

On the use of Rasch modeling in software simulations of computer system  
administration labs

Elizabeth Dalton  
Boston College

IT (Information Technology) industry certifications are offered by multiple vendors. The purpose of such certifications is to help organizations ensure the skills of their employees, who may be performing tasks of considerable financial criticality to the organization. Operating system administration is one example of such a critical task. The Standish Group reports that unplanned downtime of a system can cost thousands of dollars per minute (Murphy, 1999), The Gartner Group reports that 40% of unplanned downtime is due to operator error (Kovar, 2001), much of which could presumably be avoided if the operators had sufficient skill. One IDC report states, "Certified staff is 70% more productive than uncertified staff." (Anderson & Cushing, 2002). IT industry certification is one of the most frequently used ways organizations have of trying to ensure that their operators have sufficient skill for the tasks at hand.

Most current IT industry certifications rely heavily on multiple choice items, with some use of fill-in-blank items. A few vendors, however, have begun to experiment with the use of performance-based measures, in the form of proctored labs (exercises in which examinees perform IT tasks using actual equipment and software, and are rated by trained proctors using an instrument), remote labs (exercises in which examinees interact across the network with actual equipment and software, which may be scored by a proctor or by automated instruments), and software simulations. Software simulations involve a representation of the hardware and software used in the tasks being assessed, provided to the examinee at a testing site or via the internet. Software simulations can be expensive to create, but avoid several known problems of proctored labs, such as the expense of specialized hardware, the need for the examinee to travel to the test site, and the difficulty in training proctors to provide reliable ratings. Software simulations can also avoid the

network latency and stability issues which plague remote labs, and are more easily monitored or instrumented, and therefore more easily scored than remote labs.

Performance-based measures in general, and software simulations in particular, are an attractive alternative to existing multiple choice measures for a number of reasons:

- They promise higher concurrent validity, a form of criterion validity (Shann, 2000).
- They offer higher value to the customer/examinee due to authenticity and perceived relatedness to "real world" tasks. Although face validity is not a technical issue, it is a customer acceptance issue. The value of face validity is accepted as a motivational factor (Crocker and Algina, 1986).
- They are more resistant to cheating, due to the complexity of response required by the examinee. Cheating on IT certifications through the use of "braindump" sites has become endemic (Smith, 2004). Cheating costs the developing vendor money in the need for continual item refresh, cheapens the value of the certification (and all IT certifications), and ultimately lowers certification sales.

However, there are a number of psychometric challenges to making the switch to simulated labs in a certification situation. These challenges are largely typical to performance assessments generally, although the use of automated scoring in simulations avoids the issue of inter-rater reliability, a common hurdle in performance assessments. (Software and hardware labs, unlike more creative tasks such as essay writing, are relatively easily scored using automated means, though there are special challenges to scoring sub-tasks, as noted below.) Challenges include legal defensibility, the difficulty in equating the new exams to existing exams, the need to demonstrate the validity of the new exams, reliability concerns, unidimensionality concerns, issues of partial scoring and

local independence, and the desire to provide sub-section scoring comparable to existing certification processes. Each of these challenges will be discussed in some detail below.

For IT industry certifications to defend against legal challenges, certification vendors take a number of important steps which will need to be addressed in any performance-based variants. Exam items go through a rigorous process of technical and psychometric review and beta testing, during which formal records are kept. All production examinee responses are tracked and stored for an extended period. Comparable procedures will need to be developed for simulation items. This is not expected to be a technically difficult task. (On the contrary, because options will not need to be reviewed, only prompts, this process is likely to be somewhat simpler than the process currently used for multiple-choice items, once developed.)

The issue of equating to existing certifications is more difficult. Should performance tasks in the simulation be equated to items, subtests, or whole tests? How will simulations compare to existing certifications in production? Simulations, which do not facilitate examinee guessing, may be more difficult than multiple-choice items covering the same topics. What effect should this have on the cut score? And given the complaints about the validity of existing certifications, is it even desirable to equate the new instruments to the existing exams? These are policy questions which will need to be addressed by the vendor, based at least in part on marketplace considerations, e.g. the “brand strength” of the existing certifications.

Reliability is likely to pose some particular challenges. As with essay questions, simulation tasks are likely to take longer to complete than multiple-choice items. This means fewer tasks are likely to be assigned to each examinee, potentially reducing

reliability and also raising questions of content coverage. Sub-task scoring might raise the number of items and therefore reliability again, but carries its own considerable challenges (see below). Regardless of the position taken on item count and reliability, it will be necessary to carefully design each task to ensure adequate coverage of critical skills. The care used in this sampling will also affect content validity (McNamara, 1996). Wallace Judd, of the Performance Testing Council, suggests it may also be of value to develop an equivalent of the Guttman Coefficient of Reproducibility to aid in producing reliability estimates (private communication, April 6, 2004).

Judd also suggest that unidimensionality may be harder to achieve or prove with performance assessments (private communication, April 6, 2004). Russ Smith, a psychometrician familiar with the IT certifications of one major vendor has written that they are of narrow domain and load heavily on a single factor (private communication, April 4, 2004), but independent factor analysis of some of the data available from one specific exam of those reviewed by Smith has turned up troubling questions in this area, with extremely low loading on only one major factor (Dalton, 2004). This raises concerns about the unidimensionality of the construct(s) in question, which may make it difficult to use IRT methods to improve the measurement characteristics of simulations covering the same domain.

The unidimensionality assumption can be tested using a variety of methods (Hambleton, Swaminathan & Rogers, 1991), and McNamara (1996) reports that IRT methods are robust with respect to the unidimensionality assumption, also stressing that “measurement” unidimensionality need not be the same as “psychological” unidimensionality. Faceted models (discussed in greater detail below) may also be

helpful in identifying and analyzing multiple influences on item and person performance (McNamara, 1996).

One of the most difficult questions is that of how to score labs so as to reflect the underlying skills represented by the tasks. A given lab procedure may involve using several independent skills (each corresponding to sub-tasks, or steps). It is desirable to track these skills separately and provide some specific skill-based scoring, to provide similarity to current certifications, which provide sub-section scoring, and to provide specific feedback and remediation advice to the examinee. Should these skills be scored per lab using a partial credit model? This would, however, require ordinality of the skills being scored in the lab, which may not be the case, as many tasks consist of several steps which may be performed independently of one another, each involving discrete skills. Additionally, many tasks may be correctly completed by using any of a variety of different sets of sub-tasks, and the intent of the certification process, as a summative assessment, is to focus on outcome, rather than the particular methods of completing the tasks chosen by the examinee. We want to be careful not to look for unnecessary intermediate steps whose absence would preclude detecting an unanticipated, but correct solution. We can attempt instead to score on “outcome factors,” separate measurable elements of the desired outcome, to avoid this last problem, but there is still the issue that some outcome factors will not be ordinal, i.e. they are independent. This would make using the partial credit model based on performance of specific sub-tasks or identification of specific outcome factors difficult or impossible. But to score each of the lab steps as separate items would require that each step or factor be entirely locally independent, which is clearly not the case, as some lab steps would only be able to be completed if a

previous step was completed successfully, and the sub-solution resulting in some factors may be dependent on sub-solutions involving other factors. The issue, then, is that these performance tasks are likely to be a mix of locally dependent and independent components, amenable to neither a partial credit model nor itemization.

One approach which may merit further investigation is multiple scoring, in which the task as a whole is scored using a holistic partial credit model (in which the partial credit is based on the weighted percentage of total sub-tasks or outcome factors correctly completed, rather than the presence or absence of specific tasks or factors), but individual steps of the task or outcome factors involving specific skills are also separately scored and scores accumulated per skill for separate reporting. This separate reporting would not be combined with the main score for certification status purposes, but would only be used to present additional information to the learner (and possibly a remediation system, the certification development team, and/or other interested parties).

Another approach would involve the use of a faceted Rasch model. In the faceted Rasch model (used in Rasch analysis but not in n-PL models), "items" representing major areas of ability like "Perform User Administration" and "Manage File Systems" would be identified for each simulated task. Not every task would have every item, but the inclusion of each item in at least two tasks could create the necessary linkages to calibrate the facet analysis. We could then use this analysis to more correctly estimate the ability of the examinees based on the difficulty of the "items" being measured per task, and thus provide an estimate of the examinee's ability against that particular facet. Again, this cross-task scoring could be reported separately from overall task scoring, or could be used as an alternate final scoring algorithm. The restriction to the Rasch model here

seems reasonable, as the factor of guessing should be eliminated in a simulation environment, though the magnitude and effect of discrimination characteristics of the simulation items are still unknown.

The primary remaining issue is that of validity. We can estimate the content validity of the simulation by auditing the design against the blueprints used to create current exams, which are developed using a rigorous process of survey and interview of Subject Matter Experts (SMEs) in the field. But criterion validity is still questionable. Since the examinees, once certified, are considered ready to work, the form of validity of most immediate interest is concurrent validity. Performance-based assessments are often claimed to have higher concurrent validity (Shann, 2000) than forced-choice exams, but this claim needs to be confirmed in the context of IT industry certification. Unfortunately, concurrent validity studies have not generally been done on the existing exams, which rules out using comparison to the existing exams as a method of testing concurrent validity. A strong recommendation, then, would be to perform at least one concurrent validity study per exam to be converted, on both existing multiple-choice items and the proposed simulation-based performance items, using as broad a cross-section of IT workers currently employed in relevant positions as possible. A Job Responsibilities Scale (Ludlow, 1999) may be of interest as a possible relatively low-cost component of such a study, to help correlate performance on the items with level of experience and current responsibility in the “real world” of IT. It is hoped that a JRS may be considered less threatening or invasive and more objective than managerial or peer rating or self-rating. The JRS can also be analyzed using a Rasch model, allowing for a common point of comparison between the multiple-choice and performance-based items.

## References

Anderson & Cushing (2002). *Industry Developments and Models The Relationship Between Certifications and Partnership Programs: Alignment is Key* (Report no. 28374).

Bedford, MA: Interactive Data Corporation.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Belmont, CA: Wadsworth.

Dalton, E. (2004). Industry Certification Unidimensionality. Presented at the annual meeting of the New England Educational Research Organization, Portsmouth, NH.

Hambleton, R., Swaminathan, H. & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Productions.

Kovar, J. F. (2001, July 3). High availability means no time for downtime – solution providers balance costs with customers' needs. *Computer Reseller News*.

Ludlow, L. H. (1999). The Structure of the Job Responsibilities Scale: A multimethod analysis. *Educational and Psychological Measurement*, 59, 962-975.

McNamara, T. (1996). *Measuring Second Language Performance*. New York: Addison Wesley Longman.

Murphy, B. (1999, December). Data downtime dilemma. *Software Magazine* 19(3).

Shann, M. H. (2000). *Performance Assessment for Practitioners*. Unpublished manuscript, Boston University, Boston, MA.

Smith, R. W. (2004, April). The Impact of Braindump Sites on Item Exposure and Item Parameter Drift. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.