An NCME Instructional Module on Developing and Administering Practice Analysis Questionnaires

Mark R. Raymond, *American Registry of Radiologic Technologists*

The purpose of a credentialing examination is to assure the public that individuals who work in an occupation or profession have met certain standards. To be consistent with this purpose, credentialing examinations must be job related, and this requirement is typically met by developing test plans based on an empirical job or practice analysis. The purpose of this module is to describe procedures for developing practice analysis surveys, with emphasis on task inventory questionnaires. Editorial guidelines for writing task statements are presented, followed by a discussion of issues related to the development of scales for rating tasks and job responsibilities. The module also offers guidelines for designing and formatting both mail-out and Internet-based questionnaires. It concludes with a brief overview of the types of data analyses useful for practice analysis questionnaires.

Keywords: licensure, certification, job analysis

Each year, hundreds of thousands of individuals pay the fees required to take a licensure or certification exam for the purpose of documenting their qualifications to practice their chosen occupation or profession.[1] Some of these regulated professions are quite familiar, such as teaching, nursing, and law. Meanwhile others are relatively obscure, such as underground storage tank installation, milk testing, and crane operation. However, most have one thing in common: They require passing a high-stakes exam. Passing the exam permits the individual to use a particular title and/or to engage in certain activities associated with a profession, while failing the exam usually means just the opposite. Given the importance of the decisions made based on scores from credentialing exams, it is imperative that they be carefully developed and evaluated.

According to the *Standards for Educational and Psychological Testing*, credentialing examinations are intended to provide the public, employers, and government agencies with a reliable method for identifying practitioners who have met certain standards (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999). For much of the previous century, it was common for credentialing exams to assess the knowledge, skills, and abilities (KSAs) covered in the training programs for a profession. This was quite convenient because it meant that test plans for certification exams could be based on existing curricula and textbooks. However, in the past 30 years the goals of credentialing exams have shifted, and it is now recognized that for such exams to fulfill their mis-

sion they should cover the KSAs required for effective practice. In other words, credentialing exams should be job related (AERA et al., 1999; D'Costa, 1986; Kane, 1982; National Commission for Health Certifying Agencies, 1981; Shimberg, 1981). Although textbooks and curricula will certainly influence test content, the *Standards* note that credentialing exams should be limited to essential skills required for safe and appropriate practice (AERA et al., p. 156).

The preferred method for ensuring the job relatedness of an exam is to include a practice analysis as part of the test development process (AERA et al., 1999; Smith & Hambleton, 1990). A practice analysis—similar to a job analysis in the field of I/O psychology—is the systematic study of a profession undertaken to identify and describe the job responsibilities of those employed in the profession. Once the professional responsibilities have been identified, it is possible to determine KSAs required to effectively carry out those responsibilities. These KSAs then serve as the basis for test specifications or test plans. Practice analysis and the development of test plans are two related, but distinct, activities (Harvey, 1991; Raymond, 2001). This module focuses primarily on practice analysis. Although the development of test plans is certainly an important topic, it is also

Mark Raymond has been Director of Psychometric Services for the ARRT since 1992 and was employed by ACT, Inc. prior to that. He has directed job analysis projects and developed test specifications for numerous occupations and professions. Correspondence: 1255 Northland Drive, Saint Paul, MN 55120; mark.raymond@arrt.org.

a complex one deserving a separate article (see Raymond & Neustel, in press; Wang, Schnipke, & Witt, 2005).

The first section of this module briefly addresses project management. Although this topic is somewhat mundane and is hardly the reason why most of us went to graduate school, it is important because a detailed project plan is essential for estimating resource requirements and completing a practice analysis in a timely fashion. Next, I discuss questionnaire development, giving emphasis to the task inventory method of practice analysis. Then, I address sampling and survey administration, including discussion of Internet-based practice analysis questionnaires. The module concludes with an overview of the types of statistical analyses useful for describing professional practice.

Project Management

A practice analysis requires a substantial investment of time and resources, and can involve contributions from a large number of individuals. A thorough study typically requires from 6 to 18 months to complete; if test plans are to be developed, another 3 to 9 months can be added to the timeline. Work will likely be completed by a project director with a background in research, as well as one or more panels of subject matter experts (SMEs). The panel of SMEs should be broadly representative of the profession being studied, including entry-level practitioners, educators, administrators, and content experts. Other members of the project team might include consultants with certain types of expertise, as well as administrative and clerical support.

Any practice analysis should be guided by a statement of purpose. The purpose might have been stated as part of a request for proposal (RFP), or might be determined by project staff and SMEs once the project is under way. Either way, the purpose will influence decisions related to questionnaire content, rating scales, sample size, and data analyses. A study to be used for developing only test plans will be different from one intended to develop both test plans for a certification board and curriculum materials for training programs.

A project timeline is a very useful management tool. The project timeline identifies each activity to be completed,

when it is to be completed, and the person responsible. It might also include other details, such as the scope of the project, whether the questionnaire will be mailed or distributed via the Internet, the resources required (e.g., postage, envelopes, data entry software, printing costs), and a brief description of any products such as progress reports and technical summaries. The project schedule serves many purposes ranging from determining resource requirements to communicating expectations for staff, contractors, and consultants.

I have often found it useful to rely on two or more project schedules. A general, high-level schedule is useful for outlining major activities, meetings, and products. Then, additional schedules can be developed for parts of the project that may require a greater level of managerial control. For example, survey printing, mailing, and data entry are typically completed in a few months, but during that time many activities must be completed in a timely fashion, usually by multiple personnel. A detailed schedule just for this part of the project can help ensure that important activities are not overlooked. The importance of project schedules cannot be overstated. After all, the U.S. Postal Service cannot be expected to deliver 2,000 business reply envelopes unless someone on the project team deposited adequate funds into the appropriate account. The monograph by Bourque and Fielder (2003) offers many strategies regarding budgeting and project management.

Practice Analysis Questionnaires

Up through the 1970s and 1980s, it was common for credentialing agencies to convene a panel of SMEs and ask them, within the course of a two-day meeting, to specify the tasks and KSAs to be covered by the test plan. Although involving SMEs is essential, most credentialing agencies now recognize the importance of supplementing SME judgments with an empirical study. This helps to ensure that test plans are broadly representative of actual practice. There are various empirical approaches to practice analysis. Two procedures that have been especially effective within the context of credentialing are the critical incident technique (Flanagan, 1954) and the professional practice model (LaDuca, 1994).

However, most certification boards rely on a method of job analysis that has enjoyed widespread use for many years: the task inventory questionnaire and its variations (Newman, Slaughter, & Taranath, 1999).

A task inventory is a list of activities thought to be performed by those who work in a particular profession. The task inventory is formatted into a questionnaire and mailed to a large sample of individuals who are asked to rate each task on certain scales. The scales might ask, for example, how often each task is performed and how difficult it is. A task inventory provides an efficient way to obtain information about a variety of work-related activities from numerous individuals. This is especially important, because credentialing examinations are intended to gauge an individual's readiness for a wide range of activities in a variety of settings (Kane, 1982). In addition, response rates for task inventories are generally quite good, and they produce large amounts of data conducive to many types of statistical analyses. Such analyses can be helpful in understanding the dimensions of work that underlie professional competence, in identifying subspecialties, and in developing empirically derived test specifications (Kane, 1997; Raymond, 2001). This module focuses primarily on the task inventory questionnaire. A portion of a hypothetical task inventory questionnaire appears in Figure 1.

Development of a Task Inventory

Developing task inventory questionnaires is mostly about determining the questions to be asked and designing rating scales for eliciting responses to those questions. Because mail-out and Internet surveys do not provide an opportunity to interact with respondents, questions and rating scales need to be perfectly clear to obtain unambiguous responses (Desimone & LeFloch, 2004). This section of the module addresses editorial guidelines for writing task statements and describes some practical ways to identify the activities to include on task inventory questionnaires. This is followed by recommendations for the development of rating scales.

Editorial Guidelines. Gael (1983) defines a task as a unit of work performed by an individual that has a definite
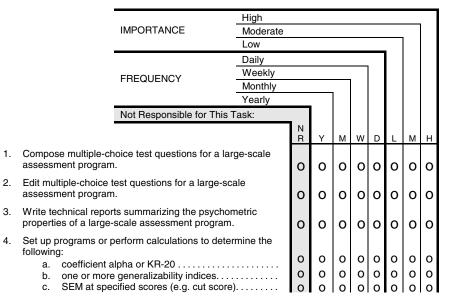
	NR	Y	M	W	D	L	M	H
IMPORTANCE — High / Moderate / Low								
FREQUENCY — Daily / Weekly / Monthly / Yearly								
Not Responsible for This Task:								
1. Compose multiple-choice test questions for a large-scale assessment program.	O	O	O	O	O	O	O	O
2. Edit multiple-choice test questions for a large-scale assessment program.	O	O	O	O	O	O	O	O
3. Write technical reports summarizing the psychometric properties of a large-scale assessment program.	O	O	O	O	O	O	O	O
4. Set up programs or perform calculations to determine the following:								
a. coefficient alpha or KR-20	O	O	O	O	O	O	O	O
b. one or more generalizability indices.	O	O	O	O	O	O	O	O
c. SEM at specified scores (e.g. cut score).	O	O	O	O	O	O	O	O

FIGURE 1. Segment of a hypothetical task inventory questionnaire. Scale definitions and instructions would precede the task inventory.

beginning and end and that results in a product or service (pp. 9, 50). This definition applies to both physical and cognitive activities (Harvey, 1991). Task statements must be written clearly and precisely if they are to provide useful information. Consider the tasks in Figure 1. One important feature of these statements is that they describe a specific activity that occurs over a short period of time. Another feature is that they have a significant cognitive component (e.g., "perform calculations"). However, even cognitively oriented activities often result in a product, service, or discernible outcome. The statements in Figure 1 also have similar syntax, following the grammatical form of subject, verb, object. In most task inventories, the subject is an implied first person that refers to the questionnaire respondent. The verb indicates the action that the respondent performs, while the object identifies the recipient of the action. Some task statements also contain qualifiers that indicate how the task is done, the tools required to complete it, its purpose, and where or under what circumstances the task is performed. Such qualifiers should be included only if necessary to clarify the task statement. For example, sometimes a task might be performed in more than one way and it is necessary to specify how the task was performed. The task "Determine measurement error associated with an observed score" might include the qualifier "based on item response theory." Sev-

eral guidelines for writing task statements appear below (also see Gael, 1983).

● One action and one object: Avoid writing task statements that contain multiple verbs or multiple objects. Consider the first two statements in Figure 1. Someone might suggest that these two statements be combined, because "compose" and "edit" are similar and require the same types of language skills. But how would someone respond to the combined statement if he or she frequently edited tests and very seldom composed them? There may be legitimate exceptions to this rule. For example, a statement may contain two verbs if two activities are almost always performed together, either because they occur in close sequence or are otherwise interdependent. Another possible exception arises when two actions or objects are joined by the word *or*, as with Task 4 in Figure 1. Is this use justifiable? I think it could be argued either way, depending on the information sought from this statement. Also note that Task 4a contains two objects ("coefficient alpha" and "KR-20"). It is hard to imagine that the use of "or" in this instance would have negative effects. All task statements with multiple verbs or objects should be scrutinized. Then decide if the multiple use hinders comprehension or if it helps. One more point; avoid the use of slashes.

● Useful level of specificity: The task, "Obtain patient's vital signs," although seemingly simple, really involves many

psychomotor and cognitive activities related to acquiring and evaluating four physiologic measurements: pulse, respiration, blood pressure, and temperature. Is it necessary to break this general task into several component tasks? There is no definitive answer, but if the personnel who perform one of the component tasks also perform the others, then numerous statements probably are not necessary. Task statements should be written at the level of specificity required to accomplish some discernible goal that is important to the profession being studied. A task inventory that consists of 25 statements is likely to provide very general information of limited utility, while a questionnaire consisting of 400 statements may be overly specific.

● Action verbs: A traditional rule for writing task statements is that action verbs should be used. This rule applies to physical and cognitive activities. Given that cognitive activities involve attending to and processing information, the task statement should describe the type of information being acted upon. The task statement might also indicate the product, information, or service resulting from the cognitive task. "Apply knowledge of generalizability theory" is hardly a useful task because it is so broad and the verb "apply" does not denote a specific action. Something like "Calculate indices of decision consistency from a table of variance components" may work better.

● Precise word choice: Ambiguity is a problem with many types of questionnaires. For example, the use of "large-scale assessment" in Figure 1 might require clarification in a footnote or in the instructions. Cognitive verbs also present special challenges in self-administered surveys. Words such as *analyze, evaluate, problem solving,* and *decision-making* may have different meanings for different respondents. It may be necessary to pilot test terms such as these to ensure that respondents have similar interpretations.

● Stand-alone content: Each task statement should make complete sense on its own. It should not depend on surrounding statements in the questionnaire to give it meaning.

● Descriptive, not prescriptive: It is tempting to write task statements like "Conduct meetings of SME panels using appropriate meeting management strategies to ensure a successful outcome." Unfortunately, we would not know what to conclude about individuals who say they seldom perform this task. Is it because

they infrequently conduct meetings, because they do not always get a successful outcome, or because they did not know which management strategies were appropriate? The purpose of the job analysis is to describe, not prescribe, practice. A related issue pertains to the use of qualifiers, such as, "as needed" or "when required," which tend to elicit positive responses even if the task is seldom performed.

Sources of Task Statements

Editing task statements is important but fairly straightforward. A more challenging activity is identifying the tasks to be included on a questionnaire. Panels of SMEs do an excellent job of this, but they need a way to structure their work, a process for completing it, and other types of support from the project director. The project director might initiate questionnaire development by mailing job-related materials to SMEs for review and input. SMEs might be given previous practice analysis reports from the same or similar professions, job descriptions, performance evaluation forms, curricula, lists of training objectives, and any other documentation that sheds light on the nature of the profession being studied. Existing records such as billing statements, patient charts, and insurance records are also excellent sources of practice-related information. Some projects might benefit from a preliminary study to help determine the content of the task inventory questionnaire. A preliminary study might involve, for example, a site visit to observe and interview employees, or possibly mailing a short survey or work diary to small samples of the population of interest.

After the project director has obtained an initial comprehensive list of activities—by whatever means—the time and effort required to finalize the list are not overwhelming. Although it is convenient to convene a meeting of SMEs, the work can be completed through the mail if necessary. It is common for early drafts of a task inventory to consist of 300 or 400 statements written at varying levels of specificity, and many of these statements will overlap. Reducing this to a manageable and coherent list with minimal redundancy requires some attention to detail. Once the task list is finalized, the project director then needs to focus on rating scales.

Types of Rating Scales

Literally dozens of rating scales have been developed for use in task inventory questionnaires over the years. The scales for rating tasks vary along several dimensions, with the most important one being the task attribute being measured (e.g., frequency, complexity). Scales vary in other ways as well, including the type of anchors (verbal or numeric), the specificity of the anchors, and the number of scale points, to name a few. The following text identifies several scales likely to provide the types of data useful for developing test plans for credentialing examinations. The presentation is not exhaustive; for examples of additional scales, see Gael (1983), Harvey (1991), Knapp and Knapp (1995), and Raymond (2001, 2002a,b).

● Task responsibility: Almost any job analysis needs to determine whether the respondent is personally responsible for performing each task. Although a simple dichotomous scale (yes, no) can be used, it is often convenient to incorporate task responsibility into other scales. In Figure 1, for example, the lowest point of the frequency scale is "not responsible for this task."

● Need at entry: This type of scale is intended to determine the extent to which each task is required of entry-level practitioners. There are many variations of this scale. Some address simple task responsibility (yes, no), while others get at attributes like level of competence required or when mastery is expected (e.g., competence not required at entry; task should be learned within first 6 months). An apparent advantage of the need-at-entry scale is that it can be completed by anyone familiar with the job—it merely asks the respondent to make a judgment. However, a more direct way to obtain information about entry-level practice is to develop a standard questionnaire using any of the other scales described here, and then being sure to include entry-level practitioners in the sample.

● Level of responsibility: A scale that measures the level of responsibility can be useful for determining the depth of knowledge at which certain skills should be assessed, with higher levels of responsibility implying the need for deeper understanding. These scales might consist of response categories such as (1) *assist with*, (2) *perform under direct supervision*, and (3) *independently perform*, and respondents would be instructed to check just one response. This type of scale assumes that response categories can be ordered from lowest to highest degree of responsibility.

● Type of responsibility: Type of responsibility scales generally allow for multiple responses. Consider a task inventory questionnaire that lists several quality control (QC) procedures. The response categories might include four options such as *recognize when to perform the QC test*, *perform the QC test*, *interpret results of the QC test*, and *take corrective action based on QC results*. When respondents can legitimately check more than one category, the end result will be a set of nominal scales with dichotomous values. Interpreting these scales can be challenging. For example, 50 tasks rated on a type of responsibility scale that consists of five response categories results in 250 dichotomous variables.

● Time spent: A common method for determining the extent of time individuals spend performing work activities is to ask them to rate the time spent on each activity compared to all others (Knapp & Knapp, 1995). The response categories look something like *much less time than on other tasks*, and so on. Although appealing, these relative time spent scales must be challenging for respondents given that they are required to first recall how often they perform a task and then compare it to some agglomeration of time spent on all other tasks. Another way to approach time spent is to have respondents estimate the percentage of time spent, making sure that the percentages sum to 100%. This type of question is effective only if the list is kept to a manageable number of activities, perhaps fewer than 10 or 15. It is particularly useful for a limited number of general job responsibilities (e.g., administration, teaching, research), as opposed to a long list of specific activities.

● Task frequency: Task frequency is one of the most common scales used in job analysis questionnaires (Newman et al., 1999). The utility of this scale derives from the notion that a credentialing exam should give greater emphasis to activities performed more often. Figures 2A and 2B present two types of frequency scales. The first is a relative frequency scale, while the second measures absolute frequency.

● Task complexity or difficulty: Scales such as these might be helpful in identifying skills that should be included on a credentialing exam because they are particularly difficult to perform or master. Although these scales are common in

A: Relative Frequency

Please use the scale below to indicate how often you personally perform each activity on the following pages. If you are not responsible for an activity, simply check NR and proceed to the next one. Check only one box for each activity.

0 = never perform
1 = seldom perform
2 = occasionally perform
3 = perform fairly often
4 = perform very often

B: Absolute Frequency

Please use the scale below to indicate how often you personally perform each activity on the following pages. If you are not responsible for an activity, simply check NR and proceed to the next one. Check only one box for each activity.

NR = not responsible for
Y = about once per year or less
M = about once per month
W = about once per week
D = about once per day
SD = several times per day

FIGURE 2. Two types of frequency scales. Absolute Frequency (B) is preferred. The response categories can be modified in various ways (e.g., exclude several times per day) to suit the purpose of the study.

business and industry, they do not yet enjoy widespread use in credentialing. See Fleishman and Quaintance (1984) for examples of task complexity scales.

• Criticality, consequences, and related: These types of scales get at the importance of a task by asking how essential it is to successful job performance, or by asking what the consequences would be if the task is performed poorly. The rationale for these scales is that credentialing exams should address those skills most crucial to public protection even if those activities are rarely performed (Kane, 1982). Respondents should be given a clear definition of the scale, and anchors must be consistent with that definition. Criticality is often defined in terms of the *risk* or likelihood of a negative consequence, or as the *severity* of the negative consequence (e.g., no harm, minor injury, extensive injury).

• Overall importance: Scales that measure overall task importance are very similar to the criticality scales just presented—only they are broader and often stand alone as a single scale. The *Standards* note that the content of credentialing exams should be defined and justified in terms of the importance of that content for effective practice (AERA et al., 1999, p. 161). Thus, it is not too

surprising that scales for overall importance are very popular in practice analysis (Newman et al., 1999). The problem is that overall importance is complex, multidimensional, and often subjective (Harvey, 1991; Raymond, 2001; Sanchez & Levine, 1989). One individual may judge a task as important if performed on a daily basis, while another may judge a task as important if it figures prominently in a supervisor's evaluation. Figure 3 contrasts a scale for overall importance (Figure 3A) with two scales that define importance in terms of consequences of poor performance (Figures 3B and 3C).

If the goal is to obtain a measure of overall importance, I recommend that two or more unidimensional scales be statistically combined into an overall composite of task. For example, a criticality scale (Figure 3B or 3C), when combined with a frequency scale, would provide a very meaningful estimate of overall importance. Sanchez & Levine (1989) found that indices task importance derived from linear combinations of two other scales were generally more reliable than holistic judgments of task importance made on a single scale. Methods for combining rating scales are briefly discussed toward the end of this module.

Rating Scale Design

When designing rating scales, it is first important to consider what information the practice analysis study requires. If the goal of the study is to determine which activities are most critical, then scales can be constructed that measure one or more aspects of task criticality. If the study needs to get at task frequency or time spent in addition to criticality, then a second scale may be needed. Some studies may have multiple purposes, say, to develop both curricula and test plans. In such instances multiple scales, or even multiple questionnaires, may be required.

Second, consider the possible sources of information (i.e., the sample of respondents) before deciding which rating scales will be used. Some scales—such as relative time spent, criticality, complexity, or overall importance—have high cognitive processing demands if taken seriously by respondents. Cognitively complex judgments are necessary for some practice analysis studies; but collecting these judgments may impact sampling, scale design, and data collection strategies. Although typical practitioners are the best source of information for ratings of task frequency or task difficulty,

A: Criticality (Overall Importance)

Use the scale below to rate how important competent performance of each activity is to the safety and protection of the public.

0 = of no importance
1 = of little importance
2 = moderately important
3 = extremely important

B: Criticality (Risk of Consequences)

Use the scale below to rate the criticality of each task to the well-being of clients, staff, or your employer. If the task were to be performed incorrectly or not at all, what would be the risk of an adverse outcome such as injury or financial loss?

0 = no risk of adverse consequences
1 = slight risk of adverse consequences
2 = moderate risk of adverse consequences
3 = very high risk of adverse consequences

C: Criticality (Severity of Consequences)

Use the scale below to rate the criticality of each task to the well-being of clients, staff, or your employer. If the task were performed poorly or not at all, what would be the consequences in terms of harmful outcomes such as complications, injury or financial loss?

L = little or no harm
M = moderate level of harm
S = serious or severe harm

FIGURE 3. Three scales for measuring criticality/importance. A is the least specific of the three. In addition to notable differences between B and C (risk vs. consequences; number of response options), there are minor differences as well.

judgments about task criticality might better be left to seasoned SMEs (Kane, 1997; Raymond, 2002a). A questionnaire is not always the most effective method for collecting job-related data. It may be more productive to gather complex judgments in the context of a live meeting where SME panels have the benefit of discussing and refining their definitions of complex rating scales such as criticality or importance.

Third, include only those scales that are necessary to accomplish the study's goals. Most job analysis questionnaires are long to begin with; those with multiple, complex scales are tedious to complete and may suppress response rates. If possible, use no more than two rating scales for each task. In particular, the simultaneous use of importance and criticality scales is not cost effective and should be avoided; those scales tend to be highly correlated and provide redundant information (Sanchez & Levine, 1989). If multiple goals dictate more than two or three scales, consider developing two versions of the questionnaire and mailing them to different samples.

Fourth, use clear and concrete verbal descriptors for anchor points on rating scales. Scales with absolute anchors are generally preferable over relative scales. Harvey (1991) spends considerable time discussing the limited statistical properties resulting from relative ratings. They are often fuzzy and are prone to response bias and halo error (Raymond, 2001, 2002b). Finally, the response categories (scale anchors) should be meaningful for the profession being studied. A study of accountants might include yearly and quarterly, as some accounting activities occur at those intervals. If it is not important for a study to differentiate between tasks that are performed daily and multiple times each day, then daily should suffice as the highest response category. Numerical anchors should be used only when a number is required to communicate the intended meaning of the scale. Anchor values such as 0, 1, 2, and 3 are probably acceptable for most ordinal scales, but are seldom appropriate for nominal scales. In many instances, letter codes may provide a more effective mnemonic for respondents (see Figure 2B).

Other Questionnaire Content

It is common to supplement a task inventory with other types of survey items. Questions that ask about respondent demographics, work setting, and types of knowledge used in practice can be helpful for understanding the profession, developing test plans, and establishing other certification requirements.

Demographics and Related Content

Most practice analysis questionnaires include one or more pages devoted to the respondent's education, experience, and work environment. These questions have at least three useful purposes. First, they are essential for describing the sample and comparing it to the population. Second, demographic questions can be used to select subgroups of respondents for detailed analysis. A study of entry-level practice may need to exclude from analysis respondents who work part-time, who have several years of experience, or who have purely administrative positions. Third, demographic questions provide the basis for comparing subgroups of respondents. It may not only prove interesting to compare the practice activities of those who work in different settings or with different types of clients, but such analyses can assist with the interpretation of results. The demographic section of a survey might include questions on practice setting (size and type of facility); population density; regional socioeconomics; employment status; educational preparation (e.g., type of degree, coursework); years of experience; hours worked; types of support personnel or colleagues in the work setting; and time spent in general practice activities or specialties (e.g., research vs. teaching vs. client services). Although it is tempting to ask pages of questions, it is important to limit the questions to those useful for making decisions consistent with the goals of the study.

Practice Context

One criticism of the task inventory is that it produces a fragmented description of practice, typically overlooking the cognitive nature of complex professions (LaDuca, 1994). One way to add depth to that description is to find out more about respondents' practice environment, including important features of the practice setting, the types of clients they see, the issues they address and problems they solve, and the tools they use in their daily work (e.g., instrumentation, tech-

nology, models/theories). In a recent survey of radiographers, my colleagues believed it was not enough to know about the types of X-rays that a radiographer was required to produce, but that we also needed information concerning the patients, their condition, their age, and the types of equipment available to the radiographer. Equipment and technology are especially important in determining the knowledge and skill demands of many professions. We included a section that listed 25 types of radiographic equipment and instrumentation, and asked respondents if each item was available to them, and if so how frequently they used it. Along a similar vein, a recent study of nurse anesthesia practice included not a single "task statement." Instead, the questionnaire had several sections related to the practice setting, patient condition, surgical procedure being performed, and anesthesia agents and technique (McShane & Fagerlund, 2004). For example, the section on surgical procedure listed over 100 procedures (e.g., cervical spine fusion, intracranial decompression), which were rated in terms of frequency and level of expertise required. Results from these types of survey questions answer questions such as: Should the exam include pediatric questions? If so, what ages? How many items should be set in a clinic versus an emergency room? What types of anesthetic agents should receive the most emphasis on the exam? Survey items that go beyond tasks to get at the context of practice can be very helpful in understanding the demands of a profession.

KSAs Required for Practice

Traditional task inventories focus on tasks that are actually performed in the practice setting. In contrast, most credentialing exams assess cognitive knowledge and skills. This means that the test plans often consist not of tasks, but of topics and KSAs—information likely to be absent from a typical task inventory. One way to expedite the process of developing test plans is to include KSAs as part of the practice analysis questionnaires. The list of KSAs for a questionnaire intended for measurement specialists might include topics such as theories of learning, personality assessment, test and item bias, and test score reliability. KSAs can be identified by reviewing curriculum materials, textbooks, and review articles.

A table of contents or an index from a comprehensive introductory textbook is a good starting point. Once a list of KSAs has been generated, it is formatted into a questionnaire complete with rating scales. Many of the scales for rating tasks also apply to KSAs. Scales that get at KSA importance or relevance appear to be the most popular choice.

Including KSAs on a practice analysis questionnaire appears to be straightforward; however, it is deceptive in its simplicity. Both questionnaire content and rating scale design require special attention. KSAs are complex abstractions that are difficult to define in a mail-out questionnaire. What is meant by the KSA *theories of reliability*, which appeared on a survey of psychologists? Does it refer to classical or IRT? One parameter or three? Single-faceted generalizability designs or multi-faceted? Thoughtful questionnaire respondents will surely wonder about such things. Consequently, care must be taken to clarify those KSAs that are broad or otherwise ambiguous. That is the first issue. The second issue is that KSAs, and the scales used for rating them—especially importance scales—are conducive to positive response bias (Landy, 1988; Morgeson & Campion, 1997; Raymond, 2001). A project I managed several years ago included a KSA for "advanced statistics," which was further clarified by providing examples in parentheses. Meanwhile, the task inventory part of the same questionnaire included two tasks corresponding to (1) conducting and (2) interpreting advanced statistics (also followed by examples). Although 90% of respondents indicated they were not responsible for performing either of the tasks, the KSA received a moderate level of endorsement. In fact, 26% of those who indicated that they never performed the tasks assigned ratings of moderately important or essential to the corresponding KSA. This anecdotal finding may very well be indicative of a more general tendency. A recent experiment convincingly demonstrated the presence of positive response bias for KSA statements and for global competency statements (Morgeson, Delaney-Klinger, Mayfield, Ferrara, & Campion, 2004).

If KSAs are included on a questionnaire, it is important that each KSA have the same meaning for all respondents. Providing definitions or

clarifying examples in parentheses can help. When possible, the use of rating scales with concrete behavioral anchors, rather than general abilities and traits, can be beneficial. In addition, it is important to ensure that those included in the sample are qualified to make the types of judgments being sought. Although entry-level practitioners can be relied on to indicate how often they apply a KSA in their work, they may not be the best group to judge the importance of KSAs or the depth of knowledge required in practice. This is another instance where a meeting of SMEs might be a more effective data collection strategy (Kane, 1997; Landy, 1988; Raymond, 2001).

Questionnaire Production and Administration

Finalizing a practice analysis questionnaire and getting it in the hands of respondents require attention to many details. During this period of time, decisions will be made regarding everything from font size to sample size. The following section addresses issues related to questionnaire format, and includes recommendations for paper as well as Internet questionnaires. Also included is discussion on sampling and pilot studies.

Format and Layout

Mail-Out Questionnaires. Practice analysis questionnaires typically consist of anywhere from 75 to 200 job-related phrases or statements, two or three scales for each of these statements, plus an additional 10 to 20 demographic questions. Most questionnaires run 6 to 12 pages in length, and some may be longer. Questionnaires this long and complicated require extra measures to ensure adequate response rates and meaningful data. If there is one rule it is as follows: Keep the questionnaire short and simple. Ask only for information that is needed to accomplish the goals of the project. Dillman (2000), Christian and Dillman (2004), and Bourque and Fielder (2003) provide many specific recommendations regarding questionnaire design. Some of those recommendations are summarized in Figure 4.

Internet-Based Questionnaires. It is becoming increasingly common to

use the Internet for survey delivery (Dillman, 2000; Montgomery & Marhafka, 2001). The Internet provides a cost-effective method for reaching a large number of recipients, and alleviates the need to print surveys, purchase envelopes, assemble mailings, and pay postage fees. Another advantage of Internet delivery is that data entry activities are essentially eliminated, and it is possible to employ automated reporting functions that are updated with each new response record. Internet questionnaires can also be conveniently tailored to each recipient's practice based on that person's responses to earlier questions (i.e., branching questionnaires).

As attractive as Internet questionnaires appear to be, their potential utility should be evaluated for each project. It is first necessary to verify that intended respondents have regular access to the Internet. To the extent that individuals with regular Internet access differ in important ways from those without, sample bias will be a problem. Also, e-mail addresses are unreliable. People change or have multiple addresses, Internet providers merge, and so on. Consequently, it still may be necessary to use conventional mail to establish contact with survey recipients. It is also important to evaluate the questionnaire for compatibility with Internet delivery. Lengthy questionnaires, task inventories with complex scales, and questions that require respondents to leave the computer to obtain answers all present challenges to Internet delivery. Even with these concerns, Internet delivery may be the method of choice. If so, thoughtful design and delivery is required to assure that data quality is not compromised. A general rule of thumb for Internet questionnaires is that they be no less convenient to complete than those printed on paper. Dillman (2000) and Christian and Dillman (2004) propose many principles for designing online questionnaires, while Montgomery and Marhafka (2001) offer guidelines specific to practice analysis. Many of their ideas are summarized in Figure 5. Some of these guidelines may seem counterintuitive at first glance. For example, the last point in Figure 5 suggests that the use of elaborate formatting features such as pop-up windows should be limited. Dillman (2000) and colleagues have conducted extensive empirical research on formatting

FIGURE 4. Ten tips for questionnaire format and administration. See Dillman (2000), Christian and Dillman (2004), and Bourque and Fielder (2003) for additional guidelines.

of both mail-out and Internet-based questionnaires. The use of "fancy" formatting—which can be distracting, may require additional time to download, and may result in snags with certain types of Internet browser—resulted in lower completion rates, longer response times, and lower overall return rates.

Pretesting and Pilot Studies

Pretesting is a way to obtain feedback on the cover letter, instructions, layout, rating scales, sequencing of questions, and questionnaire completeness. Pretesting can be accomplished by mailing the questionnaire to a small group for review and comment, by having conference calls to discuss the questionnaire, or by conducting focus groups (Bourque & Fielder, 2003). Reviewers can be asked to review the questionnaire without specific guidance or they can be prompted to address specific concerns. Pretests are especially important when a questionnaire addresses potentially sensitive topics, includes novel questions or rating scales, or is possibly long, or if there is uncertainty about certain response alternatives. Desimone and LeFloch (2004) advocate the use of cognitive interviews or think aloud protocols as a method of pretesting questionnaires. Pretesting is very common; it is hard to imagine a project that does not make time for it.

The term pilot study is usually used to describe a more extensive evaluation of the questionnaire involving a small sample of respondents (Bourque & Fielder, 2003). The purpose of a pilot study is to test all aspects of the questionnaire and its administration. Besides providing an opportunity to evaluate the questionnaire, the pilot study is useful for testing operational procedures and for estimating response rates. It is especially useful for projects that require complicated sampling plans. Although a pilot study can add a couple of months to survey development time, it is worth the additional effort for novel or intrusive survey questions or where there is some concern about the sampling plan.

Sample Composition

Several references address the topics of identifying, contacting, and weighting samples in survey research (e.g., Bourque & Fielder, 2003; Dillman, 2000; Fink, 2003; Kish, 1965). Therefore, only a couple of obvious but important points are mentioned here. First, samples should be large enough to support the types of analyses and statistical inferences required for a particular project. Although it is common for job analyses sponsored by credentialing agencies to have sample sizes exceeding 1,000 individuals, some studies have demonstrated that for uncomplicated descriptive studies, adequate generalizability can be obtained from 200 to 400 respondents (Kane, Miller, Trine, Becker, & Carson, 1995; Wang, Wiser, & Joseph, 1999). Smaller samples are acceptable for professions that employ fewer people or for job analyses conducted by individual states or local jurisdictions. For projects where more extensive analyses are required (e.g., comparisons of multiple groups, factor analysis), larger sample sizes will be necessary.

Second, it is especially important for practice analysis samples to be representative of the relevant population in terms of practice setting, ethnic background, educational level, gender, and other demographic factors. The reason this is so important is that an individual's demographic characteristics may have a significant bearing on the clients served, problems encountered, and other practice-related activities. If a certain group is undersampled or has a higher rate of nonresponse, then a biased description of practice can result.

Data Analysis

Once questionnaires have been completed and responses recorded in a data

1. Have a plan for communicating with individuals in the sample (i.e., mail and/or regular mail). Recognize that e-mail addresses may not be reliable over time. Give individuals the option to request removal from the sample.

2. Format the questionnaire so that it will appear the same on common browsers and different screen resolutions. Limit line length to eliminate the need for horizontal scrolling.

3. Provide detailed instructions for accessing the questionnaire. Anticipate common problems and describe how to solve them.

4. Start with a welcome screen that functions much like a cover letter (i.e., to motivate). Indicate what action (e.g., click or enter) will move the respondent to the next screen. Give a brief overview that allows respondents to envision the entire survey. If it requires more than a few minutes to complete, then describe procedures for stopping and restarting the questionnaire.

5. Start with an ''easy'' question that is noninvasive and simple to answer.

6. Provide instructions regarding key actions required to respond and navigate the questionnaire (e.g., radio buttons, checkboxes, scroll bar, tab key, return key). Place general directions at the beginning; directions specific to a question or section should be placed where needed.

7. Be careful with column headings that can scroll out of view. This problem is particularly annoying when completing a task inventory, and it is necessary to scroll back to the top to reread the rating scale categories (e.g., monthly, weekly, daily)

8. Allow respondents to react to questions as they could on a paper questionnaire. Respondents can be discouraged from skipping or providing multiple responses, but they should not be forced to respond.

9. To help the respondent navigate at will, utilize continuous scrolling rather than a design that presents a single screen/question. Include a ''progress bar'' to inform respondents of their location in the questionnaire.

10. Avoid excessive graphics, motion, pop-up boxes and color. Browsers and personal computers react differently to such features. Simple designs generally yield better results.

FIGURE 5. Strategies for administration of Internet questionnaires. See Christian and Dillman (2004), Dillman (2000), and Montgomery and Marhafka (2001) for additional details.

file, it is time to get down to doing what most of us were trained for. Given that many questionnaires consist of 100 or more tasks rated on multiple scales, it is not uncommon for data files to consist of 500 or more variables. Data analysis can be quick and to the point, or thorough and time-consuming. Assuming the results will be translated into test plans or other types of documentation (e.g., eligibility requirements, curriculum guides), some type of report will be needed to guide those efforts. This final section of the module provides an overview of the statistical analyses that might be conducted as part of a practice analysis project. Given that most readers are familiar with univariate and multivariate statistics, the discussion is quite general. My intent is to point out the breadth of data analyses that might be useful, to provide a rationale for such, analyses, and to highlight any issues that may be specific to practice analysis.

Routine Screening

Before computing summary statistics, it is important to verify the integrity of the data. Analyses should be undertaken to detect data entry errors and excessive nonresponse. Simple descriptive statistics and graphics are helpful for evaluating responses to questions with interval or ordinal scales. In addition to the usual process of inspecting each variable, it is also informative to compute the summary statistics for individual respondents (e.g., number of missing responses, minimum and maximum values, mean, and variance). Then tables and graphs can be produced to identify respondents with suspicious response patterns (e.g., extreme high or low means; no variance). Colton, Kane, Kingsbury, and Estes (1991) offer several other strategies for evaluating the validity of responses to practice analysis questionnaires. For example, knowing that a subgroup of psychologists are employed in a Veterans' Administration (VA) hospital might lead to certain expectations regarding their practice responsibilities and the types of clients they see. Responses that are grossly inconsistent with expectations might be viewed with suspicion.

Demographics and Response Bias

The demographic characteristics of the sample need to be described for various reasons, one of which is to evaluate the possibility of response bias. This is most feasible when samples are obtained from databases (e.g., membership files) that already contain information such as gender, age, ethnicity, practice setting, and so on. Then, it is quite easy to compare respondent demographics to the demographics based on the original sample. If demographic data are not available for the entire sample, then it is sometimes possible to compare respondent demographics to selected population parameters, if such parameters are available from other sources. Response bias, if detected, should be reported. One imperfect strategy for managing response bias is to weight responses to adjust for nonresponse, by giving greater weight to members of underrepresented groups.

Scale Transformations

The raw data file may consist of codes that do not necessarily represent the values that will be most useful for statistical analysis. Transforming responses to a useful metric may involve converting from alpha codes to numeric codes, or assigning new values to existing numeric codes. When transforming survey data, it is important to remain sensitive to the level of measurement implied by the rating scales. Although it is not necessary to review the different levels of measurement here (e.g., nominal, ordinal, interval, ratio), it may be worthwhile to acknowledge a couple of issues that arise with almost every practice analysis.

Many rating scales, such as the one presented in Figure 3A, result in data with ordinal properties. Although one can be confident that *extremely important* should receive a higher value than

moderately important, it is not possible to ascertain the magnitude of the difference between those two scale points. Similarly, it is not possible to determine if the difference between the ratings of 3 and 4 is the same as the difference between 2 and 3. It is customary in practice analysis to assign values like 0 through 3 to such scales, and then proceed with the usual statistical analyses (e.g., compute the mean importance for each task). This practice is probably acceptable in most instances, but should be carefully evaluated in others. A few issues arise when reporting summary, statistics for ordinal scales. First, statistics like means and standard deviations will not adequately summarize questionnaire responses for ordinal scales that have drastically unequal intervals. Instead of reporting summary statistics such as means and standard deviations, it may be preferable to report the percentage of respondents who selected each response category (e.g., 23% responded very important, 28% responded moderately important, and so on). A second issue concerns the use of inferential statistics like *t*-tests and ANOVAs for the purpose of comparing two or more groups of respondents. Ordinal data may not meet the assumptions required by parametric statistical procedures. There may be merit in using nonparametric procedures or other methods intended for categorical data analysis (e.g., chi-square tests or logistic models).

Another issue has to do with the practice of treating absolute rating scales as ordinal scales. Consider the frequency scale presented in Figure 2B. It is common to transform frequency ratings such as this to a simple ordinal scale where $1 =$ yearly, $2 =$ monthly, and $3 =$ weekly, and so on. However, it could be argued that the responses actually approximate an interval or ratio scale which corresponds to the number of times per week an activity is performed. For example, if a task is never performed, it is assigned a value of 0, while a task performed daily might be given a value of 5 on a times-per-week scale. Similarly, a task performed weekly would be given a value of 1. The challenge occurs when deciding what quantity to assign to *several times per day* or *yearly*. In such instances, SMEs may be able to help settle on a useful approximation. It is apparent that times-per-week is a rather coarse measurement scale; however, I believe it is

an improvement over the distortion introduced by using a simple ordinal scale that runs from 0 to 5 or 1 to 6.

Statistical Analyses of Task Ratings

After recoding and screening the data, it is customary to compute summary statistics for the task statements for the complete sample of respondents. It may be necessary to summarize results for the subgroup that most closely matches those for whom the credentialing exam is intended (e.g., entry-level, full-time employees). It is also informative to compare groups based on demographic variables, particularly those variables related to practice setting or geographic region. For example, results based on the total group may indicate that an activity is performed by a minority of practitioners, suggesting that the activity may not be addressed by a credentialing examination. However, group comparisons could reveal that the activity is performed by most of those who practice in rural settings, thereby suggesting that the activity be covered.

Statistical procedures, such as factor analysis, cluster analysis, and discriminant analysis, can also be helpful in practice analysis. Certain multivariate methods are useful for data reduction—enabling groups to be compared on, say, 15 or 20 factors (or clusters) rather than on the numerous individual tasks that comprise a task inventory. These factors may even provide a theoretically meaningful model for describing professional practice (D'Costa, 1986; Schafer, Raymond, & White, 1992). Multivariate methods are also useful in distinguishing among subspecialties and for identifying the similarities and differences among them (D'Costa, 1986; Raymond & Williams, 2004). There are potential limitations to applying multivariate analyses to task ratings; consequently the results have to be interpreted with care (Cranny & Dougherty, 1988).

Combining Ratings from Different Scales

The main reason for doing a practice analysis is to inform decisions about test plans and other credentialing requirements. The question is, how can the data be used for these purposes. Creating a test plan for a certification exam requires that decisions be

made about the topics to cover on an exam and the emphasis to allocate to each topic. Given that most questionnaires ask respondents to rate each task on two or more rating scales, the project team must determine how these scales will be combined for decision-making purposes. Numerous models, algorithms, and guidelines have been proposed for translating job analysis ratings into weights for test plans. For example, one guideline might be stated as follows: any task performed by at least 60% of the sample will be included in the test plan, with the amount of emphasis directly proportional to that task's rating on the criticality scale. Although this is a reasonable approach, it is more common to combine ratings from different scales into an index of overall importance by using a statistical model. The models can be specified to give more or less emphasis to different scales. For example, in the following additive model, criticality is weighted by a factor of three:

Overall importance

$= \text{Frequency} + (3 * \text{Criticality}).$

Additive models are popular, but certain theoretical and statistical limitations have prompted some researchers to recommend the use of multiplicative models (Kane, Kingsbury, Colton, & Estes, 1989). Although multiplicative models are more complex, requiring that the rating data first be subjected to nonlinear transformations, the added complexity may prove worthwhile. These and other methods for combining ratings from multiple scales are discussed in more detail in other publications (Kane et al., 1989; Raymond & Neustel, in press; Spray & Huang, 2000).

Concluding Comments

The goal of this ITEMS module has been to suggest guidelines for developing and administering practice analysis questionnaires. In many respects, it is very easy to assemble and administer a practice analysis survey, subject the data to various analyses, and summarize the results in a report. But, on the way to producing that report, it is important to examine the measurement procedures and assumptions that give rise to the data. Is a response rate of 65% adequate? What about 50% or 30%? Is the difference between a rating of 4 and

3 the same as the difference between 3 and 2? What does it mean when a topic such as advanced statistics is rated as "very important" by individuals whose jobs do not require them to calculate or interpret such statistics? It is hoped that some of the procedures addressed here will help minimize the ambiguity often encountered in survey research. The art and science of creating a useful questionnaire is in breaking down the job into meaningful units and using clear language to describe job responsibilities. It is also important to give careful consideration to the design of rating scales and response options. Finally, assembling these two pieces—job responsibilities and rating scales—into a user-friendly questionnaire will help maximize response rates and ensure response validity.

Although practice analyses are typically conducted for the purpose of developing test plans, the results can also be used for other endeavors related to the selection, education, and continued development of individuals who work in the occupations and professions. For example, the results can be used for curriculum development (Rosenfeld & Leung, 1999), establishing educational requirements, determining eligibility criteria, and identifying continuing education needs. It is hoped that the use of practice analysis for these other human resource functions will continue to grow.

Posttest for Practice Analysis Module

1. The content of credentialing examinations should be based primarily on:
   A. the objectives covered in major textbooks in the field.
   B. the curricula of large training and education programs.
   C. the knowledge and skills required for effective practice.
   D. the opinions of leaders in the field.
2. Which of the following are criticisms of the task inventory method of practice analysis?
   1. The focus on specific tasks often results in a fragmented description of practice.
   2. The response rates are usually too low to be acceptable.
   3. They can overlook the cognitive nature of complex professions.
   4. They result in nominal-level data that are limited to qualitative analyses.
   A. 1 and 2
   B. 1 and 3
   C. 2 and 4
   D. 3 and 4
3. In addition to the task inventory questionnaire, what other method of practice analysis is common in professional credentialing?
   A. Position Analysis Questionnaire (PAQ)
   B. Functional Job Analysis (FJA)
   C. Job Descriptive Index (JDI)
   D. Critical Incident Technique (CIT)
4. Assume that part of a questionnaire for the job of measurement specialist includes numerous activities pertaining to statistical and psychometric analyses (e.g., *Calibrate items from a test form using the 3-PL model*). Respondents are asked to rate each activity using a "type of responsibility" scale that has five response categories:
   $0 =$ No responsibility for this activity.
   $1 =$ Yes, I specify the computer code (e.g., SAS, BILOG) for performing this procedure.
   $2 =$ Yes, I run the code that performs this procedure.
   $3 =$ Yes, I interpret results of the procedure.
   $4 =$ Yes, I write reports summarizing the results of the procedure.

   Respondents choose either no responsibility or one or more of the yes categories. You have been asked to analyze data from this questionnaire. How should responses to this scale be recoded, if at all, prior to statistical analyses?
   A. Use the scale values specified above.
   B. Reorder values so "run code" receives a 1 and specify codes receives a 2.
   C. Assign new values that account for the nonlinear increase in complexity; something like "run code" $= 1$; "specify code" $= 3$; "evaluate results" $= 6$; and "write reports" $= 10$.
   D. Create four separate nominal scales (one for each yes), where 0 indicates the person does not have that responsibility and 1 indicates that the person has that responsibility.

5. Sanchez and Levine (1989) evaluated methods for determining the overall importance of each task. What was one finding of that study?
   A. Ratings on a single scale of overall task importance were *less* reliable than an index of task importance derived from combining ratings from two simpler scales.
   B. Ratings on a single scale of overall task importance were *more* reliable than an index of task importance derived from combining ratings from two simpler scales.
   C. Overall importance had *high* correlations with time spent, suggesting that the two scales are redundant.
   D. Overall importance had *low* correlations with task criticality, suggesting that the two scales are quite unique.
6. Studies of response bias in practice analysis ratings suggest that:
   A. response bias is not a problem in practice analysis studies.
   B. response bias can be controlled by using a value of 0 at the low end of the rating scale.
   C. KSAs are more likely than task statements to elicit a positive response bias.
   D. a 7-point rating scale has less bias than a 5-point rating scale.
7. Which of the following are advantages of Internet-based questionnaires?
   1. Postage and printing costs are eliminated.
   2. Easier to obtain a representative sample of the population.
   3. No need to scan returned forms or manually enter data.
   A. 1 and 2
   B. 1 and 3
   C. 2 and 3
   D. 1, 2, and 3
8. Cognitive interviews can be especially useful for:
   A. pretesting questions and rating scales on a survey.
   B. conducting follow-up telephone interviews to determine why individuals did not respond.
   C. working with statistical consultants to estimate response rates.
   D. obtaining paired comparison ratings on tasks for purposes of multivariate analyses.
9. The four-point importance scale presented in Figure 3A will result in

a scale corresponding to what level of measurement?

- A. Nominal
- B. Ordinal
- C. Interval
- D. Ratio

10–15. Assume that the following activities will be included on a practice analysis questionnaire for the job of measurement specialist. For each activity statement, comment on at least one editorial problem, and note how each statement might be revised.

10. Conduct practice analysis projects.
11. Compose/edit multiple-choice test questions.
12. Demonstrate leadership skills on a routine basis.
13. Develop scoring rubrics and train graders for an essay examination.
14. Explain test results to teachers and other school staff using appropriate terminology.
15. Apply knowledge of equating when producing scaled scores for a graduation exam.

Answers to Posttest

1 = C. Although test content will come from a variety of sources, the *Standards* and other sources recommend that the content of credentialing exams be traceable to the requirements of practice (i.e., job related). Such requirements are determined through a practice analysis.

2 = B. Response rates are generally good with well-designed surveys, and the data are generally suitable for many types of statistical analyses. However, as LaDuca (1994) pointed out, traditional task inventories focus on discrete tasks, and may not provide a coherent picture of complex professions.

3 = D. The CIT has been used in nursing, orthopedic surgery, and many other professions. The other three methods are common in I/O psychology but have little applicability to credentialing exams.

4 = D. Although B or C are tempting, the scale needs to accommodate the fact that individuals can choose more than one response category (e.g., someone might run the code and interpret results). Option D ef-

fectively manages this without making assumptions about the value of each response category. Other acceptable variations on option D also exist.

5 = A. This finding is consistent with psychometric theory on the reliability of linear composites. However, it was not replicated in a follow-up study. An important principle related to this research is that the reliability of the composite index is, in part, a function of the correlations of the scales used to create the index.

6 = C. Morgeson et al. (2004) found evidence of positive bias in ratings of KSAs and competency statements. Other studies and personal experience suggest that bias can be a problem, not so much for tasks, but for KSAs.

7 = B. Sampling may be even more complicated with Internet questionnaires. One potential problem is that some segments of the population may not have convenient access to the Internet, resulting in a systematic undersampling. The module notes other issues.

8 = A. Cognitive interviews are especially useful for evaluating the functionality of complex rating scales and sensitive questions.

9 = B. Although we can safely assume "extremely important" is greater than "moderately important" and so on, we cannot be certain of the distance between the values. The scale is ordinal, at best. The numerical values assigned to the scale in Figure 3A are pretty much arbitrary; different values could be used.

10–15. A few editorial suggestions are offered below. Other types of changes may be warranted.

10. Practice analysis projects involve many separate activities such as working with committees, producing surveys, analyzing data, and so on. Therefore, the statement is probably too general to be useful on a task inventory. It might be acceptable if the goal was to determine participation in 10 or 15 very general activities (con-

duct practice analyses, design test specifications, develop examinations, conduct statistical analyses, etc.).

11. "Composing" and "editing" are very different activities, so this activity should be split into two statements. Slashes are seldom acceptable in situations for which clarity of meaning is required. Does the slash mean "and," "or," or "and/or."

12. This really is not a specific activity, but is a more general skill or ability. This statement seems to prescribe what should be done rather than asking respondents to describe what they do. Another problem is that the phrase "routine basis" already implies that it is done with some regularity, so it would be difficult to rate on a frequency or time spent scale.

13. The use of "and" is a problem here.

14. The phrase "appropriate terminology" is prescriptive. Responses will be impossible to interpret, especially for those who say they rarely perform this activity or that it is not very important.

15. This is more of a KSA than an actual activity. The statement would probably be more informative written as "Conduct equating studies" or even as "Conduct linear equating studies based on the common item design." Also, is it necessary to specify "graduation exam?"

Note

[1] The term credentialing will be used here to refer to both licensure and certification. The word profession will be used to denote both occupations and professions.

References

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

Bourque, L. B., & Fielder, E. P. (2003). *How to conduct self-administered and mail-out questionnaires* (2nd ed). Newbury Park, CA: Sage.

Christian, L. M., & Dillman, D. A. (2004). The influence of symbolic and graphical language manipulations on answers to self-administered questionnaires: Results from 14 experimental comparisons. *Public Opinion Quarterly, 68,* 57–80.

Colton, D. A., Kane, M. T., Kingsbury, C., & Estes, C. A. (1991). A strategy for examining the validity of job analysis data. *Journal of Educational Measurement, 28,* 283–294.

Cranny, C. J., & Dougherty, M. E. (1988). Importance ratings in job analysis: Note on the misinterpretation of factor analysis. *Journal of Applied Psychology, 73,* 320–322.

D'Costa, A. (1986). The validity of credentialing examinations. *Evaluation and the Health Professions, 9,* 137–169.

Desimone, L. A., & LeFloch, K. C. (2004). Are we asking the right questions? Using cognitive interviews to improve surveys in education research. *Educational Evaluation and Policy Analysis, 26,* 1–22.

Dillman, D. A. (2000). *Mail and Internet surveys: The tailored design method.* New York: John Wiley & Sons.

Fink, A. (2003). *How to sample in questionnaires* (2nd ed). Newbury Park, CA: Sage.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51,* 327–358.

Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks.* New York: Academic Press.

Gael, S. (1983). *Job analysis: A guide to assessing work activities.* San Francisco, CA: Jossey-Bass.

Harvey, R. J. (1991). Job analysis. In M. Dunnette & L. Hough (Eds.), *Handbook of industrial & organizational psychology* (2nd ed). (*Vol. 2,* pp. 71–163). Palo Alto, CA: Consulting Psychologists Press.

Kane, M. T. (1982). The validity of licensure examinations. *American Psychologist, 37,* 911–918.

Kane, M. T. (1997). Model-based job analysis and test specifications. *Applied Measurement in Education, 10,* 5–18.

Kane, M. T., Kingsbury, C., Colton, D., & Estes, C. (1989). Combining data on criticality and frequency in developing plans for licensure and certification examinations. *Journal of Educational Measurement, 26,* 17–27.

Kane, M. T., Miller, T., Trine, M., Becker, C., & Carson, K. (1995). The precision of practice analysis results in the professions. *Evaluation & the Health Professions*, *18,* 29–50.

Kish, L. (1965). *Questionnaire sampling.* New York: John Wiley & Sons.

Knapp, J., & Knapp, L. (1995). Practice analysis: Building the foundation for validity. In J. C. Impara (Ed.), *Licensure testing: Purposes, procedures, and practices* (pp. 93–116). Lincoln, NE: Buros Institute of Mental Measurements.

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation and the Health Professions, 17,* 178–197.

Landy, F. J. (1988). Selection procedure development and usage. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government, Vols. I and II* (pp. 271–287). New York: Wiley & Sons.

McShane, F., & Fagerlund, K. A. (2004). A report on the council on certification of nurse anesthetists 2001 professional practice analysis. *Journal of the American Association of Nurse Anesthetists, 72,* 31–52.

Montgomery, L., & Marhafka, K. (2001). Innovative approach to job analysis: Internet survey delivery. *2001 Conference Handouts, NOCA Educational Conference and Annual Business Meeting.* Washington, DC: National Organization for Competency Assurance.

Morgeson, F. P., & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology, 82,* 627–655.

Morgeson, F. P., Delaney-Klinger, K., Mayfield, M. S., Ferrara, P., & Campion, M. A. (2004). Self-presentation processes in job analysis: A field experiment investigating inflation in abilities, tasks and competencies. *Journal of Applied Psychology, 89,* 674–686.

National Commission for Health Certifying Agencies (1981). *Task force report on education and certification.* Washington, DC: National Commission for Health Certifying Agencies.

Newman, L. S., Slaughter, R. C., & Taranath, S. N. (1999, April). *The selection and use of rating scales in task surveys: A review of current job analysis practice.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Raymond, M. R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14,* 369–415.

Raymond, M. R. (2002a). A practical guide to practice analysis for credentialing examinations. *Educational Measurement: Issues & Practice, 21,* 25–37.

Raymond, M. R. (2002b, April). *The influence of rating scale format on rater errors in job analysis surveys.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Raymond, M. R., & Neustel, S. N. (in press). Determining test content for credentialing examinations. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

Raymond, M. R., & Williams, C. O. (2004). Empirically mapping the subspecialties of cardiovascular-interventional technology. *Journal of Allied Health, 33,* 95–103.

Rosenfeld, M., & Leung, S. W. (1999). *The practice of radiation therapy.* Princeton, NJ: Educational Testing Service.

Sanchez, J. I., & Levine, E. L. (1989). Determining important tasks within jobs: A policy-capturing approach. *Journal of Applied Psychology, 74,* 336–342.

Schafer, L., Raymond, M. R., & White, A. S. (1992). A comparison of two methods for structuring performance domains. *Applied Measurement in Education, 5,* 321–335.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36,* 1138–1146.

Smith, I. L., & Hambleton, R. K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues & Practice, 9*(4), 7–10.

Spray, J. A., & Huang, C. (2000). Obtaining test blueprints weights from job analysis surveys. *Journal of Educational Measurement*, *37,* 187–201.

Wang, N., Schnipke, D., & Witt, E. A. (2005). Use of knowledge, skill and ability statements in developing licensure and certification examinations. *Educational Measurement: Issues and practice*, *24,* 15–22.

Wang, N., Wiser, R. F., & Joseph, M. (1999, April). *Examining the reliability and validity of job analysis survey data.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Annotated Bibliography

1. Dillman, D. A. (2000). *Mail and internet surveys: The tailored design method.* New York: John Wiley & Sons.
   This book offers practical advice on survey development. Dillman's recommendations are based on years of experience conducting survey research in social and political arenas, and on the results of numerous empirical investigations comparing methods for designing questionnaires and administering surveys. Readers are encouraged to read other reports by Dillman and colleagues published since this book was printed.

2. Fink, A. (2003). *The survey kit* (2nd ed.). Newbury Park, CA: Sage.
   This 10-volume kit includes ten monographs written by various authors. The monographs are available individually or as a set, with each offering very practical recommendations regarding some aspects of survey development and administration (e.g., survey design, sampling, data analysis, and so on). The monographs begin with a list of learning objectives and conclude with exercises and an annotated bibliography, so they are very useful for instruction.

3. Fleishman, E. A., & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks.* New York: Academic Press.
   Edwin Fleishman devoted most of his distinguished career to studying methods of job analysis and developing task and ability taxonomies. This 1984 publication is arguably the most comprehensive and scholarly treatment of job analysis methodology in the psychological literature. Although it does not offer a step-by-step guide to developing practice analysis questionnaires, reading it is sure to inspire one to approach job and task analysis with an inquisitive mind and critical eye.

4. Kane, M. T., Kingsbury, C., Colton, D., & Estes, C. (1986). *A study of nursing practice and role delineation and job analysis of entry-level performance of registered*

nurses. Chicago: National Council of State Boards of Nursing.

An effective way to learn about practice analysis is by reading technical reports of previous studies. Many credentialing agencies are pleased to share their practice analysis reports with the public. The report by Kane and colleagues is among the best available. As a bonus, it includes an appendix that describes the derivation of a multiplicative model for combining frequecy and criticality ratings into a single index of overall task importance (a revised version was subsequently published in the *Journal of Educational Measurement*).

5. Raymond, M. R., & Neustel, S. N. (in press). Determining test content for credentialing examination. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah, NJ: Lawrence Erlbaum Associates.

This chapter elaborates on many of the topics addressed in the preceding module. It also presents concrete examples for translating practice analysis results into test specifications, and describes different models for combining ratings from multiple scales into a single index of overall task importance. The chapter includes more than 160 references.