The Development of Knowledge Requirement Scales in the Health Professions

Mark R. Raymond, National Board of Medical Examiners
Nance Cavallin, American Registry of Radiologic Technologists

Introduction

Background

Construct maps (Wilson, 2005; Luecht, 2013; Mislevy & Riconscente, 2006) are useful tools for translating traditional content outlines into more detailed item (i.e., task) specifications. For each knowledge, skill, or ability (KSA) domain to be tested, a construct map arranges specific behaviors representing different levels of performance on a continuum from low proficiency to high proficiency. Construct maps are commonly presented as a graphic that depicts the relationship between the latent construct of interest and the observed behaviors to be elicited by different assessment tasks (Wilson, 2005). These graphics provide an appealing way to organize the claims to be made about examinees on the basis of their test performances.

Precursors to construct maps in educational testing can be traced to early writings on item specifications (e.g., Bormuth, 1970; Hively, 1974; Popham, 1984) and the Rasch model (Wright & Stone, 1979; Masters, Adams, & Lokan, 1994). It is also hard to ignore the influence of work on concept mapping (Ausubel, 1963; Novak & Gowin, 1984). Wilson (2005) provides an integrated and fully-developed description of construct maps. Other examples can be found in Wilson (2005), Gierl and Leighton (2010), Luecht (2013), Wyse (2013), and by Mislevy (1994), although Mislevy did not call it as such. Applications to certification testing are few; Raymond and Luecht (2013) provide a sample construct map for a hypothetical certification test, while a construct map for financial accounting is provided by Burke, Mattar, Stopek, & Eve (2014), with a variation given by Raymond (2015).

Independent of those efforts, Fleishman and colleagues developed an entire system of rating scales, not unlike construct maps, for judging and describing the human ability constructs (e.g., spatial orientation; speed of closure) required to successfully perform various jobs. These maps were initially referred to as the *Ability Requirements Scales* (Fleishman, 1975; Fleishman & Quaintance, 1984), and more recently as the Fleishman Job Analysis Survey (F-JAS; Caughron, Mumford, & Fleishman, 2012). The rigorous methodology employed by Fleishman and colleagues to develop their construct maps borrowed from both quantitative and qualitative research traditions.

At the risk of oversimplification, the Fleishman scales can be thought of as being developed in two phases. The first phase involved deriving a taxonomy of human abilities. The taxonomy was based on an extensive analysis of the experimental, correlational, and experimental literature. Many of the studies contributing to the taxonomy had been completed by the likes of JP Guilford, JB Carroll, and JW French, with many completed under the auspices of ETS. The F-JAS presently includes rating scales for 73 constructs from the cognitive, psychomotor, and affective domains. There are rating scales for constructs such as written comprehension, speed of closure, multilimb coordination, dependability, and assertiveness (Caughron et al., 2012). The second phase involved the development of rating scales for judging the relevance of each construct to a particular job or series of tasks. This required carefully defining each construct, listing sample behaviors for each, and locating each sample behavior on a vertical rating scale, with low performance (or easier behaviors) located near the bottom and high performance (or difficult behaviors) located toward the top. They relied heavily on a strategy similar to that proposed by Smith and Kendall (1963) in their now classic article on developing behaviorally-anchored rating scales (BARS). Sample scales can be found in Fleishman and Quaintance (1984).

The various approaches to construct mapping differ in nomenclature, method of development, presentation mode (textual vs. graphic), and in their level of detail or abstraction. However, they all seem to be getting at the same thing: an ordered arrangement of observable behaviors related to a construct.

Purpose

The purpose of this report is to summarize an effort undertaken in 2003 to develop a series of construct maps, referred to here as Knowledge Requirements Scales, to augment traditional test blueprint for a certification test in radiologic technology. While the paper focuses primarily on the method of development, we also describe a pilot study that involved surveying samples of educators and hiring managers regarding the level of knowledge required for each KSA domain. The paper concludes by briefly highlighting issues and obstacles to the development and interpretation of construct maps in credentialing.

Our interest in construct maps grew out of an effort to improve the job analysis scales typically used to elicit judgments from subject matter experts (SMEs) regarding the frequency or importance of the KSAs to be covered on credentialing tests. Knowing about KSA frequency or importance may provide guidance about *what topics* to test, but they are not helpful when trying to decide the *complexity* or *depth of knowledge* at which to test. It is one thing to identify topics that are important for professional practice, but something more to further explicate the depth of understanding required for each topic. The present effort was motivated in part by Kane's (1982) observation that credentialing exams are intended to assess knowledge at the *level* required for effective practice (Kane, 1982) and partly by Fleishman's work on the Ability Requirements Scales. The goal was to use the results to help inform item development efforts and possibly standard-setting exercises (Wyse, 2013).

Construct Maps in Radiologic Science: A Pilot Study

Development of Knowledge Requirements Scales

In 2003 we initiated a pilot project to better define the level of knowledge and skill required of radiologic technologists. The Content Specifications for the Radiography Examination developed by the American Registry of Radiologic Technologists (ARRT®) comprise five major content domains:

A. Radiation Protection
B. Equipment Operation and Quality Control
C. Image Acquisition and Evaluation
D. Imaging Procedures
E. Patient Care and Education

The first four domains are primarily technical in nature, consisting of both semantic and procedural knowledge. The last domain encompasses areas like communication, professionalism, and legal issues. Each major domain consists of three to four additional levels of detail (e.g., C.1.b.2.a). For this project we concentrated on 33 mid-level domains (e.g., radiobiology; digital image enhancement; legal and ethical issues; radiographic positioning: shoulder). The 33 domains covered basic science topics, imaging techniques, and knowledge of radiographic positioning.

Scale development required two meetings of a panel of six SMEs and considerable effort from project staff and SMEs over a period of about nine months. SMEs were assigned to work on those KSA domains with which they were most familiar. The effort proceeded as follows:

1. For each knowledge domain, SMEs were asked to write examples of knowledge and skill that could be expected of radiographers at low, moderate, and high levels of proficiency, focusing exclusively on the cognitive domain. SMEs were asked to produce more statements than would ultimately be used, understanding that the best examples would be kept and that others would be discarded. The goal was not to exhaustively describe a domain, but to sample from it.
2. Project staff edited the cognitive behaviors for consistency and eliminated redundancies, contacting SMEs for clarification as needed. Staff then identified two to four statements at each level of proficiency (low, moderate, and high) to retain for further development.
3. After culling and refinement, the behavioral statements were returned to SMEs who were then asked to rate, on a scale of 1 to 7, the level of cognitive complexity associated with each behavior where each behavior fell (1 = basic comprehension; 7 = in-depth understanding). The mean rating for each behavior across SMEs was used to locate that behavior on the scale. SMEs understood that the statements and ratings would be discussed later during a meeting.
4. SMEs met to discuss ratings, and finalize the sample behaviors to be included on the final scale, and revise the wording of some statements in minor ways. Behaviors with notable disagreement across SMEs were discarded, leaving four to six sample behaviors for most scales. Staff suggested the idea of engaging in another round of ratings, but SMEs were sufficiently comfortable with results of the first round; thus the original means for the retained statements were used to locate each statement on its scale.

The process described above was employed for each of the 33 scales. Figure 1 presents the Knowledge Requirement Scale created for the KSA domain of *infection control*. The sample behaviors cover a range of concepts and principles that span the infection control domain. In this example, the top two behaviors pertain to

specific aspects of tuberculosis, while the bottom two scales address the general transmission of pathogens. Figure 2 is the Knowledge Requirement Scale for radiographic positioning (shoulder). For both scales it is evident that the behaviors progress from a basic understanding of simple concepts, through application, and end up with analysis.

Knowledge Requirements Survey

Development & Administration. To help determine the level of knowledge at which test items within each of the 33 KSA domains ought to be targeted, we surveyed two groups of stakeholders: educators and hiring managers in medical imaging departments. Knowledge Requirements Scales were created for all 33 mid-level knowledge domains and formatted into a questionnaire, with each scale requiring a full page for display and instructions. Just above each scale, but following the definition of each knowledge and skill domain (Figure 4) the following instructions appeared:

> **Instructions:** With the preceding definition in mind, use the 1 to 7 scale below to indicate the depth of knowledge and skill you expect of radiographers who graduate from your program. Please provide two ratings.
>
> First, place the letter ***T*** on the scale to indicate the level of knowledge and skill you expect of your *typical* (or *average*) graduate.
>
> Then, place an ***M*** on the scale to indicate the minimal level of knowledge and skill you would accept from the *marginal* graduate – the *lowest* level you are willing to accept from a graduating student.

The 33 maps were formatted into a survey and field tested with radiologic science educators (n = 303). An alternate version of the questionnaire was completed by healthcare facility managers who hire radiographers (n = 31). The alternate wording focused on the level of proficiency expected of the typical and the marginal *new hire* rather than the graduating student. Each scale required a full page for presentation, meaning that the survey consisted of 33 pages.

Survey Results. As expected, mean ratings for the *typical* student (or employee) were higher than ratings for the *marginal* student across all scales. In addition, educators consistently provided higher ratings than hiring administrators, indicating that educators generally had higher expectations than administrators. The table below summarizes the findings across all KSA Rating Scales.

Table 1: Mean Ratings Across all 33 KSA Domains[1]

	Target Group	
SME Group	Typical Student or Employee	Marginal Student or Employee
educators (n=303)	5.6	3.9
managers (n=31)	5.0	3.2

Although the mean differences between the two SME groups are notable, they rank ordered the KSA domains similarly (r = .75 for typical; r = .73 for marginal). Differences between the two groups were smaller on the more applied, job-relevant domains (e.g., positioning for a shoulder radiograph; image processing), and larger for less directly applied domains (e.g., radiobiology) or domains that were the primary responsibility of other health-care personnel (e.g., medications and drug administration). Across most KSA domains, the difference between typical and marginal ratings hovered around the constant of 1.7 for educators and 1.8 for administrators, but this was not always the case.

As one example, consider Figure 1. The educator means for the typical and marginal student were 5.6 and 4.2 (1.4 difference) while corresponding means for hiring administrators were 5.3 and 3.3 (2.0 difference). Keep in mind that the administrator sample size was 31 and ratings had standard errors in the region of .20. As another example,

[1] Standard errors of mean ratings averaged about .05 for educators and .19 for administrators.

the educator means for shoulder positioning in Figure 2 were 6.0 and 4.8, while the administrator means were 5.6 and 4.1. These within-scale comparisons are interpretable and can be useful. However, while it is tempting to compare means across scales (e.g., shoulder positioning to infection control), such comparisons are difficult to interpret because the scales are specific to the verbal descriptors that make up a KSA domain.

Discussion

This paper provided a glimpse into one attempt to clarify the level of knowledge required for the KSAs covered by a certification test in the radiologic sciences. While the primary motivation for the original effort was to provide a tool to better target item writing, other applications come to mind. For example, the results could be used to better inform standard setting panels regarding performance expectations (Wyse, 2013). In this way, the scales might function as performance level descriptors (PLDs) in education testing. Another application would be to incorporate these scales into item classification efforts, whereby SMEs judge each newly written item in terms of their location on the relevant scale. Such ratings could be useful for test assembly and for predicting item difficulty (e.g., Luecht, 2013). Another potential use would be as adjunct information to gather as part of a content relevance study. Such studies require that SMEs judge the relevance of test item to job performance; the depth of knowledge ratings could provide a way to identify and discard items that are quite distant from their ideal level of complexity.

The survey provided a way to obtain valuable information from stakeholders. While the survey was useful, the lessons we learned pertain more to the methodology used for developing the scales and the issues to be addressed when replicating such an effort. We believe that the general approach used here – a modification of the method used for developing BARS (Smith & Kendall, 1963) – would benefit the development of construct maps in both education and credentialing. Key features of the Smith and Kendall method include:

- ***Multiple SMEs.*** The present investigation relied on 6 SMEs assisted by two staff, and the SMEs were knowledgeable of the performances under consideration. Ideally, more than 6 SMEs would be involved to (a) further divide the work among; (b) have broader representation; and (c) obtain more stable estimates of each behavior on its scale.
- ***Specific, concrete, and verifiable behaviors.*** This is an important feature and can be assured through proper selection of SMEs who know or have witnessed the behaviors of interest. With increased specificity comes a greater number of behaviors. This issue was addressed by designing each scale to include a *sample* of behaviors, rather than to present an exhaustive list of behaviors. We comment on this further below.
- ***Systematic methods for eliciting and collating SME statements and judgments.*** SME were given guidelines for generating behavioral statements; staff edited them for consistency. Ratings were used to identify and cull ambiguous or ill-fitting statements, and to locate statements on the construct map.
- ***Iterative process.*** In the present study, there were typically 3 to 5 rounds for developing behavioral statements (SMEs at meeting – staff – SMEs – staff – SMEs at meeting). The present effort included only one round of ratings to locate behaviors on each scale. An additional round would have been beneficial but staff sensed that SMEs were growing impatient with the process.

  The original Smith and Kendall procedure involved a step called "retranslation," which is a consistency check that requires SMEs to reclassify behavioral statements (after being blinded to the parent KSA domain from which the statements originated). This step is probably unnecessary for well-circumscribed knowledge domains because the domain to which a behavior belongs will usually be obvious. However, for complex domains, ambiguous constructs, or soft skills (e.g.,interpersonal skills; teamwork), the retranslation step is critical.

A few issues arose in the course of this pilot effort. First, we are not convinced that the numerals on each scale are necessary. While a numeric approach provides a vehicle for locating behavioral statements on the scale, there are other, albeit more tedious, approaches to the same end. Ratings do provide a way to compare groups, to assess reliability, and conduct other analyses; however, comparing one scale to another in effort to determine which KSA domain demands a deeper level of understanding is not warranted. That is, within-scale comparisons have some meaning, but between-scale comparisons generally do not.

Second, the maps depicted in Figures 1 and 2 present *sample* behaviors. It is assumed that item writers will write items to numerous cognitive skills that were not listed. In contrast, concept maps in evidence-centered design

(ECD) or assessment engineering (AE) often are considered exhaustive lists of requisite behaviors (for an example, see Burke et al, 2014). A third and related issue pertains to the specificity of the KSA domains (or constructs) to be mapped. For the present effort, we chose 33 mid-level KSA descriptors from a content outline (e.g., radiobiology). This meant that the list of behaviors within a KSA domain consisted of quite different topics. In retrospect, it would have been more informative to drill into the content outline one level deeper (e.g., radiosensitivity; photon-matter interactions). However, this would have meant developing well over 150 construct maps – an ambitious undertaking. And even that may not have been sufficient specificity to, say, develop task models.

This leads into a more general issue when applying construct maps and other aspects of principled test design (e.g., ECD, AE) to credentialing. While principled test design has influenced educational testing in significant ways (Huff, Steinberg & Matts, 2010; Pellegrino, 2013), it has not gained much traction in credentialing. On one hand, the time and effort rigor required to develop construct maps and related tools can undoubtedly increase the quality of credentialing tests. On the other hand, developing just one tool— construct maps— for each of the numerous KSAs required for just one profession is an expensive and time-consuming endeavor. The cognitive skills required for many professions span numerous KSA domains (e.g., statistics, anatomy, pathology, biochemistry, and so on), and for each domain one can envision that scores of construct maps would be required. While the development of so many construct maps is certainly doable, the question becomes, "To what extent will such efforts improve the accuracy of the inferences and decisions made?"

References

Ausubel, D.P. (1963). *The Psychology of Meaningful Verbal Learning.* New York: Grune and Stratton.

Bormuth, J.R. (1970). On a theory of achievement test items. Chicago: University of Chicago Press.

Burke, M., Mattar, J., Stopek, J. & Eve, H. (2014). Modeling Complex Performance Tasks. Annual Meeting of the National Council on Measurement in Education, Philadelphia

Caughron, J.J., Mumford, M.D., & Fleishman, E.A. (2012). The Fleishman Job Analysis Survey. In M.A. Wilson, W. Bennet Jr., S.G. Gibson, & G.M. Alliger (Eds.). *Work analysis: methods, systems, applications, and science of work measurement in organizations* (pp. 231-246). New York: Routledge.

Fisher, K.M., Wandersee, J.H., & Moody, D.E. (2000). *Mapping biology knowledge*. Boston, MA: Kluwer Academic Publishers.

Fleishman, E.A. (1975). Toward a taxonomy of human performance. *American Psychologist, 30*, 1127-1149.

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

Gierl, M., & Leighton, J (2010). *Developing construct maps to promote formative diagnostic inferences using assessment engineering*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education*, *23*, 310-324.

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

Luecht, R. M. (2006, May). *Engineering the Test: Principled Item Design to Automated Test Assembly.* Paper presented at the Annual Meeting of the Society for Industrial and Organizational Psychology.

Luecht, R.M. (2013). An introduction to assessment engineering for automatic item generation. In M.J. Gierl & T.M. Haladyna (Eds.), *Automatic Item Generation: Theory and Practice* (pp. 59-76). Routledge: New York.

Masters, G.N., Adams, R., & Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research*, *21*, 595-609).

Mislevy, R. J. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R.J. & Riconscente, M.M. (2006). Evidence-centered assessment design. In S.M. Downing & T. M. Haladyna (Eds.). *Handbook of test development* (pp. 61-90). Mahwah, NJ: Lawrence Erlbaum Associates.

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science*, *340*, 320-323.

Novak, J.D., & Gowin, D.B (1984). *Learning How to Learn.* New York and Cambridge, UK: Cambridge University Press.

Popham, W.J. (1984). Specifying the domain of content or behaviors. In R.A. Berk (Ed.), *A Guide to Criterion-Referenced Test Construction*. Baltimore, MD: Johns Hopkins University Press.

Raymond, M. R. & Luecht, R. L. (2013). Licensure and certification testing. In K.F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology*. Washington, DC: American Psychological Association.

Raymond, M.R. (2015). Job Analysis, practice analysis and the content of credentialing tests, in S. Lane, M.R. Raymond, & T.M. Haladyna (eds.), *Handbook of Test Development, 2nd ed*. New York, NY: Routledge.

Smith, P.C. & Kendall, L.M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, *47*, 149-155.

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Wright, B., & Stone, M (1979), Best Test Design Chicago: Mesa Press.

Wyse, A.E. (2013). Construct Maps as a Foundation for Standard Setting. *Measurement: Interdisciplinary Research and Perspectives, 11(4),* 139-170.

Figure 1: Construct Map for Infection Control

Infection Control. To protect patients, other staff, and themselves from infectious disease, radiographers are expected to have an understanding of the following topics: common terminology (nosocomial, asepsis, etc.), cycle of infection, routes of transmission (airborne, vector, etc.), and types of precautions (standard, transmission-based, and isolation).

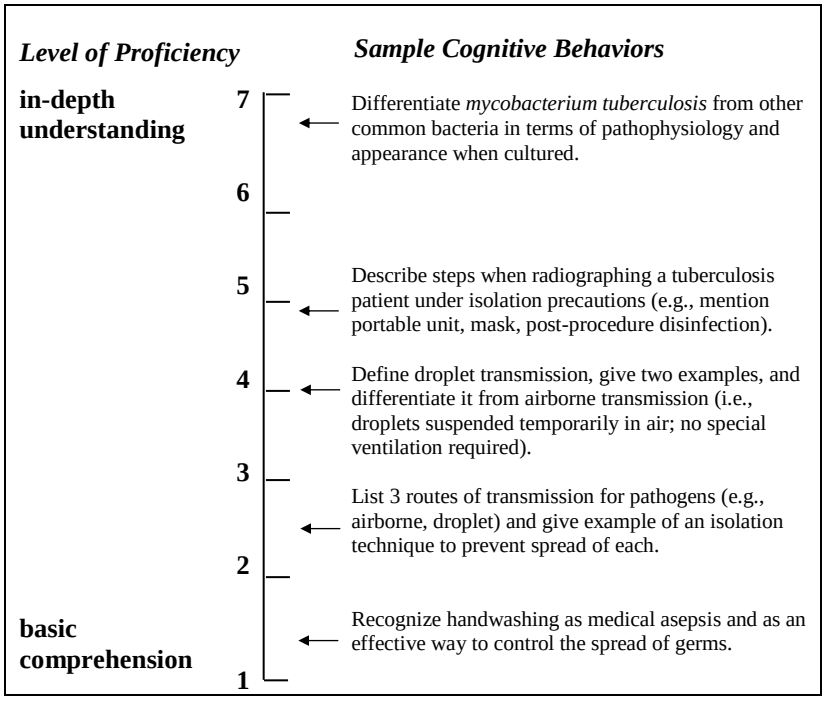

Figure 2: Construct Map for Radiographic Positioning

Positioning: Shoulder. Careful and precise radiographic positioning is required to ensure high-quality images and accurate interpretation, while minimizing dose and the need for repeat examination. This requires knowledge of anatomy, physiology, topographic landmarks; optimal the path of the central ray; use of immobilization devices or other techniques (e.g., breathing); and correct specification of technical factors, including modifications for body habitus, trauma, pathology, or other circumstances. The scale below lists *sample* behaviors corresponding to different levels of skill related to positioning of the shoulder.