Osterlind, S. J. (1997). *Constructing Test Items* (2nd edition). Boston, MA: Kluwer Academic Publishers.

Rosen, G. A. (1996). Test Administration. In A.H. Browning, A.C. Bugbee, Jr., & M.A. Mullins (Eds.), *Certification: A NOCA handbook*. Washington, DC: National Organization for Competency Assurance.

Sicerci, S. A., & Biskin, B. H. (1992). Measurement practices in national licensing examination programs: A survey. *CLEAR Exam Review*, III (1), 21-25.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Linda A. Althouse
Training Sales and Marketing
SAS Institute, Inc.
SAS Campus Drive
Cary, NC 27513
919-677-8000
919-677-4444
linda.althouse@sas.com