BUILDING AND MANAGING SMALL EXAMINATION PROGRAMS[1]

By

Roberta N. Chinn, Ph.D.
Norman R. Hertz, Ph. D.
and
Barbara A. Showers, Ph.D.

[1] This resource brief was based in part on the session entitled "Building and Managing Small Examination Programs" presented at the annual meeting of the Council on Licensure, Enforcement, and Regulation in Las Vegas, NV, September 2002.

TABLE OF CONTENTS

INTRODUCTION 1

PSYCHOMETRIC CONSIDERATIONS 2

- Practice (Job) Analysis 2
- Test Specifications 3
- Item Development 3
- Test Construction 4
- Test Equating 5
- Cut Score 5
- Test Administration 6
- Item Analysis 6
- Feedback 7

PRACTICAL CONSIDERATIONS 7

- Collaborative Arrangements 7
- Focus Groups 9

COMMONLY ASKED QUESTIONS 11

BIOGRAPHICAL SKETCHES 13

INTRODUCTION

Small examination programs must find innovative ways to fulfill their obligations to candidates and consumers. These programs have the same duties as their large-scale counterparts with fewer resources and limited budgets. Their responsibilities include identifying minimum candidate qualifications for eligibility, developing candidate handbooks, determining test specifications, establishing cut scores, scoring or grading examinations, maintaining item banks, evaluating examination results, sending score reports, and revoking licenses or certificates of individuals who violate professional codes of conduct.

There are many benefits of creating and maintaining collaborative relationships with other regulatory agencies and professional associations. Programs that pool their resources can produce higher quality examinations than those that could have been done by the programs themselves. While the obvious reason to pool resources is monetary, small programs can share information regarding candidate handbooks, administrative procedures, legislation, and recruitment of subject matter experts. They can also share psychometric expertise if they hire a psychometrician to provide services to their programs. A psychometrician with experience in regulation can be invaluable in assisting small programs to meet standards by performing an objective evaluation of the program and offering recommendations for enhancement of that program.

Several psychometric and practical considerations will be addressed to ensure the reliability and validity of examinations, regardless of program size or candidate volumes. These considerations are discussed in the context of collaborative relationships with other agencies or organizations.

PSYCHOMETRIC CONSIDERATIONS

Small examination programs should be aware that they must have the same elements of validity as large programs in order to meet professional testing standards[2]. The major elements include:

- Foundation and structure of the examination (practice analysis, test specifications),
- Procedures necessary for creating items (item development),
- Procedures for constructing and equating forms (test construction, test equating),
- Procedures for establishing cut scores,
- Procedures for administering the examination (test administration), and
- Procedures for evaluating the examination (item analysis, feedback).

Practice (Job) Analysis

A practice analysis defines practice in terms of the job competencies required by the job. Job competencies can be identified by obtaining practice analyses from other jurisdictions or from professional organizations and extracting information relevant to the jurisdiction of interest. The job content may not be the same for every jurisdiction, but it is a place to start. The ultimate focus of the job competencies should be competencies critical to public protection.

Once the job competencies are identified, a survey questionnaire can be developed and distributed to a representative sample of persons who work in the field. Professional organizations, government agencies, and universities can be helpful in providing lists of people who work in the field. If there are not sufficient persons to provide meaningful data, a focus

[2] Professional testing standards include the Standards for Educational and Psychological Testing (1999) developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education.

group of subject matter experts can be convened to determine the critical competencies. The focus group strategy is ideal for new or emerging professions.

Test Specifications

Test specifications define the subject matter areas as well as the percentage of the subject matter areas (weights) to be covered on the examination. Once defined, test specifications provide a definitive source from which item writers can develop the items for the examination and candidates can become familiar with subject matter areas covered.

Test specifications can be developed in one of two ways. The subject matter areas and their weights can be derived empirically from the data obtained from a practice analysis survey questionnaire. If there will not be sufficient data to define the specifications, the alternative is to convene a focus group of subject matter experts to determine the subject matter areas and their weights.

Item Development

Item development should be a systematic process that includes numerous opportunities for independent review. Regardless of candidate volume, several procedures are essential to the development of quality items if the items are to meet professional testing standards. There is a tendency for small programs to minimize the role of independent review and systematic procedures in the item development process as a cost-saving measure. However, in the event of a challenge, programs that take shortcuts may discover that they have flawed items.

First, the item writers should be formally trained in item construction techniques so that the items produced can survive independent review. Second, there should be opportunities for independent review of the items by persons other than those who wrote the items. Independent review improves item quality because issues are considered from fresh perspectives. Third, the

items should be maintained in a database (item banking software) to permit instantaneous inventories of item status, history of the items, statistical measures of item performance, etc.

Fourth, reliable scoring procedures should be developed and built into item development. For example, conventional multiple-choice items are usually worth the same number of points per item. This means that item writers should develop items of equal difficulty rather than trying to "weight" the difficulty of one item over another by making the content of one item more complex than other items. With essay, practical or oral formats, item writers must be able to understand what the scoring criteria are and how the scoring procedures will be applied when structuring the content of the items.

Test Construction

Test construction (publication) is often overlooked as a key procedure in the development of a valid examination. If the procedure is conducted in a systematic manner, it can help ensure that multiple forms of an examination are more equivalent in difficulty and content; and conform to the test specifications. Because there may not be sufficient candidate volumes to use item statistics to guide selection of items, the easiest way to construct an examination is to convene a standing committee or a group of 6-8 subject matter experts to select the items for each form of the examination. Under the direction of a psychometrician, the subject matter experts would select the items based on the difficulty of content, coverage of content outlined in the test specifications. Several other factors should also be considered in the test construction:

- The item is written at the level of minimum competence;
- The item content is comparable in difficulty;
- The item content is relevant to actual job situations;
- The key and distractors are clear, unambiguous;

- There is minimal overlap of content; and,
- The percentages prescribed by the test specifications are covered.

Test Equating

Equating different forms of an examination can be a challenge for small examination programs. Statistical equating methodologies such as item response theory (IRT) and equal percentile may be inappropriate for small programs because there is insufficient candidate data to use such methodologies.

One approach is to equate the tests by means of a cut score study. In the cut score study, a focus group is convened to establish difficulty ratings for an entire bank of items. Then, another focus group constructs the tests according to the difficulty ratings of individual items. There can be a different mix of items on different forms of the examination so long as the sum of the difficulty ratings is the same for a given form and the content of the items is considered comparable.

Cut Score

The cut score, or passing score, is by far the most visible aspect of an examination program and one of the critical elements in determining the validity of the examination. The most defensible cut score is one that is established according to criterion-referenced methods, e.g., Angoff, Nedelsky, Ebel, etc. The criterion-referenced methodologies focus on minimum standards for competent practice and result in a cut score that reflects the difficulty of the items contained in the examination. Since the methodologies are largely dependent upon judgments of a group of individuals and skill of the facilitator, it is advisable to enlist the assistance of a qualified psychometrician in establishing the cut score and providing thorough documentation of

the process. In the event of a challenge, a program will have validity evidence that the cut score was set according to professional testing standards.

Test Administration

Test administration procedures are as important as any of the procedures in an examination program in establishing the reliability and validity of the tests and ensuring that the examinations are fair and unbiased. Good test administration procedures will ensure that all candidates are afforded the same testing experience. Standardized administration procedures should be established for candidate registration, proctor training, candidate seating, inventory of examination materials, storage of examination materials, and incident reports. The greatest benefit of standardized procedures is examination security because there is a way to track candidates throughout the administration process.

Item Analysis

There is no set number of candidates on which to perform an item analysis of multiple-choice items. Item analyses can be performed with as few as 20 candidates to obtain information about item performance, e.g., if the items are too easy or too hard, if the items do not discriminate between high scoring and low scoring candidates, which distractors are not effective, etc. The caveat, however, is that the item statistics for small numbers of candidates provide a gross measure of item performance that must be subject to expert judgment (subject matter experts) before making substantive revisions to the items. Small programs are well advised to aggregate data for items over multiple administrations. Aggregating data is the same process used by large-scale programs, but it may take several administrations to accumulate data sufficient for obtaining stable statistics.

For essay, practical, and oral examinations, statistical analyses on as few as 20 candidates can assist in identifying situations where graders or examiners are not using the same scoring criteria. Statistical analyses may include reliability of examiner ratings, percentage of agreement between pairs of examiners, and descriptive statistics of candidate scores.

Feedback

It is important for small programs to obtain feedback because the results from statistical analyses may not be definitive for several administrations. One source of feedback is from comment sheets available for candidates at the examination. Another source of feedback is from subject matter experts. Finally, feedback obtained during routine item review or cut score workshops can be helpful in identifying items that may need revision in light of current standards of practice, current laws and regulations, etc.

PRACTICAL CONSIDERATIONS

Collaborative Arrangements

The key to successful collaboration is having interested parties involved throughout all aspects of practice analysis, test specifications, and item development. This means that the terms of the collaborative arrangements should be clearly documented to minimize misunderstandings. All parties should enter into arrangements with a spirit of mutual trust and cooperation as well as willingness to reach a compromise regarding process and procedures that do not sacrifice adherence to testing standards.

The arrangement should also be documented and reviewed by each party's legal counsel. Clauses in the agreement can address the following issues:

Intent of the examination. All parties should agree on the purpose of the examination, e.g., whether it is to be used for assessing minimum competence or advanced certification.

Ownership of the items. Ownership of the items can have an impact on the security of the items and how the items are used. The items may have initially been the property of one party, but after the items have been validated for all jurisdictions, the items would be transferred to and maintained by all parties under secure conditions.

Validation of the items. Formal validation will ensure that the content of the items is relevant to a specific jurisdiction. Traditional methods, e.g., focus groups of subject matter experts, can be used for establishing job-relatedness of items.

Maintenance of the items. Maintenance of item content should be a shared responsibility so that the items are current, although one party should be designated to maintain the master item bank.

Test security. The items and all examination materials will need to be maintained under secure conditions. Access to the items and examination materials should be restricted to maintain test security. All persons who have access to the examination should sign security agreements or nondisclosure agreements.

Item development. All parties should share the responsibility to develop and review the items. All parties can share the costs of psychometric services whether those services are provided by in-house staff or an independent psychometrician.

Cut score. The cut score can be the same for all jurisdictions if all parties agree to use a common practice analysis, use the same test specifications, and establish the cut score with representation from each jurisdiction.

Test scoring, item analysis and score reporting. Item statistics can be aggregated to provide a larger sample and thus stable statistics. Each party can determine what shall be included in score reports and release its own score reports to its candidates.

Separation of testing functions from other functions. There may be issues that arise when government jurisdictions and professional organizations use the same examination. Candidates should not be required to hold membership in the professional organization as a prerequisite to take the examination. Moreover, the examination functions of the organization should be handled independently from membership and membership functions. Government jurisdictions should check their specific regulations and the membership organization structure before entering into this type of arrangement.

Remuneration of subject matter experts. Each jurisdiction should decide whether or not to provide honoraria and travel expenses to subject matter experts. It is advisable to pay honoraria and travel expenses to subject matter experts to emphasize the importance of the work to be performed.

Focus Groups

Because small examination programs rely heavily on focus groups, their composition is very important. The following factors should be considered when selecting focus groups of subject matter experts:

Relationship to the board. The majority of the experts in a focus group should not be board members or board staff. This will help maintain independence of the examination and board functions.

Practice specialty. The experts in the group should reflect the practice specialty mix of the profession.

Understanding of regulation. There should be a clear understanding that the examination will be used for regulatory (minimum competence) rather than for hiring or promotional purposes.

Region of practice. The experts in the group should represent as many geographic regions of practice as possible.

Tenure as a licensed practitioner. The experts in the group should include both newly licensed and experienced practitioners. In selecting experienced practitioners, it would be preferable to select practitioners who sat for an examination rather than those who were licensed under grandparent statutes. There are some practitioners granted a license under grandparent statutes that may not understand minimum competence or current training and experience needed to practice.

Independent perspective. Some credentialing bodies try to use the same subject matter experts for item writing, item review, and cut score focus groups. While continuity is important to group productivity, it is always wise to have several subject matter experts who have not participated before. This is particularly true for cut score studies because the experts can lose their sense of objectivity thereby unduly affecting the resultant cut score.

COMMONLY ASKED QUESTIONS

QUESTION Is it really necessary to hire a psychometrician? We've heard that they are very expensive. Some of our board members are experienced educators and believe that our board can get by without one.

ANSWER There are many benefits of hiring a psychometrician. While some psychometric organizations work only with large-scale programs, there are some psychometricians who have considerable experience working with small programs.

There is a general consensus that boards are better off allocating resources to develop a valid examination than trying to mitigate situations created by problem licensees. By hiring a psychometrician, a board can devote its efforts to the administrative functions of the program. A board can retain the rights to approve final content of the examination and the cut scores but not get caught up in the intricacies of test development.

The most important benefit is technical expertise. Board members and other interested persons may have good intentions but lack the skills to oversee development and maintenance of examinations. It should be emphasized that development and maintenance of high stakes examinations requires training and experience in order to meet professional testing standards.

QUESTION: How many forms of an examination should we develop? Our board has a very small budget and we might not have sufficient funds to develop more than one form.

ANSWER: There should be a minimum of two forms developed for each administration of the examination. Why two forms? Within an examination cycle, one form is given to all candidates on the designated date and the other form is reserved in the event that the examinations are stolen or compromised, persons retake the examination or take the examination on an alternative date, etc.

QUESTION We do not have sufficient candidate volume to pretest items as is done in large-scale programs. How can we determine the quality of our questions?

ANSWER It is true that pretesting is not feasible for programs that administer examinations to only a few candidates per year. The best thing to do is to convene a focus group of subject matter experts who represent the practice specialty mix of the profession and have them review the items.

QUESTION: It would greatly simplify our scoring reporting process if we could use a fixed percentage as the cut score rather than go to the trouble of conducting a cut score study. What would you advise?

ANSWER: The best option for a defensible cut score is to take the time to conduct a formal cut score study. In the event of a challenge, validity evidence will be available to support the cut score process.

The problem with using a fixed percentage is that it may not reflect the difficulty of the items in the examination. Therefore, persons who take an examination comprised of difficult items are not being treated the same as persons who take an examination comprised of easy items. A criterion-referenced cut score study will yield a standard of practice that remains the same regardless of the difficulty of the examination.

If the item bank contained items of equivalent content and difficulty, the cut score could be used as a percentage score from which to equate other forms of the examination.

QUESTION: We've been thinking about administering a take-home examination that covers relevant laws and regulations. This would be a cost-effective means of administering the examination. What factors should be considered?

ANSWER: To the layperson, take-home examinations appear to provide an attractive alternative for small boards with tight budgets. The decision to provide a take-home versus an examination offered under secure conditions depends largely upon the expectations regarding the outcomes of the examination.

If the intent is to protect the public, it is likely that a high-stakes examination should be developed and maintained under secure conditions. If the intent is to assess candidates' knowledge of laws, regulations or ethical standards that can be looked up, it is likely that the examination can be administered as a take-home examination.

BIOGRAPHICAL SKETCHES

ROBERTA N. CHINN, Ph.D. is the general partner of HZ Assessments in Folsom, California. She was formerly the lead technical specialist at the Office of Examination Resources at the California Department of Consumer Affairs. She received her Ph. D. from Louisiana State University in experimental psychology. She has conducted validation studies; written, performance and oral examination development; and established cut scores for over 40 professions including dental auxiliaries, engineers, geologists, marriage and family therapists, psychologists, social workers, speech language pathologists, and veterinarians. She is a member of the American Psychological Association; American Educational Research Association; National Council on Measurement in Education; and the Council on Licensure, Enforcement and Regulation. She has authored numerous articles and publications in the measurement field and has presented research at various professional conferences.

NORMAN R. HERTZ, Ph.D. is a licensed psychologist and the managing partner of HZ Assessments in Folsom, California. He was formerly the Chief of the Office of Examination Resources at the California Department of Consumer Affairs. He received his Ph.D. from University of Memphis in industrial psychology. He has extensive experience in private industry and government settings and has conducted validation studies, developed licensure and certification examinations, and established cut scores for more than 80 professions ranging from construction trades to medical specialties. He has also served in various national examination committees. He is a member of the American Psychological Association; Society for Industrial Organization Psychology; American Educational Research Association; National Council on Measurement in Education; and the Council on Licensure, Enforcement and Regulation. He has authored numerous articles and publications in the measurement field and has presented research at various professional conferences.

BARBARA A. SHOWERS is the Director of the Office of Education and Examinations for the Wisconsin Department of Regulation and Licensing in Madison, Wisconsin. She received her Ph. D. from Michigan State University in measurement evaluation and research design. She has managed credentialing programs for 30 national examinations and 55 in-house examinations. She has extensive experience in job analysis, examination development, validation, test administration, computer-based testing, and contract negotiation. She has served on various national examinations committees. She is a member of the American Educational Research Association; the Council on Licensure, Enforcement and Regulation; and the National Council on Measurement in Education. She has authored, co-authored, and contributed numerous articles and publications and presented research at various professional conferences.