Practice Analysis Studies: A Literature Review of Definitions, Concepts, Features, and Methodologies

(Version: 8-30-2011)

Amanda Corby Soto
University of Massachusetts Amherst

A report prepared for the Medical Council of Canada.

Licensure and Certification Examinations: An Introduction

"The development of licensure and certification examinations requires accurate information about the kinds of problems encountered in practice and about the knowledge, skills, and judgment required to handle these problems" (Kane, 1997, p. 5). Without a clear understanding of the demands of practice, the credentialing agency would have no basis for selecting content for the examinations and no way to gauge the appropriateness of examination content. Information about the habits and demands of practice is necessary whether we are using written tests, simulations, or performance assessments to readiness to enter the profession. A practice analysis study, then needs to provide an overall description of practice *patterns* for the studied profession. The goal is to develop an understanding of the shared general knowledge and skill base across practitioners and settings. "Ideally, practice analysis provides not only an accurate and parsimonious description of this complex situation but also one that can be readily translated into test specifications" (Kane, 1997, p.7).

Beyond building a description of the patterns and routines required for practice, an empirical practice analysis study should provide data about the breadth and depth of knowledge required by practitioners at the specific career level under study; for example the practice analysis to guide a test to enter supervised practice will be different than one to guide certification into an advanced specialty. These practice data will be used to guide the selection of important items (and content) for the test plan (Clauser, Margolis, & Case, 2006). Thus practice analysis data will provide an empirical link between the examination content and the demonstrated demands of independent practice. This link can be called upon to defend the eventual examination content and structure and to establish the cognitive level of the examination.

In the medical field credentialing exams include those for both licensure and certification. The two types of examination are used to delineate differing levels of knowledge and skill within the

field; "licensure differs from certification in that licensure is a legal requirement for practice…by contrast, certification denotes a status that exceeds the minimum legal requirements for practice" (Clauser, Margolis, & Case, 2006, p.701). This distinction is muddied in practice as in some cases certification may become a *de facto* license; for example in the case of a hospital requiring specialty certification in order to grant privileges the certification becomes the minimum standard for practice instead of the license. On licensing examinations, in the interest of protecting the public, "minimum standards [are] set as the requirement whereas certification serves to identify a professional group with rigorous entry requirements" (D'Costa, 1986, p. 138).

Report: Purpose and Design

This report will outline practice analysis methodologies preferred for use by credentialing agencies as well as the theoretical background for using a practice analysis study for the design or revision of a test plan. The first section will outline the background of practice analysis, as a practice, and the use of practice analysis in credentialing procedures. The second section will transition to a discussion of practice analysis data as a type of validity evidence for content validity of a credentialing examination and the various methodologies that are currently preferred in licensure and certification; this also includes commentary on applied concerns of practice analysis (managing the process) and managing the data generated from a practice analysis study. The third section addresses how to implement the various practice analysis methodologies; including building and managing a task inventory (or 'practice') questionnaire from tasks and knowledge, skill, and ability (KSA) statements. And the final section will address the use of practice analysis methodologies for the revision of the Medical Council of Canada Qualifying Exam (part 1) blueprint.

This report will focus on licensure procedures, or ensuring that passing individuals have demonstrated the knowledge and skills necessary for safe and effective practice. Building a practice

analysis study to assess the minimum competencies required for independent practice is a different endeavor than assessing more expert knowledge of a specialty. This distinction is not drawn directly in the majority of the practice analysis literature, however. Throughout this report the terms credentialing, licensure, and certification are used interchangeably following the original author's use, though this paper speaks to the construction of a practice analysis study for licensure purposes.

Following Kane (1997) and Knapp and Knapp (1995), the term "practice analysis" will be used throughout this report instead of the broader term job analysis, or the more specific role analysis, role delineation study, task analysis, or process analysis. The term practice analysis can be used more broadly to reflect the multi-faceted demands of professional practice as opposed to the task focus of a traditional role analysis study (Smith & Hambleton, 1990). This difference has grown as task analysis/role delineation studies and practice analyses have diverged in purpose. Traditional job analysis, with the goal of assessing responsibilities and duties within a job, is distinct from a credentialing-focused practice analysis.

Practice Analysis: Purpose & Background

The Standards (AERA, et al., 1999) suggest that the primary obligation of licensure examinations is to protect the public; credentialing serves to "protect the profession by excluding persons who are deemed unqualified to do the work of the occupation. The tests used in credentialing are designed to determine whether the essential knowledge and skills of a specified domain have been mastered by the candidate" (p. 156). An empirical practice analysis has been argued to be the best way to establish a tie between the tasks and demands of professional practice and the credentialing examination (AERA et al., 1999). Practice analysis data are the preferred basis for developing certification examination content (Kane, 1982, 1997; McGaghie, 1991) and are considered the primary basis for assessing the validity of licensure examinations (AERA et al., 1999).

The domain of practice, particularly in medicine, is extensive and complex; professionals may practice in a very large number of settings and face diverse client needs and presentations. Each setting (and population) will pose a unique challenge. The practice analysis study serves as the unifying element across professional settings by representing the common demands and procedures (and required skills) across settings. This might include a decision-making procedure leading to the selection of a treatment, or a shared set of steps that are repeated across clients leading to diagnosis. The shared demands across each setting and client population should be represented on the eventual test blueprint.

As mentioned above, licensure exams are used to ensure that individuals who enter a profession have met the necessary standards and are qualified to begin practice in a safe and effective manner (AERA et al., 1999; Clauser, Margolis, & Case, 2006; Shimberg, 1981; Smith & Hambleton, 1990). Beyond the philosophical requirements that certification tests accurately sort examinees, in practice this requires empirical justification of the examination scores and the subsequent classification decisions (Messick, 1989). Practice analysis data serve as the link between practice and test content (D'Costa, 1986; Levine et al., 1988), forming the foundation of an argument in support of the score inferences.

Practice Analysis: Role in Building Validity Evidence For Score Inferences

There are differing views in the literature about the usefulness of a practice analysis study as a tool for gathering validity evidence. The overall consensus is that conducting a practice analysis is important for justifying the eventual test content, for the defensibility of the examination building process (Knapp & Knapp, 1995), and for making defensible classification decisions based on examination scores (D'Costa, 1986).

As Hecht (1979) suggests, "…predictive criterion-related validation studies would be the type most closely fitting the expressed purpose of licensure exams: that of assuring minimal competency

on the job for the protection of the public. Interest is with the criterion not yet obtainable at the time of testing" (p. 21). Given the challenges of demonstrating that individuals that pass the exam are more prepared for practice than those who do not, it is necessary to build an argument in support of using the scores for classification decisions. Kane suggests that scores be interpreted "as…evidence of an examinee's present competence on specific abilities that are considered critical for practice" (Kane, 1982, p.911 emphasis added) instead of using them to predict successful practice. These "critical abilities" should be those known to affect client outcomes (following the mandate to protect the public) and should be drawn from the practice analysis study.

> "The critical abilities for a profession include cognitive abilities involving knowledge and the ability to apply knowledge, as well as psychomotor skills involving the ability to do various things for clients. The abilities may be quite general (e.g., communication skills) or quite specific (e.g., the ability to carry out a particular diagnostic procedure). It is not assumed that any specific set of critical abilities would guarantee success in professional practice, but it is assumed that the absence of these abilities in professional practitioners would pose a definite risk to the clients of these practitioners" (Kane, 1982, p. 914).

Given the known challenges to building predictive validity evidence (because those that do not pass the examination do not enter practice, preventing comparisons), it cannot be known whether the examination can predict performance. However, building an examination based on the abilities critical for practice allows for a reasonable standard for interpreting credentialing examination scores. "If the examination ensures that professional practitioners have these abilities and if the lack of these abilities would lead to poor practice, the examination provides an effective safeguard against some causes of poor practice" (Kane, 1982, p. 918).

The link between a test's content and the eventual demands on professionals relies on a practice analysis study that accurately reflects practice and is accurately translated into a test plan. The test plan will shape the work of the item writers, test preparation programs, and study choices of future examinees (Raymond, 1996). If this link is strong, and content on the test plan reflects practice, inferences from examination scores will be defensible in court – a particular concern for

credentialing bodies (AERA, et al., 1999). This report will continue by discussing the methodological approaches to practice analysis that are preferred for medical credentialing. Before considering what form a practice analysis study should take, however, a plan for how it will be managed should be implemented.

Practice Analysis: Managing the Process

Raymond (2005) suggests that a statement of purpose be crafted to guide the study design and data collection; this statement can originate from the credentialing agency or subject matter experts in the field. This is important since practice analysis data can be collected for a range of possible purposes, even within a field or testing program. "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" (Raymond, 2005, p. 30). The purpose statement should reflect the requirement that new practitioners are prepared for independent practice and "often takes the following form: *This test is designed to measure whether or not an examinee has the requisite knowledge and skills that are required of new practitioners"* (Clauser, Margolis, & Case, 2006, p. 704 emphasis in original). The statement should also reflect the meaning of scores, if they are provided (perhaps they measure the *extent* to which the examinee possesses the knowledge and skills). The important distinction here is that the skills being measured are those required for practice (e.g. those required to protect the public) not necessarily those learned in school.

Given the different components in a practice analysis study (design, building a model for practice, recruiting a respondent sample, data preparation, analysis, and implementation) the process should be led by an individual and/or a committee and guided by a timeline of tasks and deadlines that delineates who will participate in each step of the process. The entire process can take between

6 and 18 months with the possible addition of 3 to 9 months should be allowed if the data are going to be used to create a new test plan (Bourque & Fielder, 2003).

Practice Analysis Methods

Practice analysis methods vary across the types of practice related information they collect, their orientation towards tasks or human factors, and the level of specificity in the included questionnaire (Raymond, 2001). There is not a clear consensus in the research community that one job analysis method is preferred over another (Levine et al., 1980) though several authors have suggested tools for comparing methodologies (Brumback, Romashko, Hahn, & Fleishman, 1974). Other authors have argued that regulatory programs and credentialing examinations are more politically influenced than they should be, with undue bias introduced by subject matter experts, practice analysis consultants, and a misplaced emphasis on job relevance instead of protecting the public (Nelson, 1994).

These are legitimate concerns, yet several authors (Kane, Kingsbury, Colton, & Estes, 1989; Sanchez & Levine, 1989) have focused their work on practice analysis for credentialing instead of job analysis to explore the unique issues related to building a study that is *both* job related and attuned to the credentialing mandate to protect the public.

The research community has not come together to conclusively recommend one methodological approach over another, yet the clear indication is that the methods selected should be linked to the eventual use of practice analysis data. Levine et al. (1980) found "few substantial effects" on "exam plan contents, exam plan quality ratings, or costs encountered in developing exam plans from the reports" (p. 534) across the critical incident technique, position analysis questionnaire, task analysis and the job elements approach. The practice analysis study, as discussed

above, should be described in by study's statement of purpose. This statement, along with the intended use of the data, should guide the decision between methods.

The selected methodology is not what assures the validity and defensibility of the test plan, instead, it is the execution and documentation of the process, whichever one is chosen, that will provide the foundation for defending potential legal challenges (AERA et al., 1999). Knapp & Knapp (1995) suggest that for licensure, a practice analysis methodology should be 1) relatively easy to conduct, 2) cost effective, 3) easily replicated to reflect changing professional knowledge and competency requirements, and 4) useful for obtaining 'buy in' from stakeholders in the licensing process. They suggest that the methods used for credentialing need to be tailored to the needs of the licensing board and profession; this may mean including different combinations of SMEs, practitioners, and educators on the steering committee or using a pre-existing data collection instrument from a similar profession instead of creating one.

Practice Analysis: In Credentialing Versus Industry

At the most pragmatic level, practice analysis studies used in credentialing often require a larger and more diverse sample of respondents than those used in industry. This is largely due to the fact that professionals practice in diverse settings and with diverse groups of clients, breadth that must be represented by the study data. The other differences between practice analyses in credentialing and industry are more directly related to methodological decisions. For example, the purpose of a credentialing practice analysis may be to develop a test plan for a knowledge-intensive profession as opposed to a test plan for personnel selection or performance evaluation. The content also will differ as a measurement consultant may be familiar with the content of an industrial practice analysis and will be able to design the study without the input of SMEs unlike in the majority of credentialing fields where the study design and content development will occur collaboratively between SMEs or others familiar with the field and measurement experts.

Practice analysis methodologies that focus on human abilities and personality traits as well as those designed for human resource activities (evaluation, personnel selection, etc.) are preferred in industry but are not often used in professional credentialing (Raymond & Neustel, 2006). It is argued that the social nature of professional life should distinguish professional practice analyses from those conducted in industry. LaDuca (1994) identifies the professions as both dynamic and social, requiring sound judgment and strong practical knowledge. Following Brain (1991), professional work is distinguished from other work by its collective orientation – as professionals practice and interact with the governing body, the notion of acceptable practice (and sanctioned practice decisions and actions) is developed. "This is the core of the profession's practical knowledge, which constitutes their capacity to identify a job to be done, to know how to go about doing it, and to recognize when it has been done appropriately" (Brain, 1991, p. 260).

A practice analysis in a profession, especially when intended to provide the basis for a credentialing exam, should be designed to assess this type of knowledge as opposed to a practice analysis study in industry, which may be used (and designed) to classify jobs within an organization or for employee evaluation. This speaks to the different applications of practice analysis data in the two domains as well as the different nature of professional and technical jobs. Credentialing examinations also assess an individual's knowledge and skill unlike a job re-design study that would focus on his or her traits and attitudes (Raymond, 2001).

Practice analyses rely on information from individuals and sources that are knowledgeable about the profession under study (supervisors and trainers, educators, practitioners, and possibly researchers). However, it is not necessary to limit data collection to human sources as documentation from records and research studies can provide valuable information. Depending on the methodology chosen, sources like patient charts, insurance records, and textbooks can provide valuable information for a practice analysis. Particularly charts and insurance records can provide

insight into the diagnostic and treatment approaches favored by physicians. Data collection methods may also include observing and interviewing experts, asking professionals to complete a work diary or time-log, or conducting focus groups and meetings with SMEs.

Practice Analysis: Credentialing and Methodological Orientation

Practice analysis methodologies vary according to how they characterize practice (Fine, 1986; Fleishman & Quaintance, 1984; Raymond, 2001) in task statements/KSAs. *Task-oriented* methodologies describe professional practice more directly by examining problem-solving activities and the related outcomes. *Person-oriented* approaches focus "on the human attributes required to effectively carry out the tasks required of a job" (Raymond, 2001, p. 373). This includes KSAs or an individual's characteristics that contribute to successful practice. Deciding the study's orientation is an important step in design; the orientation will influence the choice of rating scales, the participant sample selection, and the translation of the practice analysis data to the test plan (Raymond, 2001).

Two methods that have been especially effective and popular for practice analyses in credentialing are the critical incident technique (Flanagan, 1954) and the professional practice model (LaDuca, 1994), however certification boards have also commonly employed the task inventory questionnaire and its variations (Newman, Slaughter, & Taranath, 1999). Downing and Haladyna (2006) suggest that the critical incident technique and the task inventory are both common and appropriate for credentialing. Raymond (2002) and Raymond and Neustel (2006) identify the three methods above as preferred for credentialing purposes because of their structure and flexibility. These characteristics are preferable for a credentialing practice analysis because they allow for data collection across professional settings and roles and are not limited to analyzing the demands of a single job in complement to other jobs in the same organization.

The **Critical Incident Technique** or CIT was introduced by Flanagan (1954) and has been used extensively for military purposes but has also been used in the health professions; in general

medicine (Hubbard, Levit, Schumacher, & Schnabel, 1965), medical education (Edwards, Currie, Wade, & Kaminski, 1993), nursing (Jacobs, Fivars, Edwards, & Fitzpatrick, 1978), anesthesiology (Altmaier et al., 1997) and orthopedic surgery (Levine, McGuire, Miller, & Larson, 1968). This approach entails gathering anecdotes that document incidents that were 'critical' in that they had a strong positive or negative effect on client (patient) outcomes. These anecdotes should share structural elements, like the practice context, the effective or ineffective behavior, and the consequence of those behaviors for clients. They are typically collected through a questionnaire but focus groups and interviews have also been used. Subject matter experts (SMEs) then summarize and categorize the anecdotes into more general performance categories that might serve as an outline for test plans or curricula. By its nature, this type of data limits the sampled practice domain to the "most critical" incidents or those that occur most frequently or are most important for client outcomes. This approach then limits the scope of practice that can then be represented on a credentialing exam.

The test plan drawn from a CIT study may represent the practice analysis findings by representing the demands of practice related to important client outcomes, but may exclude elements of practice that are crucial in protecting the public but are less common or more routine for practitioners. Given the specific situational focus of the CIT, the incidents described in the practice analysis may also be used to guide item writing (Levine et al., 1968) and be broken down into prompts for a task inventory (Cascio, 1982). The CIT has been found to provide better data than the functional job analysis or position analysis questionnaire, yet is still highly labor intensive (and costly). Despite the labor and cost, the data may not provide for a better test plan (Levine et al., 1980) or a strong representation of professional practice (Knapp & Knapp, 1995). For a detailed description of the critical incident technique (as applied to dentistry) see Fitzgerald et al. (2007).

The **Professional Practice Model** (LaDuca, 1994) perhaps should not be regarded as a specific method of practice analysis but instead as a conceptual framework that can be used to organize the data from another practice analysis method (Raymond, 2002). This approach has been used primarily in the health professions (McGaghie, 1980). Based on professions theory and facet design, the PPM begins with a theory of the target profession to guide answers to the following three questions: "What is the character of the situational domain (i.e. ecology) of the professional who is seeking the license? What is the nature of this professional's cognition? What is the nature of the relationship between the professional's cognition and the ecology in which this person functions competently?" (LaDuca, 1994, p. 180). Raymond (2001) describes a potential model for physician practice as containing two facets, or dimensions, care setting (hospital setting or private practice) and medical problem (perhaps organized by organ system) with the possible inclusion of other dimensions including patient age, treatment approach, or severity of the medical problem. "The cells produced by crossing the facets give rise to numerous practice-related situations…the PPM uses very global practice descriptors (i.e., each facet and its levels) to ultimately produce very specific descriptions of practice problems" (Raymond, 2001, p. 375).

Based on the argument that task analysis is highly inappropriate for complex professions like status or law (LaDuca, 1980, 1994), the PPM attempts to tap into both the professional's knowledge and relevant modes of thinking. One of the most important products from a practice analysis with the PPM is how the facets provide scripts for the KSAs of the specified practice situations (LaDuca, 1994). This allows the PPM to provide both task- and person-based practice descriptors by using information from physician charts (to assess the medical techniques employed) and interviews (to assess the person-related side of practice). LaDuca, Downing, and Henzel (1995) suggest that one of the advantages of the PPM is the ease with which the practice model can be used to create test

items and performance-based examinations. This model is similar to Kane's 1997 Model-Based Practice Analysis.

The **Task Inventory** and **Comprehensive Practice Analysis** appear to be the most common methods of practice analysis (Newman, Slaughter, & Taranath, 1999) and is one of the more efficient and cost-effective methods available. This method is based on the construction of a task inventory questionnaire that is then mailed to a large sample of respondents to judge whether the included tasks are essential to safe performance in the profession. This method appears to be particularly important and preferred in credentialing, as these examinations are intended to gauge an individual's readiness for a wide range of activities in many different settings (Kane, 1982) and establish a candidate's ability to demonstrate the necessary knowledge and skills regardless of his or her future practice setting. A task inventory was employed in the most recent practice analysis of physician assistants (Arbet, Lathrop, & Hooker, 2009).

Though this method provides a large amount of data that are conducive to statistical analysis (and thus to being used to establish weights for test plans and developing models of practice), the focus on discrete tasks may be seen as limiting when a key factor in professional success lies in decision-making and cognitive skill. Task inventories have the potential to overlook the complex nature of many professions because of their reliance on discrete tasks or functions (LaDuca, 1994). This is partially addressed by the comprehensive practice analysis questionnaire, a variation on the task inventory that more directly links the tasks to a practice model.

A Comprehensive Practice Analysis involves multiple approaches and data collection sources to build an understanding of practice. For example the study may build a list of professional tasks from current practitioners using the CIT and then use SMEs to design a list of KSAs and/or task ratings. This approach combines other methodologies to build an understanding of practice context as well as the habits of practice; in such it shares characteristics with the PPM (LaDuca, 1994) and

the CIT. This approach is described as a 'descendent' to the task inventory (Raymond, 2001), but because the tasks in the inventory questionnaire are drawn from a model and are more clearly linked to the demands of practice it could more accurately be described as a combined approach that links a model to the task inventory or vice versa.

This approach has been used in practice analysis studies for the licensed psychologists (Rosenfeld et al., 1983) and the nurse anesthetists (Zaglaniczny, 1993). D'Costa (1986) also employed a comprehensive approach in a practice analysis study using elements of the functional job analysis and the critical incident technique to build a multi-faceted practice model. The distinguishing feature between the "model-driven questionnaires from traditional task inventories is their use of behavioral descriptors that emphasize types," characteristics, and solutions to practice problems (Raymond, 2001, p. 378). Raymond (2002) recommends a comprehensive practice analysis, or an approach that combines multiple methodologies, in order to capitalize on the strengths of each method.

Preferred approaches for credentialing: in practice

> "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 mission they should cover the KSAs required for effective practice" (Raymond, 2005, p. 29).

As discussed previously, in order to accurately assess readiness for practice a licensure examination's content should be rooted in the knowledge and skill demands placed on practitioners. The Standards echo this, suggesting that a licensure examination should be built on a model for practice and be "job related" (AERA, et al., 1999; D'Costa, 1985; Kane, 1982). Kane (1997) and LaDuca (1984, 1994) both advocate building the "job relatedness" of a practice analysis by developing a model of practice, evident in the Professional Practice Model and the Model-Based

approach to practice analysis. Both of these approaches begin with the construction of a model for practice by SMEs, researchers, or the credentialing body. In medicine, several databases have been developed to organize data about clinical practice across specialties (general practice, general internal medicine, general surgery, general pediatrics, family medicine, emergency medicine, and general obstetrics and gynecology). These databases (specifically the National Hospital Discharge Survey, the Ambulatory Medical Care Survey, and patient-encounter self report, among others) might be used as initial data in building a practice model.

The practice model is used to conceptualize the relationship between practice and ecology (as in LaDuca, 1994) or the intersections of cognitive demands, practice setting, and client presentation (as in Kane, 1997) or a different, profession-specific framework that ties practice demands to the required knowledge, skills, and abilities for effective performance. The model-based approach can include any number of dimensions; for example a practice analysis for the nurse anesthetists (McShane & Fagerlund, 2004) included facets specific to nurse anesthetists. The 2001 study was based on a practice model including patient condition, nursing procedures, equipment and instrumentation, and the relative knowledge necessary for performance. Overall, this approach "characterizes practice by the types of problems that professionals are called on to solve and the contexts in which those problems occur" (Raymond & Neustel, 2006 p. 191). The professional practice model has been used for practice analysis studies in medicine, occupational therapy, and medical dietetics (LaDuca, 1980, 1984; LaDuca et al., 1978) and the USMLE Step 3 exam (LaDuca, 1994).

Like the PPM, the task inventory can be used in conjunction with the CIT. SMEs can be gathered to discuss the incidents and break them down into component tasks. This process incorporates the anecdotes to determine the domain of practice, and then the rating process of the task inventory method. The strength of this approach would be a test plan with an emphasis on the

tasks deemed most relevant for client outcomes (based on the critical incidents). The process itself would require a large investment on the part of SMEs and the credentialing body as it would require two large-scale data collections (first for the critical incidents and second for the task inventory ratings). Theoretically it seems possible to use the CIT to build a practice model, though it does not seem to be the most efficient method.

Newman et al. (1999) found that 27 out of 30 practice analyses incorporated a large-scale questionnaire or survey as a data collection instrument. The authors report that most certification boards rely on the task inventory (or one of its variations) when performing a practice analysis. A group of SMEs may generate the list of tasks (based on a practice model, focus groups, or a pre-existing collection of competencies), responsibilities, and cognitive patterns thought to be relevant to practice or they may be solicited from current practitioners, supervisors, or educators in the field with a survey instrument. The task inventory may be used in several ways; first as the foundation for a survey or questionnaire and second as the foundation for a practice model. If it is being used as the foundation for a survey, the task statements will be reviewed and formatted into a questionnaire and mailed to a representative sample of professionals (for example a group of physicians in their first three years of practice after licensure). These individuals then are invited to rate each task on how often it is performed, how difficult it is, and/or how important it is for successful practice. If the task inventory is to serve as the foundation for the practice model and eventual test blueprint the tasks will be organized into categories to build a practice model (and test outline). Since the tasks are known to be relevant to practice, the task inventory may be used to build categories for the test blueprint and to guide the eventual weighting of items across categories.

The task inventory / practice analysis questionnaire is preferred for credentialing for two main reasons. First, the questionnaire serves as a simple way to collect data from professionals

across myriad practice settings. And secondly, the task inventory allows for efficiency and more objective (statistical) methods for data analysis.

Task Inventory: Efficiency in Data Collection

- Task analysis ratings are easy for respondents and are not overly time consuming (depending on the number of tasks included in the inventory).
- Given the ease of disseminating the task inventory through a professional organization or member list, the burden on the credentialing board to include a large and diverse sample is relatively low.
- The returned task inventories are also useful for comparing the requirements and patterns in practice across settings (Cascio, 1982; Raymond, 2001). "The breadth of coverage provided by a task inventory questionnaire is especially important for credentialing examinations, which are intended to indicate an individual's readiness for a wide range of activities in a variety of settings" (Raymond, 2001, p. 374; Kane, 1982).

Task Inventory: Efficiency in Data Analysis

- The task inventory method provides data that may be used to create a practice model or differentiate between practice habits in different specialties or settings (Raymond, 2001).
- The rating scale data can be combined and weighted by various empirical methods to build and weight categories on a test plan (Kane, 1997; Raymond, 1996).
- The task inventory provides clear evidence for the job-relatedness of the test plan (Thompson & Thompson, 1982), which is useful in defending the examination content in legal challenges.

Raymond and Neustel (2006) provide guidelines for structuring a task analysis questionnaire (p.184-185) with the caution that "the purpose of practice analysis is to describe practice, not prescribe it. Using a task inventory questionnaire as a vehicle for communicating effective or appropriate practice, even though well intentioned, produces ambiguous data" (p. 185). The limitations of the task inventory include the possibility that the respondents will misinterpret the survey items – no matter how carefully they are written and reviewed. There is also the concern, prevalent throughout the test design process that some types of information do not translate well to the selected format. Particularly when gathering data for credentialing in the health professions, where the intersection of cognitive skill and professional judgment is particularly important, not all of the dimensions of practice can be represented in a task inventory. As Clauser, Margolis, & Case

(2006) argue, "even after a list of topics or tasks is defined and the data are collected; judgment is required to turn the data into something useful" (p. 705).

Though the limitations are well known, as LaDuca (1994) argues "task analysis is highly inappropriate with complex, high status professions such as medicine or law" (p. 183), the task analysis procedure can provide detailed data about professional practice depending on how the included tasks are created. If based on a practice model or even a CIT study the task inventory provides an additional piece of evidence linking the test plan to professional practice. The advantages of this general approach can be "distilled into three fundamental benefits: numerous practitioners and positions can be sampled, multiple types of descriptors can be used, and the data lend themselves to multiple methods of analysis and interpretation" (Raymond, 2001, p. 379).

D'Costa (1986) has employed the functional job analysis method for practice analysis in credentialing, though it does not appear to be very common given its emphasis on job requirements over cognitive processes (Raymond, 2001). Functional job analysis does provide a strong foundation for the writing of task statements, however, with each statement following a very rigid format including the action performed (the task), the tools or resources used, and whether the task is performed autonomously (Fine, 1986). The tasks are then rated by SMEs in terms of the worker functions required based on "the extent to which each task involves working with people, data, and things, and …language, math, and reasoning skills as well as the physical strength required" (Raymond, 2001, p. 377). D'Costa's (1986) approach combined the functional job analysis method with a comprehensive practice analysis questionnaire to provide additional depth.

Practice Analysis : Data Collection

Landy & Vasey (1991) suggest that in business and industry a job analyst would conduct the practice analysis study. Historically in credentialing, a consultant would lead a small group of SMEs in designing a test plan based on their unique knowledge about the field (McGaghie, 1991). A more

modern and 'job related' approach includes combining the expertise of a measurement consultant, SMEs, and practice data into the test plan design process. Though SMEs will still be involved in the construction of the test plan from the practice analysis data, the eventual test plan content will be more closely linked to practice (from the practice analysis) and less linked to the particular group of SME participants which can introduce bias (Morgeson & Campion, 1997). The first step in the data collection process (after the task inventory or comprehensive practice analysis questionnaire has been designed) is to build a sample of individuals who will provide task or KSA ratings.

The primary concern in building a sample should be the qualification of the sample members to provide strong ratings of the tasks or KSAs based on their experience and expertise. Raymond (2001) notes that this is a particular concern if the practice analysis requires different types of data, for example criticality and frequency ratings may be better provided by different groups. Educators or supervisors will better know how important a task is to the professional's overall performance while an entry-level practitioner will know how often he or she is required to perform a task. The sampling process for a practice analysis study should be similar to that undertaken for any large study requiring a representative sample; this requires basic knowledge about the current population of professionals including a sense of the various practice settings and representation of various groups within the profession.

Knapp & Knapp (1995) emphasize the importance of including a large and representative sample of practitioners in a practice analysis sample. They suggest including up to 30-40 participants in a pilot test for a field with 500-1,000 practitioners and allowing the sample individuals time during a structured meeting to discuss the survey, instructions, and the completeness and accuracy of the content (p. 103). Kane et al. (1995) and Wang, Wiser, & Joseph (1999) suggest that a sample can contain 200-400 respondents in order to provide adequate generalizability, though Raymond (2001) suggest that this should be a starting point, with actual sample size relating more directly to the

specific research questions. A larger sample will improve the reliability and precision of the statistical analyses and will lend greater credibility to the outcomes. A large sample is necessary if comparisons are to be made between practice settings. A smaller sample, on the other hand, may be justified in the case of a practice analysis study with the purpose of revising an existing test plan, instead of creating a new one. A smaller sample may also be appropriate in a subspecialty or field that employs very few people. The strongest rule of thumb seems to be that the sample needs to be representative of practitioners (Knapp & Knapp, 1995) to clearly establish the link between practice and test specifications.

Depending on the requirements on the eventual exam, the sample should include SMEs, newly licensed professionals, and others. Raymond (2001, 2002) and Clauser, Margolis, & Case (2006) recommend building a representative sample of the practitioner population in terms of practice setting, urban/rural location, ethnic background, educational level, gender, and geographic region along with other demographic factors as necessary. Diversity within the sample is necessary, as individual practitioners' characteristics may influence his or her practice setting, the clients and problems encountered in practice, and other systematic differences across the profession. The simplest approach to sample selection involves inviting professionals who are members of a professional association or similar body to participate. This approach does require a commensurate effort to recruit individuals who, for example, practice in nontraditional settings (abroad, humanitarian relief, college or university health center, alternative clinic).

Building a Task Inventory Questionnaire

Since most practice analyses in credentialing will be based on a questionnaire of some type targeted toward the practicing population, this section will briefly discuss the process of selecting rating scales and analyzing scale data (see Tombleson, 2000, and De Champlain, 2007, for more detailed discussions). Given that the specific nature of scale selection depends on the scope,

purpose, timeline, and budget of the practice analysis study, this section will serve as a general overview. See Raymond (2002) and Fleishman & Quaintance (1984) for a more in-depth discussion about selecting and refining the scales for a task inventory/practice analysis questionnaire. The primary decision to be made at this point is whether the questionnaire will include task statements, KSAs, or some combination thereof. "The choice of conceptual foundation [tasks or KSAs] will effect everything…from item writing to score reporting" Raymond (2001, p. 391).

The responses practitioners provide on the practice analysis questionnaire (or task inventory) provide empirical evidence to link the KSAs included on the exam to the demands of practice (D'Costa, 1986; Harvey, 1991; Raymond, 2002). Initial scale selection is guided by the purpose of the examination. Since a credentialing examination should only include "activities actually performed in practice, or should give greater emphasis to activities performed more often or for greater periods of time" (Raymond, 2002, p. 27), a practice analysis questionnaire needs to provide data about which tasks are performed as well as how often. The scales selected for the questionnaire will need to contribute to the specific goals of the practice analysis study; for example a criticality scale will be useful for the creation of a new test plan but may not be as appropriate for revising a test plan. The statement of purpose and orientation of the study (towards people or tasks) will guide the design of scale items. As previously addressed, *task-oriented* approaches examine problem-solving activities and outcomes while *person-oriented* approaches emphasize the KSAs required for practice.

Task-oriented test plans are common in credentialing because the examinations emphasize knowledge, and task-oriented plans describe practice requirements in terms of semantic knowledge (Millman & Greene, 1989; Shimberg, 1981). Commonly included scales for a task-oriented study include responsibility / time on task, criticality / importance, and task acquisition / difficulty (addressing where the task was learned and how difficult it is to perform). Including a criticality scale is valuable to discern which skills are most crucial to protecting the public even if they are not

often performed (Kane, 1982). Criticality is challenging to measure (Morgeson & Campion, 1997; Raymond, 2001) and a task's criticality also varies across practice contexts. Given the complexities of assessing task criticality it is preferred to seek this type of rating from SMEs or other experts in the field (Kane, 1997) instead of entry-level practitioners. Entry-level practitioners on the other hand will be better suited to provide responsibility, time on task, and difficulty ratings. Identifying the more difficult skills in the practice analysis may be particularly useful for designing a credentialing exam because these skills are both necessary for practice *and* are challenging for incoming professionals to master (Raymond, 2002). This type of information may be useful for developing curriculum standards and for differentiating between entry-level and advanced practitioners' skills.

The standards as well as Kane et al. (1989) suggest that if two scales are to be used, frequency and criticality are preferred (AERA, et al., 1999, p. 160), while if only one may be used (because of the questionnaire length) overall importance to the profession is preferred (Newman et al., 1999). Sanchez and Levine (1989) found that overall importance scales are not as reliable as a composite importance scale derived from ratings of criticality and frequency. Task-oriented plans do pose several limitations. Primarily, the relationship between a particular task and practice may not be obvious to examinees or item writers. Knowledge requirements in turn can only be inferred from task data, requiring additional SME input into the creation of the test plan (Raymond, 1996). Gael (1983), Knapp and Knapp (1995), Raymond (2002), and Raymond and Neustel (2006) provide examples of common rating scales for practice activities and KSAs with a specific focus on building practice analyses for credentialing examinations.

Using criticality or frequency rating scales to gather information about practice activities is important, however most written credentialing exams do not directly assess an individual's ability to perform a task but instead "assess the cognitive knowledge and skills that provide the foundation for

competent task performance" (Raymond, 2002, p. 27). Knowledge requirements can only be inferred from the task-focused data (Raymond, 1996) requiring SME judgments to build an empirical link between the analysis and the test plan. Both item writers and examinees tend to emphasize the knowledge and cognitive content, justifying the use of KSA scales in addition to task ratings on a practice analysis questionnaire. It is difficult to write a useful test item about the frequency of taking a patient history - even though that skill may be critical for practice - while it is much easier to write items based on the cognitive skills and KSAs necessary to take a patient history. Using KSA ratings fits with a person-oriented study, since they address the individual knowledge and characteristics that are necessary for professional competence.

"[C]omplex professions are characterized not only by tasks and practice responsibilities but also by the KSAs that professionals bring to bear on those tasks" (Raymond, 2001, p. 384). Most credentialing examinations assess cognitive knowledge and skills in place of the tasks that are actually performed in the practice setting. Thus, most test specifications consist of KSAs making it logical and sometimes necessary to include KSA ratings in a practice analysis study. These data can be used identify major cognitive competencies necessary for practice (See Rosenfeld, Shimberg, & Thornton, 1983, for an example of KSAs for psychological practice).

KSA rating scales may be applied to assess the level of an individual's knowledge (of a KSA necessary for practice) in order to determine the level required for minimum competency (Kane, 1982), similar to the work of Rosenfeld et al. (1983) who applied a level-of-knowledge scale based on Bloom's taxonomy in their practice analysis of psychologists. There are methodological concerns related to classifying items and by cognitive level (see Cizek, 1995; Haladyna, 1994), however they are largely alleviated by the use of KSAs in a test plan (since they reflect cognitive processes instead of factual knowledge). See Kane (1994) and Haladyna (1994) for a further discussion linking the validity of score inferences (and interpretive arguments) to the use of KSAs.

KSA ratings in particular are thought to be very useful for the construction of most written credentialing exams. The "most significant limitation is that the link between a KSA and its application to clinical practice is not always apparent" (Raymond, 2001, p. 397). This weakness can be addressed by convening a SME panel to tie KSAs to practice tasks and facilitate the transition from 'expert' KSA language to practice activities (Fleishman & Quaintance, 1984). "Establishing a link between practice responsibilities and the KSAs on the credentialing exam can strengthen the validity evidence of the test inferences" (AERA, et al., 1999). The linkage process between practice analysis data and a test plan built from KSAs is much simpler than one based on tasks since the test plan. This approach provides strong evidence for content validity because the practice analysis will include the same descriptions or categories of practice as the test plan.

The specific wording and architecture of each task and KSA statement should be reviewed by SMEs and a pilot group of practitioners to ensure that the statements are written at the same level of specificity, are written clearly, and can be easily comprehended by respondents. The anchors (whether behavioral or numerical) should also be reviewed by SMEs before the questionnaire is released to ensure that the range and scale selected are both comprehensive and relevant to practice. Bourque and Fielder (2003) and Fink (2003) discuss specific recommendations for designing a pilot study and building a questionnaire for practice analysis.

Some attention should be given to the basic psychometric features of the included rating scales. First, the scale points need to be clearly defined and anchored. For a frequency scale, for example, this could mean building a scale that includes relevant time points from multiple times a day to quarterly (depending on the profession and the included tasks). Raymond and Neustel (2006) suggest that numerical labels should be avoided for scales categories that are not necessarily ordered and that verbal descriptors may be more easily interpretable by respondents. They also suggest that scales with absolute anchors are preferable to relative scales (p. 190) to reduce the potential for

response bias and to facilitate rating comparisons across people or practice settings. The selection of a nominal or ordinal scale should be guided by the intended methods for combining rating scale data; "the practice of treating ordinal scales as if they have interval or ratio properties is probably acceptable in most instances" (Raymond & Neustel, 2006, p. 191) but can have a significant impact on the process of building category weights from rating scale data (Spray & Huang, 2000). Task or KSA statements should be reviewed multiple times to ensure that they are all written at a similar level of specificity *and* that they address the minimum competency level required for practice (see Popham, 1992).

Though useful, KSA ratings pose different psychometric issues than task ratings. Combining data across KSA scales can be challenging especially if they are being used to differentiate across settings and/or are being used with a frequency scale (Harvey, 1991). Halo error is also an additional concern when using KSAs instead of (or in addition to) task ratings (see Rosenfeld, et al., 1983). "Halo error occurs when raters attend to general, more salient attributes of the entity being rated rather than considering the specific attribute implied by each rating scale" (Raymond, 2001, p. 388). Positive response bias is also a concern with KSA ratings as many practitioners struggle with identifying cognitive patterns or knowledge that is *not* crucial for successful practice (Raymond, 2001).

Methods for developing scales that are both job-related and linked to observable behaviors are provided by Fleishman & Quaintance (1984) and Raymond (2001, 2002). They suggest including KSA rating data along with task-based rating scale data in a practice analysis in order to maximize the collected data about criticality and the level of knowledge required for beginning practice. This may require two separate data collection steps as KSA ratings are most easily collected in a SME meeting or focus group setting, given their abstract nature (Harvey, 1991; Morgeson & Campion, 1997). Though a focus group meeting may not be as efficient for data collection as a survey or

questionnaire, the group setting will allow participants to discuss the relationships between the KSA and practice situations in order to reach a shared judgment of the overall importance of the KSA. Task-based rating scales, however, are well suited to a mail or internet-based questionnaire if up to 100 tasks are included (even across multiple scales) (Raymond & Neustel, 2006). Beyond 100 tasks may require a more complex sampling design (see Fink, 2003) as incorporated into the most recent accounting practice analysis (AICPA, 2008). If a practice analysis questionnaire consists of 100 task or KSA items across multiple rating scales it is easy to conceive of the eventual data set consisting of hundreds of individual variables (one for each item/scale pair).

Task-oriented plans appear to be more common for initial qualifying examinations, like the Step 1 USMLE examination (NBME, 1999; Raymond, 2001) and the Fundamentals of Engineering Exam (Raymond & Neustel, 2006) as the examinations are content based.

Developing content categories

A test plan based on either tasks or processes (KSAs) is valuable for credentialing though it has become increasingly common to capitalize on the strengths of both by creating a two-dimensional test plan based on content *and* process. "It is common for content outlines to be developed as a matrix with one dimension listing the topics to be covered. The other dimension can vary, but it commonly represents either the cognitive level or task…" (Clauser, Margolis, & Case, 2006, p. 705). Raymond (2002) suggests building a content by process matrix, perhaps with the rows corresponding to the content (or the medical problem and organ system) and the columns corresponding to the practice related behaviors (drawn from KSA ratings or a critical incident study (p. 32).

In credentialing the process dimension rarely refers to the cognitive level of the item (as in educational testing) but refers to the practice behaviors (diagnosis, communication skills, assessment, treatment planning). The matrix can include additional dimensions (like cognitive level) as employed

by De Champlain et al. (2007) in a practice analysis of physician assistants or additional practice related dimensions. La Duca et al. (1995) also suggest using cognitive objectives to enhance the content dimension of process-oriented test plans since they can serve to link the process (KSAs) to medical content.

Raymond (2002) sets forth his work with the orthopedic surgeons as an example of a multi-method test plan including content and practice. This test plan includes two sections, the first being a topical outline of academic disciplines related to medicine and second a process-based outline addressing the medical management of various orthopedic problems. De Champlain, et al. (2007) incorporated two dimensions; patient acuity (at three levels) and an overall importance rating for each statement related to safe and effective PA practice (p. 6).

There are several statistical methods in practice for modifying data from multiple rating scales to create the categories represented on a test plan. Factor analysis has been a popular tool for managing and organizing task rating data (D'Costa, 1986; Fleishman & Quaintance, 1984; Kane, et al., 1986) though others (Knapp & Knapp, 1995; Raymond, 2002) argue against factor analysis in favor of multivariate procedures. Factor analysis or multidimensional scaling procedures are useful for creating content categories based on respondent ratings, and the categories can then be used to guide the creation of a test plan. These procedures can also be used to identify shared processes or activities across practice subspecialties or settings. Raymond (2002), however, suggests using expert judgments to affirm and revise these categories (with caution because a small group of experts may introduce bias, see, Levine, et al., 1968; Morgeson & Campion, 1997). If judges are to be used for establishing content categories for the test plan, they can provide similarity ratings (for eventual use with multidimensional scaling or cluster analysis) which seems to utilize both the strongest features of judgmental methods with the organizational structure of an empirical method which allows for the judgmental information to be integrated systematically into the test plan (Raymond, 2002). This

approach appears in several practice analyses including emergency nursing (Schaefer, et al., 1992), nursing (McCloskey & Bulechek, 1992), and psychology (Rodgers, 1988).

Developing section weights

Once the categories have been established to organize the test plan content, judgmental and statistical methods are employed to decide on the number of test items allocated to each category, or weight, on the test plan. Weighting the various scales is an added challenge as data from any individual scale gives an incomplete picture of how professionals use their time. The data from a frequency or time-spent scale, for example, do not indicate the importance of various activities for the protection of the public.

Test plan category weights can be established through a judgmental process or by statistical methods. The most systematic approaches for judgmental weighting approximate a standard setting process; judges meet in panels to discuss the relevant information and provide a judgment, they receive feedback, provide an additional judgment, and then attempt to reach a consensus (Raymond, 2002, p. 34). While a less formal process may consist of inviting SMEs to suggest a number of test items (or percentage representation) for each category based on their knowledge of the field. "It is not uncommon for test designers to convene a group of content experts or policymakers to generate a content outline by consensus…Those involved in defining the examination content should include a diverse group of practitioners; among others, factors such as ethnicity, race, gender, urban/rural setting, and geographic region should be considered in selecting members for the group" (Clauser, Margolis, & Case, 2006, p. 705).

It is also popular to generate 'top down' judgmental data using a questionnaire representing the test plan. Each category is assigned a percentage of the test plan (such that the categories sum to 100%) and then the sub categories are allotted a section of the test plan (such that *they* sum to 100%) and the process continues until each level of categories has received a weight. Each percentage is

then assigned a number of the total test items. The 'bottom up' process is similar, except begins with the most specific level and moving toward the most general category (Raymond, 2002). This process presents a few issues, particularly when important topics at the lower levels are misweighted because their allotted percentages are dependent on the percentages assigned to the parent category. Judgmental methods may also pose problems if the SMEs do not have an accurate idea of how many test items a percentage of the examination represents (for example 50% of the exam may represent 100 questions or 20) (Raymond, 2002).

Beyond judgmental methods, statistical methods are popular for assigning category weights; they provide an empirical method for combining ratings from a task inventory or practice analysis questionnaire into test category weights based on the selected rating scales (Kane et al., 1989; Sanchez & Fraser, 1992; Slaughter & Newman, 1999; Spray & Huang, 2000). Statistical methods "provide a way to combine ratings from multiple scales into a single index of overall task importance" and "convert the index of overall task importance into a weight that corresponds to the number of test items in each category" (Raymond, 2002, p. 35). The most popular statistical procedures beyond factor analysis are the additive model, the multiplicative model, hierarchical ranking, and the Rasch rating scale model (Raymond & Neustel, 2006, p. 195). The models differ in terms of which rating scales they can incorporate and the way the scales are combined.

In creating an overall importance scale, Kane et al. (1989) advocate for the **multiplicative model** based on the "premise that criticality can be regarded as the importance of an activity per occurrence of a task...and that the overall importance of a task can be obtained by summing criticality over all occurrences of that task. Further if task frequency can be assumed to correspond to all occurrence of a task, then overall importance can be estimated by multiplying task criticality by frequency" (Raymond & Neustel, 2006, p. 195). This model allows for data to be aggregated across

respondents or for vectors of mean ratings (across tasks). The multiplicative model takes the most basic form of;

$$I_i = C_i F_i$$

where I_i refers to the importance for task *i*, C_i refers to the mean criticality rating for task *i*, and F_i refers to the mean frequency rating for task *i*. It indicates that criticality and frequency influence overall importance equally (and that the criticality and frequency scale have equal variances). In the case that they do not, Kane et al. (1989) suggest a modified form;

$$I_i = C^a_i F_i$$

where *a* refers to the square root of the variance in F_i divided by the variance in C_i.

Additive models are the simplest and most common approach (Raymond & Neustel, 2006; Sanchez & Fraser, 1992). Each task's "overall importance is determined by summing across all rating scales included in the model. For example frequency, criticality, and time-spent ratings may be summed into a single rating. This model is fully compensatory: a low rating on one scale can be offset by a high rating on another scale" (Raymond & Neustel, 2006, p. 195). This can create a problem when a task is performed frequently but is not critical in protecting the public; the high frequency rating will mask the low criticality rating. This model allows for data to be aggregated across respondents or for vectors of mean ratings (across tasks). More on the application of the additive model can be found in Sanchez & Levine, 1989), but in its most basic form the model takes the form of

$$I_i = C_i + F_i$$

where I_i refers to the importance for task *i*, C_i refers to the mean criticality rating for task *i*, and F_i refers to the mean frequency rating for task *i*. It is common to weight the terms to reflect SME judgment of their contribution to the overall importance of the task. For example if criticality were

more important than frequency the C term would be weighted. To use this model to generate weights for test specifications each task's weight can be determined by dividing its importance rating over the aggregate importance rating (resulting in weights that sum to 1.0).

The multiplicative model is preferred to even a weighted additive model in the case of scales with quasi-ratio properties (like absolute frequency or criticality) and a meaningful zero point because of the ease of interpretation and the additional variability possible in task weights (Raymond & Neustel, 2006).

Ranking methods use a conversation table, like the hierarchical ranking scheme proposed by Spray & Huang (2000). This method is used to recode responses from multiple scales in pairs, nested such that the most important scale takes precedence. For example if criticality ratings are considered to be more important than frequency ratings, recoding would be based on a criticality scale with nested values based on the possible responses for frequency. Hierarchical ranking requires that data are aggregated at the individual level creating combined ratings for each respondent.

The **Rasch Rating Scale Model** (Spray & Huang, 2000; Andrich, 1978) requires rank order data (produced by most practice analysis scales) and can be used as a method for combining scales (see Joseph & Taranath, 1999), though according to some researchers, the RRSM is not a preferred method for the following reasons: In the RRSM the overall importance factor is determined by the statistical properties of the data, which may be an issue if the scales measure criticality and frequency (which should be left to SME judgment, see Kane et al., 1989). The data also may violate a few key assumptions of item response theory, namely that the ratings are not locally independent (Spray & Huang, 2000) and the combined scales may not represent a unidimensional construct (Raymond, 2001). The RRSM, like the hierarchical model, requires that data be combined at the individual level, not the task level.

Even after the model has been selected, SMEs and/or researchers should provide input into how the data will be used to design the test plan and how individual scales' relative contributions should be weighted in the combining process. For example in a licensure exam criticality may be weighted more heavily than frequency given the overall purpose of licensure to protect the public and by ensuring that practicing professionals have the knowledge critical for practice. The properties of the data also will influence the manner in which the data are combined; for example scales that are highly correlated or provide similar information one should be selected (Sanchez & Fraser, 1992). If the additive or multiplicative model is selected the variance and covariance of each scale will influence the composite scale as the scale's contribution to the index may vary from the assigned contribution (Raymond & Neustel, 2006 refer to this as the scale's nominal contribution).

The eventual content category weights (and relevant categories or subcategories) will serve as the foundation for the test blueprint. This will guide the work of item writers and will also be used by educators and future examinees as they prepare for the test.

Ties to MCC Blueprinting process

Several different bodies facilitate Canadian medical assessments-- medical schools, the Medical Council of Canada (MCC), and provincial regulatory authorities. The MCC is currently engaged in the revision and validation process for the MCC Qualifying Examination part I; the examination used to assess the competency of candidates who have earned the M.D. degree and are seeking a position in a post-graduate training program (residency). Given the position of the MCCQE part I in the trajectory of medical education and assessment, the level of the examination is relatively clear (individuals who pass should be capable of performing basic medical tasks independently but will be entering supervised practice). The MCCQE part I assesses clinical knowledge / application through 196 multiple-choice questions administered in a computer-

adapting testing mode. The content specifications for this examination are comparable to its paper and pencil predecessor; the candidate is presented with an equal number of questions in each of six competency areas (General medicine, Obstetrics/Gynecology, Pediatrics, PHELO, Psychiatry, and Surgery). These questions are not specified or categorized beyond the competency area.

The proposed practice analysis will address how the current test plan (blueprint) needs to be modified to reflect current practice demands on newly practicing residents (based on both practitioner input, SME input, MCC objectives, and provincial databases) (Touchie, 2009). The previous discussion of practice analysis methodologies is particularly relevant to this study because the data to build several models for practice or to guide the creation of a task inventory or critical incident study already exist in varying levels of detail. Drawing from the CanMEDS competencies and MCC objectives, SMEs could create a model for professional practice consistent with LaDuca (1994) or Kane (1997)'s approach, thus creating an additional layer of protection against judgmental bias in the practice analysis process. The PPM and model-based practice analysis methods can be combined to guide the creation of a model for practice that will then undergo revision in questionnaire format – allowing for individual practitioners to weigh in on whether or not the proposed model reflects their practice.

If this process were to follow a comprehensive methodology, common in the medical professions, the practice analysis study would consist of three major parts; the practice model, the questionnaire, and the revised practice model/basis for the test blueprint. Once the practice model has been constructed (based on MCC objectives, the CanMEDS competencies [for KSAs], and/or other data sources) by SMEs, it will serve to categorize tasks (and KSAs) on the practice analysis questionnaire. The task ratings will then be combined and will be used to revise the practice model and task categories. These categories and tasks will then be used to construct the test plan following a weighting procedure (led by SMEs).

Given the nature of the proposed practice analysis, and the specific target population (first year residents), it is advisable to begin with a clear practice model. The NBME CRU study of first year residents determined that though most residents performed prevalent activities under general supervision, some were expected to perform these activities independently (NBME, 2009). If those activities are part of practicing safely and effectively (and physicians are expected to perform those tasks independently) competence in those activities should be represented on the credentialing examination. Additional studies indicate that there may be a technical mismatch between resident preparation and supervision, indicating that an updated practice analysis and revised examination may be useful (Baldwin, Daugherty, & Ryan, 2010; Langdale, Schaad, Wipf, Marshall, Vontver, & Scott, 2003).

Overall the MCC is well positioned to undertake a practice analysis study to revise the MCCQE part 1 blueprint given the data available to SMEs. If the shifts in resident practice are similar in Canada to those in the United States (as observed by the CRU study), an updated blueprint will guide the creation of an exam that more accurately reflects current practice.

Appendix A: Existing Data Sources (adapted from Touchie, 2009)

Data Banks

Health Statistics:

1. Community Health Survey (CCHS), 2007 Questionnaire (Statistics Canada)
2. Canadian Survey of Experiences with Primary Health Care, 2008 Questionnaire (Statistics Canada)
3. Population Patterns of Chronic Health Conditions in Canada, December 2007 (Health Council of Canada)
4. Health Indicators, 2008 (Statistics Canada and Canadian Institute for Health Information)
5. Discharge Abstract Database (Canadian Institute for Health Information)
6. National Ambulatory Care Reporting System (Canadian Institute for Health Information)
7. Hospital Mental Health Database (Canadian Institute for Health Information)
8. United States Department of Health & Human Services Agency for Healthcare Research & Quality
9. The National Board of Medical Examiners Data Commons project (currently under development; will contain a collection of data repositories from medical education and assessment sources)

Critical Incident Data:

1. Canadian Medication Incident Reporting and Prevention System (Canadian Institute for Health Information)
    a. Drug Claims by Seniors: An Analysis Focusing on Potentially Inappropriate Medication Use, 2000-2006
2. Canadian Medical Protective Association
3. Canadian Patient Safety Institute
4. Critical Incident Monitoring in Emergency Medicine Database (Critical incident monitoring in the United Kingdom)
5. American Society of Anesthesiologists Critical Incidents Reporting System (2009)

Appendix B: Practice Analysis Methods not preferred for credentialing

There are many other methodological approaches to practice analysis that have been used more commonly in industry and the military and are not appropriate for credentialing given their focus on job classification or limited scope (to jobs in skilled labor, for example). The **Functional Job Analysis** (Fine & Wiley, 1971; Fine, 1986) is a popular method in industry and has been used to categorize jobs for the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1977) and is useful for the purpose of categorizing jobs and roles within an organization. This method is generally performed by first identifying the purpose of the occupation, then identifying the acts performed by the worker and the processes (physical, mental, etc.) involved. The occupations are then sorted based on "the degree to which it deals with data, people, and things" (Knapp & Knapp, 1995, p.97). The specificity required by the FJA "far exceeds the level of specificity required to describe a profession for licensing purposes" (Knapp & Knapp, 1995, p. 97) and will not provide adequate information for designing a knowledge- and skill-based test plan.

The **Position Analysis Questionnaire** (McCormick, Meeham, & Jeanneret, 1972, 1977) is generally used to compare job characteristics across different occupations based on the job's relative emphasis on information input, mediation processes, worker output, interpersonal activities, work situation and job context, and miscellaneous (including methods of payment, scheduling, etc.) (Knapp & Knapp, 1995). This type of analysis is performed either by the individual currently working at the job or a job analyst. Knapp and Knapp (1995) cite several authors that argue that the PAQ approach may be too general and that it may be difficult to apply the general questionnaire items (particularly those about worker output or equipment use) to professional activities (see Landy, 1989 and Cornelius, Schmidt, & Carron, 1984).

The **Job Elements** Method (Primoff, 1975) begins with a SME meeting to discuss and generate a list of the "skills, knowledges, abilities, and other worker characteristics (the job elements) required to perform the job in question" (Levine et al., 1980, p. 525). These elements are then rated

by SMEs to establish their relative importance to performance in the job. The ratings are then combined and analyzed to establish the most important elements that are then used for job classification and employee interventions. This is not necessarily applicable to credentialing in the health professions given the complex cognitive demands required for successful performance and strong client outcomes. Similarly, the **Fleishman Job Analysis** Survey (formerly the Ability Requirements Scales, Fleishman & Quaintance, 1984) is used to assess 52 employee abilities (including cognitive, physical, psychomotor, and sensory) to determine the level of functioning required for successful performance in a specific position (Fleishman & Mumford, 1991). The required abilities (and the corresponding levels) are then used to classify a position within an organization. This is a job analysis and classification method and is not easily applied to job candidate evaluation.

The **Job Diagnostic Survey** (Hackman & Oldham, 1975) is designed to 'diagnose' jobs before they are re-designed as well as to research the impact of redesigned jobs on the employees. This measure is based on the idea that positive work outcomes are possible when an employee finds his or her work meaningful, is responsible for the work outcomes, and is knowledgeable about his or her work activities. These conditions are influenced by five core job dimensions (skill variety, task identity, task significance, autonomy, and feedback). The JDS itself is used to measure the motivating potential of a position based on a combination of the job's score on the five core dimensions. This measure is not necessarily applicable to practice analysis in credentialing due to its design. Instead of assessing the knowledge, skills, or abilities required for minimum practice it is designed to assess the motivation and potential motivation of an employee in a position (Hackman & Oldham, 1975) thus the results are focused on the employee/job match instead of competence. However this type of measure could be potentially useful in evaluating the responsibilities and demands placed on current professionals *within* a setting.

References

AICPA (2008). Practice analysis technical report. December 2008. Accessed online from http://www.aicpa.org/BecomeACPA/CPAExam/ExaminationContent/ContentDevelopment/DownloadableDocuments/2008_Practice_Analysis_Technical_Report.pdf on June 2, 2011.

Altmaier, E.M., Pearson, K.S., Gorbatenko-Roth, K.G., & Ugolini, K.A., (1997). A prospective study to select and evaluate anesthesiology residents: Phase I, the critical incident technique. *Journal of Clinical Anesthesiology, 9*(8), 629-636.

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

Andrich, D. (1978). A rating formulation of ordered response categories. *Psychometrika, 43,* 561-573.

Arbet, S., Lathrop, J., & Hooker, R.S. (2009). Using practice analysis to improve the certifying examinations for PAs. *Journal of the American Academy of Physician Assistants, 22*(2), 31-36.

Baldwin, D. C., Jr., Daugherty, S. R., & Ryan, P. M. (2010). How residents view their clinical supervision: A reanalysis of classic national survey data. *Journal of Graduate Medical Education*, 37-45.

Berk, R.A. (Ed.). (1984). *A guide to criterion-referenced test construction*. Baltimore: Johns Hopkins University Press.

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

Brain, D. (1991). Practical knowledge and occupational control. *Sociological Forum, 6*(2), 239-268.

Brumback, G.B., Romashko, T., Hahn, C.P., & Fleishman, E.A. (1974). *Model procedures for job analysis, test development and validation* (Vol. 1): *Final report of the first phase.* Washington D.C.: American Institutes for Research.

Cascio, W.F. (1982). *Applied psychology in personnel management*. Reston, VA: Reston.

Cornelius, E.T. (1988). Practical findings from job analysis research. In S.Gael (Ed.), *The job analysis handbook for business, industry, and government, Vols. I & II* (pp.48-64). New York: Wiley.

Cornelius, E.T., Schmidt, F.L., & Carron, T.J. (1984). Job classification approaches and the implementation of validity generalization results. *Personnel Psychology*, *37*(2), 247-260.

De Champlain, A.F., Cuddy, M.M., LaDuca, T. (2007). Examining contextual effects in a practice analysis: An application of dual scaling. *Educational Measurement: Issues & Practice, 26*(3), 3-10.

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

Edwards, J.C., Currie, M.L., Wade, T.P., & Kaminski, D.L. (1993). Surgery resident selection and evaluation: A critical incident study. *Evaluation and The Health Professions, 16*, 73-86.

Fine, S. (1986). Job analysis. In R.A. Berk (Ed.), *Performance assessment: Methods and applications* (pp.53-81). Baltimore: Johns Hopkins University Press.

Fine, S., & Wiley, W.W. (1971). *An introduction to functional job analysis.* Washington D.C.: Upjohn Institute for Employment Research.

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

Fitzgerald, K., Seale, S., Kerins, C.A., & McElvaney, R. (2007). The critical incident technique: A useful tool for conducting qualitative research. *Journal of Dental Education, 72*(3), 299-304.

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

Fleishman, E. A., & Mumford, M. D. (1991). Evaluating classifications of job behavior: a construct validation of the ability requirement scales. *Personnel Psychology, 44*(3), 523-575.

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

Hackman, J. R., & Oldham, G. R. (1975). Development of the Job Diagnostic Survey. *Journal of Applied Psychology, 60*(2), 159-170.

Haladyna, T.M. (1994). A research agenda for licensing and certification testing validation studies. *Evaluation and the Health Professions, 17,* 242-256.

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

Hecht, K.A. (1979). Current status and methodological problems of validating professional licensing and certification exams. In M.A. Bunda and J.R. Sanders (Eds.), *Practices and problems in competency-based education.* Washington, D.C.: National Council on Measurement in Education.

Jacobs, A.M., Fivars, G., Edwards, D.S., & Fitzpatrick, R. (1978). *Critical requirements for safe/effective nursing practice.* Washington, DC: American Nurses' Association.

Jenkins, S.M., & Curtin, P. (2006). Adapting job analysis methodology to improve evaluation practice. *American Journal of Evaluation, 27*, 485-494.

Joseph, M.J., & Taranath, S.N. (1999, April). *Obtaining task measures from job analysis data.* Paper presented at the meeting of the National Council on Measurement in Education, Montreal, Quebec, Canada.

Kane, M.T. (1982). The validity of licensure examinations. *American Psychologist, 37*(8), 911-918.

Kane, M. (1994). Validating interpretive arguments for licensure and certification examinations. *Evaluation and the Health Professions, 17,* 133-159.

Kane, M.T. (1997). Model-based practice analysis and test specifications. *Applied Measurement in Education, 10*(1), 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 and The Health Professions, 18*(1), 29-50.

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. (1980). The structure of competence in the health professions. *Evaluation and The Health Professions, 3,* 253-288.

LaDuca, A. (1994). Validation of professional licensure examinations: Professions theory, test design, and construct validity. *Evaluation and The Health Professions, 17*(2), 178-197.

LaDuca, A. (2006). Commentary: A closer look at task analysis: Reactions to Wang, Schnipke, and Witt. *Educational Measurement: Issues and Practice, 25*(2), 31-34.

LaDuca, A., Engle, J.D., & Risley, M.E. (1978). Progress toward development of a general model for competence definition in the health professions. *Journal of Allied Health, 7,* 149-155.

LaDuca, A., Taylor, D. D., & Hill, I. K. (1984). The design of a new physician licensure examination. *Evaluation and The Health Professions, 7,* 115-140.

Landy, F.J. (1989). Psychology of work behavior (4th ed.). Pacific Grove, CA: Brooks/Cole Publishing.

Langdale, L. A., Schaad, D., Wipf, J., Marshall, S., Vontver, L., & Scott, C. S. (2003). Preparing graduates for the first year of residency: Are medical schools meeting the need? *Academic Medicine, 78*, 39-44.

Levine, E.L., Ash, R.A., & Bennett, N. (1980). Exploratory comparative study of four job analysis methods. *Journal of Applied Psychology, 65*(5), 524-535.

Levine, E.L., Thomas, J.N., & Sistrunk, F. (1988). Selecting a job analysis approach. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government (Vols. I & II;* pp.339-352). New York: Wiley.

Levine, H.G., McGuire, C., Miller, G.E., & Larson, C. (1968). *The orthopaedic training study: Final report.* Chicago: University of Illinois, Center for the Study of Medical Education.

McCloskey, J.C., & Bulechek, G.M. (Eds.). (1992). *Nursing interventions classification.* St. Louis, MO: Mosby-Year Book.

McCormick, E.J., Jeanneret, P.R., & Mecham, R.C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire. *Journal of Applied Psychology, 56,* 347-368.

McCormick, E.J., Meeham, R.C., & Jeanneret, P.R. (1977). *Technical manual for the Position Analysis Questionnaire (PAQ) (System 2).* West Lafayette, IN: University Book Store.

McGaghie, W.C. (1980). The evaluation of competence: Validity issues in the health professions. *Evaluation and The Health Professions, 3,* 289-320.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed. pp. 13-103). New York: American Council on Education and MacMillan.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.) *Educational measurement* (3rd ed. pp. 335-366). New York: American Council on Education and MacMillan.

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

National Conference of Examiners for Engineering and Surveying. (1999). *Fundamentals of engineering (FE) examinations: Principles & practice of engineering examinations*. Clemson, SC: Author.

Nelson, D. (1994). Job analysis for licensure and certification exams: Science or politics. *Educational Measurement: Issues & Practice, 13*(4), 29-35.

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 meeting of the National Council on Measurement in Education, Montreal, Canada.

Patterson, F., Ferguson, E., & Thomas, S. (2008). Using job analysis to identify core and specific competencies: implications for selection and recruitment. *Medical Education, 42,* 1195-1204.

Popham, W.J. (1992). A tale of two test-specification strategies. *Educational Measurement: Issues & Practice, 11*(2), 17-17, 22.

Primoff, E.S. (1975). How to prepare and conduct job element examinations. Washington D.C.: U.S. Government Printing Office (GPO No. 006-000-00893-3).

Raymond, M.R. (2001). Job analysis and the specification of content for licensure and certification examinations. *Applied Measurement in Education, 14*(4), 369-415.

Raymond, M.R. (2002). A practical guide to practice analysis for credentialing examinations. *Educational Measurement: Issues & Practice, 21*(3), 25-38.

Raymond, M.R., (2005). An NCME instructional module on developing and administering practice analysis questionnaires. *Educational Measurement: Issues & Practice, 24*(2), 29-42.

Raymond, M.R., & Neustel, S. (2006). Determining the content of credentialing examinations. In S.M.Downing, & T.M. Haladyna, *Handbook of test development* (pp. 181-223). Mahwah, NJ: Lawrence Erlbaum Associates.

Rodgers, J.L. (1988). Structural models of the American Psychological Association in 1986: A taxonomy for reorganization. *American Psychologist, 43,* 372-382.

Rosenfeld, M., Shimberg, B., & Thornton, R.F. (1983). *Job analysis of licensed psychologists in the United States and Canada.* 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.

Schaefer, 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.

Shimberg, B. (1990). Social considerations in the validation of licensing and certification exams. *Educational Measurement: Issues & Practice, 9*(4), 11-14.

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

Tombleson, P., Fox, R.A., & Dacre, J.A. (2000). Defining the content for the objective structured clinical examination component of the professional and linguistic assessments board examination: Development of a blueprint. *Medical Education, 34*, 566-572.

Touchie, C. (2009). EDAC blueprint whitepaper. Unpublished Manuscript.

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

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

Wang, N., Witt, E.A., & Schnipke, D. (2006). Rejoinder: A further discussion of job analysis and use of KSAs in developing licensure and certification examinations: A response to LaDuca. *Educational Measurement: Issues and Practice, 25*(2), 34-38.

Wang, N., & Witt, E.A. (2005). Use of knowledge, skill, and ability statements in developing licensure and certification examinations. *Educational Measurement: Issues & Practice, 24*(1), 15-22.

Zaglaniczny, K.L. (1993). Council on certification professional practice analysis. *Journal of the American Association of Nurse Anesthetists, 61*, 241-255.