Graduate Management Admission Council®

Validity and Decision Issues in Selecting a CAT Measurement Model

James B. Olsen
Alpine Testing Solutions
and
C. Victor Bunderson
EduMetrics Institute

Presented at the CAT Models and Monitoring Paper Session, June 7, 2007

Abstract

This paper discusses validity and decision issues that should be addressed in selecting a computerized adaptive testing (CAT) model. The paper begins with a historical perspective on the expected benefits of computer-based testing. It summarizes six computer-based testing innovations by the authors and shows how these innovations extend traditional theory and practice of measurement science andCAT. These six innovations include performance work models, job analysis and synthesis, continuous learning progress pathways, validity-centered design and documentation, and logical measurement opportunities in performance tasks and simulations. The paper then discusses specific decision issues relevant to the selection of a CAT model and provides current research related to those decision issues. Three illustrated examples are provided for using item response theory and CAT elements leading toward a performance testing measurement model. Conclusions and recommendations for future research are presented.

Acknowledgment

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.

Citation

Olsen, J. B. & Bunderson, C. V. (2007). Validity and decision issues in selecting a CAT measurement model. In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved [date] from www.psych.umn.edu/psylabs/CATCentral/

Author Contact

James B. Olsen, Alpine Testing Solutions, 51 West Center Street, #514, Orem, UT 84057, Email: jim.olsen@alpinetesting.com; or C. Victor Bunderson, EduMetrics Institute, 3305 North University Avenue, Suite 250, Provo, UT 84604, Email: vbunderson@edumetrics.org

Validity and Decision Issues in Selecting a CAT Measurement Model

This paper addresses validity and decision issues involved in selecting a computerized adaptive testing (CAT) model and has five primary objectives:

1. Present a historical perspective and a validity centered design and documentation model
2. Highlight the need to generalize the definition of test items in CAT to allow for use of integrated, functional simulation and performance tasks within a CAT system environment.
3. Present a practical approach for the various decisions involved in selecting a CAT testing measurement model.
4. Provide statistical results from three examples of performance testing measurement models.
5. Provide recommendations and conclusions for future research investigation.

Historical Perspectives

In 1988, Samuel Messick, a psychometrician and test score validation theorist, stated that validity is the primary issue for tests delivered by all media (italics have been added by the authors for emphasis to the current paper):

> "Over the next decade or two, computer and audiovisual technology will dramatically change the way individuals learn as well as the way they work. Technology will also have a profound impact on the ways in which knowledge, aptitudes, competencies, and personal qualities are assessed and even conceptualized." He specified that there would be:
>
> 1. "New and more varied *interactive delivery systems* in education and the workplace,
> 2. Heightened *individuality in learning* and thinking,
> 3. Increased premium on *adaptive learning,*
> 4. Heightened emphasis on *individuality in assessment,*
> 5. Increased premium on *adaptive measurement,* perhaps even the *dynamic measurement* of knowledge structures, skill complexes, personal strategies and styles as they interact in performance and as they develop with instruction and experience" p.33.

Sam Messick envisioned that future tests would become more interactive, more adaptive, more individualized, and more dynamic.

Bert F. Green (1970) gave the following statement in his conference presentation, "Comments on Tailored Testing," after a presentation by Fred Lord entitled "Some Test Theory for Tailored Testing" (italics again added for emphasis).

> The computer has barely started to establish itself in the testing business. As experience with computer-controlled tests accumulates, we can expect important changes in the technology of testing. Most of these changes lie in the future. Lord's results, clear-cut and devastating as they are, will in the end seem a minor skirmish in *the inevitable computer conquest of testing*." p. 194

During the early 1970s, David J. Weiss and his graduate students at the University of Minnesota initiated the definition and development of the professional field of computerized adaptive testing (CAT) to refer to a test that is dynamically adjusted or tailored to the examinee after each test item, and in which comparable scores can be computed even if different examinees were administered different sets of items. In a graduate student seminar David J. Weiss called for the relegation of the paper-and-pencil test to a museum (D. J. Weiss, July 3, 2007, personal communication).

Robert L. Brennan, as editor of *Educational Measurement, Fourth Edition* stated in his 2006 essay entitled, "Perspectives on the Evolution and Future of Educational Measurement"

> "In the 20th century perhaps the single most important technological development was E. F. Lindquist's invention of the optical scanner…It is interesting to speculate how Lindquist, the editor of the first edition of *Educational Measurement*, would perceive the influence of his invention. My own guess is that he would strongly encourage innovative use of the computers, even if doing so resulted in less use of his invention. Whereas, the optical scanner primarily impacts only one aspect of testing, computers have the potential to impact virtually all aspects." p.11.
>
> "… I think that eventually computers will have a major impact on measurement. My own belief is that the role of technology and computers in testing is partly evolutionary and partly revolutionary." p. 11 (see Bunderson, Inouye, & Olsen, 1989; Cohen, 2006, and Drasgow, Luecht & Bennett, 2006).

Bunderson, Inouye, and Olsen (1989), in the Third Edition of *Educational Measurement*, envisioned four generations of computerized educational measurement. These generations were:

1. *Computerized Testing (CT)*: computers are used to automate sequentially delivered tests with static items. "The first, or CT generation, is defined as the translation of existing tests to computerized format, or the development of new, non-adaptive tests that are similar to manually administered tests but utilize computer capabilities for all or most test administration processes." p. 374.
2. *Computerized Adaptive Testing (CAT)*: a computerized test with adaptive delivery of item and/or task sequence, immediate scoring, and adaptive decisions to stop the test. "The second, or CAT, generation of computerized educational measurement is defined as computer-administered tests in which the presentation of the next task, or the decision to stop, is adaptive. A task can be an item or a more complex standardized situation involving one or more responses. To be adaptive means that the presentation of the next task depends on calculations based on the test taker's performance on previous tasks… Item response theory provides a psychometric foundation for one kind of CAT test, that which adapts primarily on the basis on the item difficulty and/or discrimination parameters. Three types of adaptive tests are possible: *adapting item presentation*, based on item response theory parameters, particularly the item difficulty parameter; *adapting item presentation times*, based on previous response times; and *adapting the content or composition of the item* based on previous choices. In any of these cases a separate adaptive decision can be made: *adapting test length*, based on the consistency of previous performance." p. 381

3. *Continuous Measurement (CM)*: The third, or CM generation "uses calibrated items and tasks to continuously and unobtrusively estimate and profile dynamic changes in examinee proficiency. Tasks measured may be items, item clusters, exercises, unit tests, or independent work assignments. [Note the generalization of what constitutes the measurement administration unit.] Changes may be observed in the amount learned, the proficiency on different tasks, changes in the trajectory through the domain, and the student's profile as a learner. The differentiating characteristic of CM is the ability to specify dynamically a learner's position on the simple and complex scales that define a growth space. Continuous measurement produces a trajectory over time for the individual who is working to master a domain of knowledge and task proficiency. Measurement is accomplished by assessing the performance of each individual on tasks calibrated to serve as milestones of accomplishment. The milestones that make CM possible are embedded into a curriculum, training or education program so that measurement is unobtrusive." p. 387.

   "The definition of the CM generation assumes a two-part definition of a curriculum, training or educational program: (a) a course of experiences laid out to help the learner grow toward certain educational ends, that is a path through the domain; (b) a set of course markers, or standards, that serve as milestones of accomplishment along the way, that is, beginning, intermediate, and terminal markers." p. 387.

4. *Intelligent Measurement (IM)*: The fourth or IM generation "is defined as the application of knowledge-based computing to any of the processes of educational measurement… The knowledge and expertise of measurement professionals can be captured in a computer memory in a symbolic form called a knowledge base. This knowledge can be used to replicate, at multiple sites through a computer or other technology, the expertise of humans, who are otherwise physically restricted to one site at a time. Thus, less expert humans, with the aid of intelligent computing systems, can perform measurement processes that require considerably more knowledge and experience than they presently have." p. 398.

   "Applications of intelligent measurement can be classified within the following areas. *Test Development Processes:* computer tools for job and task analysis with advisor, computer tools for developing test specifications with advisor, and item and test development programs with advisor. *Test Administration Processes*: programs for administration of individually administered tests with advisor to guide the paraprofessional, natural-language-understanding expertise for scoring constructed responses, programs for helping to interpret profiles, and intelligent tutoring within a task when additional practice is needed. *Analysis and Research Processes*: statistical programs with intelligent advisor, intelligent scheduling and calibrating of experimental items, and intelligent data collection for research and validity studies." p. 399

   "Intelligent measurement can use machine intelligence to (a) score complex constructed response involved in items and in reference tasks, (b) generate interpretations based on individual profiles of scores, and (c) provide prescriptive advice during continuous measurement to learners and instructors, to optimize learner and examinee progress through a curriculum." p. 399.

> "Intelligent advice during learning is the most promising contribution of IM for learners and teachers. Its goal is the optimization of learning. It requires a curriculum administered in association with a continuous measurement delivery system. It requires that human expertise be acquired in a computerized knowledge base, analogous to that of the expert counselors who interpret individual profiles of static scores…Intelligent advice during continuous measurement is the epitome of computerized educational measurement. The optimization of learning in a growth space of calibrated educational tasks represents a challenge for educational measurement scientists and practitioners that will require great effort over many years." pp.400-401 "Intelligent measurement will make possible adaptive and intelligent advice based on individual trajectories and learning profiles. Before this goal is achieved machine intelligence will be used to score complex constructed responses automatically and to provide complex interpretations of individual profiles made of static measurements." p. 403.

A Seasoned Response to the Four Educational Measurement Generations

The authors of this earlier chapter have noted that during the last eighteen years the need for improved test security with computerized and online testing and improved measurement models are natural drivers from Generation 1 CT to Generation 2 CAT involving the implementation of expanded and rotated item banks to ensure adequate security of items and tasks, and improved item selection and proficiency estimation approaches. Improved measurement of examinee competence is a natural driver from Generation 2 CAT to Generation 3 CM to help ensure competence by learning and training and not by screening out those who are unqualified. Measurement of generated items, tasks, and work models noted later in this paper is also a natural driver from Generation 2 CAT to Generation 3 CM. The administration of dynamic simulations and use of live application or performance environments can be viewed as a Generation 3 CM or Generation 4 IM form of adapting the delivery presentation and the score processing procedures in response to dynamic changes in the simulated or live application system and performance tasks. The movement from administering and scoring discrete academic items to more complex item clusters and sets, to testlets and to performance testing is another significant branch of Generation 4 IM educational measurement highlighted in this paper.

Six Testing Innovations Provide the Foundation of Our Work

Six testing innovations and their respective creation dates provide the foundation for our work presented in this paper. These six innovations are briefly introduced below:

1. Performance Work Models (1981)
2. Job Analysis & Synthesis (1981)
3. Logical Measurement Opportunities (1999-2000)
4. Continuous Learning Progress Pathways (2002)
5. Validity Centered Design & Documentation (2003)
6. Logical Measurement Opportunities within Performance Tasks and Simulations (2003)

Performance Work Models (1981)

A Performance Work Model is defined as an integrated exercise that allows replication of both information and interactions. It may correspond to one or a group of scaled performance

tasks. Performance work models provide for generalization of tasks and items into meaningful integrated measurement units. The examinee/evaluator can assess different levels of the quality of the performance. There is built-in feedback of results to examinee/evaluator. The performance work model is focused on providing realistic experience with emphasis on fidelity of the model to the actual work domain.

Performance Work Models are reviewed in two papers by Bunderson, Gibbons, Olsen, and Kearsley (1981) and Gibbons, Bunderson, Olsen, and Robertson (1995). Gibbons and Fairweather (1998) provided a discussion of how performance work models help solve the problem of instructional and assessment fragmentation vs. integration. In the abstract of the original paper defining performance work models the authors noted, “While instructional objectives have provided a cornerstone for the practice and science of instruction, they have also locked us into a lexically based conceptual system. In order to realize the interactive potential of computer-based instructional systems, we need a new way of representing performance. This paper presents the concept of a work model which is a unit of practice which allows replication of both information and interaction. In addition, the work model idea also addresses some of the fundamental problems with objectives such as their inability to capture the richness of terminal behaviors or how to relate objectives to content.” (Bunderson, Gibbons, Olsen & Kearsley, 1981, p. 205, paper abstract)

The concept of a work model implies that the teaching or assessment system should provide “working models” in which the learner or examinee can perform. A work model can be defined as a single integrated unit of practice and performance. It may correspond to one or a group of performance tasks. Work models provide settings in which the learner can converse using the new vocabulary and concepts, perform the new procedures, and make predictions and solve new problems. These work models will have visible results so that the learner, and sometimes other learners and the teacher or supervisor who are observing, can obtain information about the success or failure of the performances.

After a fourteen-year period Gibbons, Bunderson, Olsen and Robertson (1995) reviewed work models and noted that work models were still beyond objectives. “Not only does [the work model] construct encourage designers of computer-based instructional systems to create syllabi with uniquely integrative and performance-based qualities, but it encourages also the construction of families of increasingly complex microworlds that challenge traditional views of the syllabus and curriculum.” p. 221, paper abstract

We refer to these job tasks represented in functional contexts as "work models or performance models of work." A performance work model is defined as a representation of the essential, integrated performance situations in a domain of expertise that well-qualified individuals are able to perform to a high standard. Essentially, we are trying to measure what individuals “can do” as well as what they “know” and to measure these critical performances with carefully structured task environments with realistic job situations and integrated and increasingly more complex job tasks.

Job Analysis and Synthesis for Performance Task Development (1981)

The standard job analysis or task analysis that is performed in a domain of expertise typically identifies a series of major tasks or practices that are performed in the given domain. Each task or performance is analyzed to determine the required subtasks, and the associated knowledge, skills, and behaviors that are required to perform that task. The job or practice analysis breaks the job down into elementary units in an analytic fashion. What is needed after the job and task analysis process is a synthesis process that collects and clusters the work elements into meaningful, integrated and critical worthwhile tasks. In most work

settings a given subtask is not performed separately but is often part of an integrated work flow process that produces an integrated work output. Without the synthesis process the job and task analysis often lacks meaningful context and tasks are often fragmented into minute, unrelated job elements and subtasks rather than meaningful integrated work flow units.

The implementation of performance testing requires principled task design models for identifying key job performance elements and their relationships and structuring them into meaningful, integrated job like performance situations. These performance tasks and situations allow for assessing realistic and dynamic problem-solving, troubleshooting, and diagnostic-reasoning skills.

Following are several identifiable characteristics of performance assessment tasks:

1. Employs constructed actions or response series.
2. Assesses actions and responses directly within context.
3. Referenced to criterion or standards of quality.
4. Focuses on the process of solution as much as the result.
5. Involves judgments of qualified performers in determining test scoring.
6. Performers understand criteria, process, and products on which they will be judged.
7. Used for individual or group measurement.

When performance tasks are created, the task of item writing is significantly altered from traditional processes of test item writing. Authoring performance tasks includes a set of entering arguments (variables) that are placed into a typical scenario, case study, or task environment. One performance task can typically cover the subject domain areas for which five or more traditional items would attempt to predict mastery. With performance tasks the technical review/edit process involves subject matter experts performing the tasks and creating a useful scoring model for alternative paths or end state actions within the modeled system or performance environment.

Logical Measurement Opportunities (1999-2000)

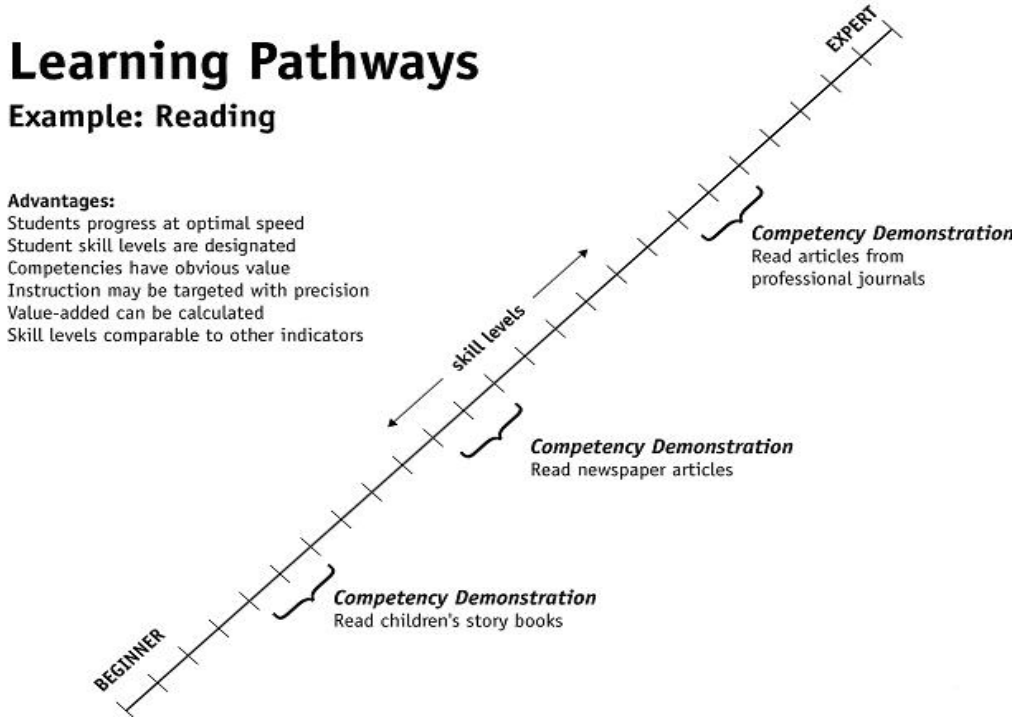
As we began using more complex test designs and innovative testing item formats, we found the need to introduce the concept of logical measurement opportunities. Logical measurement opportunities are logical combinations of answer selections or constructed actions that can be evaluated according to subject matter expertise in deciding whether or not the logical combination of selected and constructed results is true or false, or correctly answers the question or task at hand. Logical measurement opportunities allow for measuring results like the following examples: “A and C and B and D but not E, F, G, or I” or correct actions for “subtasks A, C, and G but not subtasks B, D, E, and F”. The introduction of logical measurement opportunities allowed for the assessment designer to determine how a general performance or work model task could be evaluated by using a series of logical actions that occur through interactions with the performance or work model task.

Continuous Learning Progress Pathways (2002)

In 2002 the authors were asked to propose a competency based system for statewide instruction and assessment. In this analysis we determined that the educational standards and objectives across the grades were not articulated or correlated well to allow for effective measurement within and across grades. At that time we proposed the development of

continuous learning progress pathways with designated competency demonstrations as markers across the continuous increasing skill levels from beginner to expert within a content domain. As an example, we found that the competency for learning to tell time was spread across a five grade curriculum span rather than taught as an integrated competency with extensions. Figure 1 provides an illustration of a sample learning progress pathway in reading. This pathway would allow students to progress along the pathway at optimum speed and direction. The competency demonstrations along the pathway indicate examinee skill levels at differing segments of the pathway. For example, a learner near the beginning of the pathway could read children's story books, a learner in the middle could read newspaper articles, and a learner at the upper end could read articles from professional journals. Increasing skill levels along the learning progress pathway are evaluated with competency demonstration assessments.

Figure 1. Learning Progress Pathways



Many curriculum and training programs exhibit gaps in content and skill competencies where learners are apt to fall behind and do not have sufficient chance to practice and catch up. These curriculum and training programs are often not continuously measured. In contrast, a learning progress pathway would have competencies defined with smoothly increasingly levels of difficulty, progress feedback is provided on valid measures all along the pathway, practice with feedback is provided across the pathway, and learner progress is easy to interpret by reporting competencies completed and competencies in progress. With a learning progress pathway, measures can be created that sample skill competencies from the domain appropriate to each segment of the achievement pathway. At the present time, the authors are not aware of any state or provincial assessment system that is built on competency demonstrations within and across the educational grades. Nor are we aware of any professional certification and licensure testing system that is based on a learning progress pathway system.

In his paper on performance assessment, Samuel Messick (1994) identified three key questions that should be asked as we develop all types of tests and assessments.

1. What complex of *knowledge, skills and abilities* should be assessed?
2. What *behaviors or performances* will reveal the constructs and skills to be tested?
3. What *tasks or situations* should elicit those behaviors?

Messick's advice is very useful in helping curriculum and training designers and assessment specialists to envision realistic competence-focused learning progress pathways with benchmark or reference tasks that illustrate the knowledge, skill, and ability complexes that should be assessed at differing locations across the learning progress pathway.

Validity-Centered Design and Documentation (Bunderson, 2003, Olsen, 2006)

Validity-Centered Design and Documentation (VCDD) is the beginning of a principled design process for designing and developing improved learning theories of progressive attainments or competence in specific domains of expertise. These domain-specific learning theories are often referred to as "domain theories" among those using validity-centered design. VCDD is also used to develop the construct-linked measurement scales associated with each domain theory and to document evidence for a validity argument for the domain. The validity argument is not accomplished all at once (and indeed, never ends), but is improved step-by-step as we complete work on each respective aspect of validity. VCDD also includes planning for future activities to improve other aspects of the validity argument in an ongoing process. In essence, the way we measure learning progress must be validated. Because validating scales in a social setting where concepts, ideas, etc., change continuously, the measurement process must be validated continuously.

Validity-centered design is focused on designing and implementing the ideas of Samuel Messick on validity in educational and online educational and assessment environments. Messick (1988, 1989a, 1989b, 1998a, 1998b) has been perhaps the most influential validity theorist of the past 15 years. He developed the unified validity framework that showed that construct validity is the central, unifying concept among the variety of different views or perspectives on validity. Construct validity deals with the invisible traits or constructs that intelligent observers have formulated and "constructed" in words, diagrams, and so forth; how these invisible constructs are made visible through responses to items and performance situations; and how these responses are turned into meaningful scores. Construct validity is the link between the invisible theoretical ideas about important human qualities (sometimes called "latent traits") and the scores on some instrument or measurement procedure designed to produce numbers reflecting differences in the unobservable human qualities. These numbers represent more or less of the latent trait or construct in question.

Validity-centered design and documentation is a set of methods and tools used at each of several stages of a design process to develop a learning progress measurement system, to implement it in a computer or online environment, and to keep continuously improving it. The learning progress measurement system is used by instructors and learners, and the measurement system is improved based on both qualitative and quantitative results. Data are collected at each of several cycles of implementation and during the design process itself. Over time, data and documentation provide an increasingly strong "validity argument" for the quality of the learning progress measurement system. The idea is that validity cannot be proven once and for all, but that evidence and argument threads can be assembled to show how well a given learning progress measurement system, when used in certain ways, meets the multifaceted ideal of the unified validity model. Validity is much more complex and unified than usually understood. Validity-centered design identifies nine different but

interrelated aspects of validity. These incorporate the six aspects of construct validity identified by Messick (1998, 1989b): (1) content, (2) substantive processes, (3) structure, (4) generalizability, (5) external, and (6) consequential.

Validity-centered design and documentation restructures these six aspects and then adds three new aspects: (7) overall appeal, (8) usability, and (9) value and positive consequences. A comprehensive validity argument can be organized around these nine integrated aspects of validity. Table 1 shows the nine validity aspects structured into three primary classifications, each with three interrelated elements. Each of the aspects and elements of Table 1 is further discussed below.

Table 1. Validity-Centered Design Elements for Assessment Systems

I. Design for Usability, Appeal and Positive Expectations (User-Centered Design)
- A. Overall Appeal
- B. Usability
- C. User values and Positive expectations

II. Design for Inherent Construct Validity
- A. Content
- B. Substantive thinking processes
- C. Structural (number and meaning of dimensions)

III. Design for Criterion-Related Validity
- A. Generalizability
- B. External Validity (convergent/discriminant)
- C. Consequential (positive and negative)

Validity-centered design and documentation aspires to do more than guide the design of assessments, although it is well adapted to do this. A learning progress system includes a *measurement* system in context with an instructional system, evaluation system, and implementation system. This integrated system is used for ongoing, cyclical evaluation of not only the learners, but also the measurement system itself; the training and instructional materials delivered on the same computers as the measurement system; the adaptive research system that includes adaptation to individual differences; and the strategic implementation of learning and content management systems. The learning-progress measures are part of this comprehensive system that integrates measurement with instruction, but are not all of it.

To accomplish this, validity-centered design leads to an interpretive framework with domain-spanning unidimensional scales for monitoring and measuring learning progress.

I. Design for usability, appeal, and positive expectations. This aspect of validity has the first and highest priority in view of those who will be using it and who will be influenced by it most. This category of design is not enough, as the instrument and its theory will fail unless the basic core of inherent construct validity is also considered from the first. This priority is generally required before organizations will invest in creating some new measurement instrument. Activities leading to this aspect of validity are often found in treatments of user-centered design. Common characteristics are the following:

1. Overall appeal.
2. Usability. The instrument must be easy to use, understandable, quick and efficient.

3. Perceived value to the target users, perceived positive consequences.

Design for appeal and usability can establish superficial face validity, but in order for users to continue to perceive true value and positive consequences, there must be a strong foundation of inherent construct validity. Without a good blueprint, and continued improvement (as established in category II below), the instrument might be so off-target that the perceived value cannot be achieved, nor will positive consequences occur. Users quickly notice when real value is not forthcoming.

II. Design for inherent construct validity. Construct validity is the link between reality and the scores or measures produced by an instrument. This aspect of validity starts with the blueprint. Are we measuring important but invisible thinking processes related to the valued human practices that are hypothesized to exist in users? Do the scales we construct through scoring questions connect with important aspects of reality? Construct validity is the technical, statistical component of validity that, while it will be conducted by the psychometricians, its results will continue to improve upon the measurement and analysis methods. We use data results of our measurement and analysis methods to continually improve the instrument. There are three aspects to inherent construct validity.

1. Content coverage and appropriateness
2. Substantive thinking processes—the important but typically invisible mental processes used by those whom we would wish to score as more successful on an instrument, or affective attributes of persons such as their beliefs, attitudes, and values. It is only through theories of the cognitive, linguistic, affective, or perhaps psychomotor processes, that we can design appropriate questions or performance tasks to get at different degrees of these usually invisible thinking and style processes.
3. Structure of the constructs. The starting number of questions or tasks is expected to collapse into a smaller number of separate unidimenstional measurement scales. The scales we design should correspond with a hypothesized, then increasingly validated domain structure.

III. Design for reliability and for evidence of criterion-related validity. This aspect of validity is attained through analyzing the data from using the instrument—along with other measures. Except for reliability and generalizability to different groups of examinees, scores from the instrument must be correlated with other measures—other instruments and outcome criteria.

1. Generalizability. Evidence that the scoring methods and scores are reliable, and generalize to different occasions, settings, genders, racial groups, national groups, etc., both within the same institution and from institution to institution.
2. External. Evidence that the scores predict other valid criteria for what is being measured. External validity examines evidence that other instruments correlate and do not correlate as would be expected by the nature of their constructs.
3. Consequential. Evidence that positive results (consequences) do occur over time, and that unexpected but negative consequences do not occur over time. This is an extension of the perceived positive consequences listed under category I, above. In this aspect of validity, we obtain evidence of the actual occurrence of such positive or negative consequences.

Validity-centered design is a work in process. A familiar discipline with design in its title is experimental design. It is well established across many social science disciplines and is broadly interdisciplinary. Validity-centered design, although new, is interdisciplinary as well,

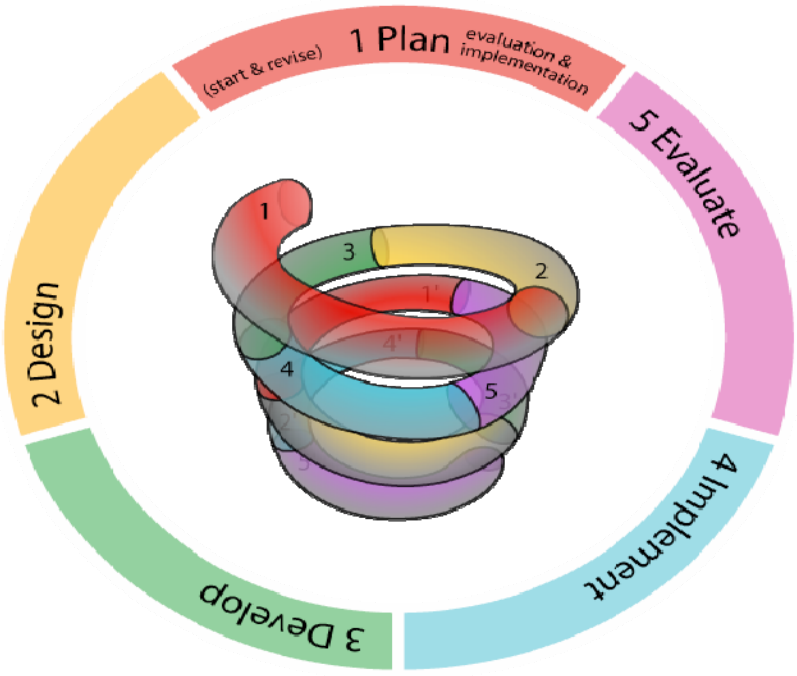
and reaches out not only to tough-minded experimental logic, but also encompasses ways to design and develop at least two sets of artifacts: learning and training and instructional materials in domains of interest, and measurement systems for measuring progress in the same domains.

An indication that the disciplines of design are entering education more broadly is a fairly new approach to educational assessment that uses design in its title, "Evidence-Centered Design" (ECD). While VCDD looks at many aspects of a total system that integrates learning and assessment, ECD is more sharply focused on the assessment of individuals. It is more narrowly aimed at category II, and secondarily at category III above of the validity argument for educational assessments (Mislevy, Steinberg, & Almond, 1999, 2003, Almond, Steinberg, & Mislevy, 2003). VCDD benefits from the excellent work of ECD and builds upon it in developing better student assessments and learning experiences.

Validity-centered design includes both a design and development process and a documentation process. The design and development process provides a focus on the instrument content and improvements, the tasks and questions created to measure the content, the scoring algorithms, test administration procedures, and preparation of user-centered score interpretive materials, and preparing evaluation plans for the next improvement cycle. The documentation process provides the validity evidence argument with documentation on the design for usability, appeal and positive consequences, design for internal construct validity, and design for external criterion-related validity.

Figure 2 shows that validity-centered design is an ongoing, cyclical, and improving process. It begins with a planning process, moves to a design process, to a development process, to an implementation process, and then to an evaluation process, and then a revised planning process and the cycles continue.

Figure 2. The Validity Centered Design Process



Logical Measurement Opportunities within Performance Tasks and Simulations (1997)

Over the last decade the authors have been investigating the use of logical measurement opportunities within performance tasks and simulations. Subject matter experts and psychometricians can jointly determine the appropriate scope and sequence of performance tasks, the defining of logical scoring elements within the performance tasks, and the appropriate weighting and aggregation of the logical scoring elements within and across tasks. These activities require interdisciplinary collaboration and research between the disciplines. In 1984 the first author visited a full flight simulation development organization and discussed the use of psychometric models for the development of computerized adaptive simulation and live performance tasks.

Computer-Based and Adaptive Testing Reference Sources

This section of the paper discusses design and implementation decisions involved in selecting an appropriate model for CAT. We recommend that the adaptive testing models need to be general enough to accommodate integrated performance and simulation tasks derived from performance work models. There are several quality reference sources for innovative computerized and computerized adaptive testing systems: Bunderson, Inouye and Olsen, 1989; Cohen & Wollack, 2006; Drasgow, Luecht & Bennett, 2006; van der Linden and Glas, 2000; Williamson, Mislevy, & Bejar, 2006; Wainer, Bradlow & Wang, 2007, van der Linden & Hambleton, 1997, and van der Linden, 2005.

Evaluate and Select a CAT Item Calibration Model

A key decision in implementing CAT is evaluating and selecting an appropriate CAT item calibration model. The CAT item calibration model selected forms the foundation for the measurement of examinee trait levels and the interpretation of scores from the test. We recommend use of an item response theory (IRT) measurement model. IRT employs a theoretical model for relating the examinees performance success on a test item across the continuum range of examinee traits.

IRT is based on the following four assumptions (Hambleton & Swaminathan, 1985). (Italics have been added for emphasis).

1. "*It is assumed that a set of k latent traits or abilities underlie examinee performance on a set of items*. The *k* latent traits define a *k*-dimensional latent space, with each examinee's location in the latent space being determined by the examinee's position on each latent trait." p. 16. Research is currently underway on multidimensional item response theory models but most useful application work concerns applications of unidimensional models of ability or proficiency.
2. "*There is an assumption equivalent to the assumption of unidimensionality known as the assumption of local independence*. This assumption states that an examinee's responses to different items in a test are statistically independent. For this assumption to be true, an examinee's performance on one item must not affect, either for better or worse, his or her performance to any other items in the test." pp.22-23.
3. "*An item characteristic curve [item response function] is a mathematical function that relates the probability of success on an item to the ability measured by the item set or test that contains it*. In simple terms, it is the nonlinear regression function of item score on the trait or ability measured by the test." p. 25.

4. "An implicit assumption of all commonly used item response models is that the *tests to which the models are fit are not administered under speeded conditions*. That is, examinees who fail to answer test items do so because of limited ability and not because they failed to reach test items." p. 30.

IRT is a statistical approach to measuring test scores that attempts to model the statistical characteristics of items and then aggregate the results on individual items to form the test score. IRT develops a theoretical item characteristic curve or item response function that shows the probability of successful response to the item based on each of several different levels of a trait. Different item response models are defined based on the number of parameters that are used in estimating the shape of the theoretical item response functions. Examinee test scores computed under IRT explicitly involve weighting of items by the item difficulty or other associated item response parameters such as item discrimination or lack of model fit.

IRT models are defined by the number of statistical parameters estimated for the measurement model. The one-parameter (Rasch) model is used for items that differ in difficulty (*b* parameter), a common numerical value is often assumed for the item discrimination, and no parameter is estimated to accommodate guessing.

The two-parameter IRT model is a generalization of the one-parameter model in which items can differ in item difficulty (*b*), items can also differ in discrimination (*a* parameter), and the pseudo-guessing parameter (*c*) is set at a common value or assumed to be 0.

The three-parameter IRT model is also a generalization of the two-parameter model in which items can differ in item difficulty (*b*), items can also differ in discrimination (*a*) and items can also differ in the pseudo-guessing parameter **(***c*).

Investigating Model Fit

The investigation of the fit of the empirical data to the theoretical measurement model is a key decision in selecting and implementing a CAT model. If the empirical data show that the item or item group does not fit the measurement model then the item or item group should be examined and possibly eliminated from the test. To assist the psychometrician with evaluations of model fit, the various IRT calibration programs provide various indices of model data fit. For the Rasch model, "outfit" and "infit" item statistics are provided by the WINSTEPS program (Linacre & Wright, 1991-2000). The outfit item statistic is an outlier fit statistic. The statistic is sensitive to unexpected examinee score patterns on items that are very easy or very difficult. The infit statistic is an inlier pattern sensitive fit statistic. The infit statistic is sensitive to unexpected examinee score patterns for persons with items that are targeted near their trait level. Model fit with the one- parameter model can be examined with the item parameter standard errors for the Rasch measure (*b*) values, comparing empirical item information functions with expected information functions, examination of item residuals after fitting a principal components analysis and extracting the item difficulty values, and the magnitude of the person and item reliability computations.

For the two- and three-parameter models, investigations of model fit can be evaluated by comparing the relative sizes of the item parameter standard errors for the *a*, *b*, and *c* parameters. The item calibration programs XCALIBRE (Assessment Systems Corporation,1996), BILOG-MG (Zimowski, Muraki, Mislevy & Bock, 1996, 2003), and MULTILOG (Thissen, 1991, 2003) provide item response functions and item information functions, test information and test standard error functions, comparison of the empirical and theoretical item response functions, plots of answer option response functions for multiple-choice or polytomous score models, and examination of items that are flagged by chi-square

goodness-of-fit indices. A widely used test for comparing alternative measurement models is the log-likelihood test computed by comparing -2 times the log likelihood for the various measurement test models (one parameter, two parameter, three parameter). Various reliability indices are available to determine the consistency or predictability of examinee scores. These indices include the alpha item reliability, empirical score pattern reliability, and the estimated reliability from the test information curves.

The CAT Process

The CAT process includes five major process steps:

1. Estimate initial trait level (θ) for an examinee.
2. Select an item or item set from an item bank or information matrix at the estimated θ level.
3. Use the selected IRT CAT model to update the θ estimate and its standard error.
4. Select and administer a new item or item set at the updated θ estimate.
5. Evaluate the selected test termination criterion.

Estimating initial examinee θ. To initiate a CAT, the initial θ for the examinee must be estimated. One widely used approach is to select the mean or average θ of the calibration population. With standard IRT, this mean value is typically at 0.0 or the middle of the calibration population distribution. A short "locator test" can be administered with 4-5 items selected from differing difficulty levels across the θ domain and the score from the locator test used to estimate the examinee's initial θ. If information is available from previous exam levels or available collateral information provides information regarding group identity information, this information can be used to determine the examinee's initial θ. To ensure that test items from the item bank near the center of the score distribution are not overexposed or underexposed, the initial θ can be selected by using a random θ location near the mean θ of the calibration population.

Item(s) selection procedure. With the initial θ estimate specified, an item is selected from an item bank to match the initial examinee θ estimate. The test items which provide maximum information at the examinee θ estimate are included in an item selection set. The test items are often selected from an item information matrix that provides an ordered set of the most informative items at 20 or more θ values. The information matrix is typically on the order of 20 or more items at each of 20 or more θ levels. The θ levels can range from -4.00 to +4.00, or less extreme boundary values, in increments of 0.125. This item selection procedure is often defined as *maximum information item selection*. The optimum item for selection is the item that has not been selected yet that provides the most information at the current θ estimate. The optimum item will be matched to the examinee's estimated θ level and provides the most information at that θ level.

A generalization that we would like to introduce is to select an item set rather than a specific test item. The item set can include one or more items in a structured administration unit. In 1987, Wainer and Kiely proposed the definition of testlet as a packet or group of items that are administered together and thus carry their context with them. Testlets can be items associated with a given textual passage or visual diagram. Testlets can also be small networked clusters of structured items. In their seminal manuscript on testlet response theory Wainer, Badlow and Wang (2007) noted,

> "The key idea is to use a multi-item testlet as the fundamental unit of test construction and test administration. We shall define a *testlet* as a group of items that may be developed as a single unit that is meant to be administered together. Although the path through a testlet could be branched, our focus at least for now, is on linear testlets that contain *n* items, where *n* could be as few as one, but more typically would have four or more items. All examinees are presented with a particular testlet would be confronted with the same items in the same order." p. 52-53.

A testlet cluster of items can be selected with testlet parameters that are closest to the current ability estimate. We recommend an additional elaboration of the testlet concept to refer to a series of logical scoring opportunities within performance tasks tied to synthesized work models. These scoring opportunities can be defined by subject matter experts within the given performance domains. These logical scoring opportunities from performances might become the set or group of test items within a performance testlet. The testlet approach allows for accommodating the lack of local independence among a set of performance tasks where one step leads to other related or linked steps.

The maximum information item selection procedure leads to predictable sets of items selected at each of the various ability values. Hence, strategies have been adopted to reduce item exposure by selecting a satisfactory item or item set randomly (i.e., select one of 5, then select 1 of 3 then 1 of 2) from a designated set of most informative items or item sets.

Chang and Ying (1997, 1999, 2001) suggested stratifying the item bank by discrimination values and then selecting an item from designated strata of discrimination as the adaptive testing process continues. The item bank is partitioned by the item discrimination index and then by fixed size strata within the discrimination order. Several items are selected from each stratum for possible administration. The item selection algorithm proceeds first with lower discriminating items and moves to higher discriminating items as the test progresses**.** This process insures that the more discriminating items are used at a later stage of the adaptive testing process when they are more beneficial in increasing precision of the ability estimate rather than used early in the adaptive testing process when the ability level is fluctuating more and the most discriminating items are administered too soon.

Wim van der Linden (2000, 2005) has proposed a shadow test approach as a general framework for adaptive testing. van der Linden uses linear programming variables (0 and 1) to address logical constraints on the adaptive testing process. Using the shadow test approach at each stage of the adaptive testing process, one or more complete optimal test designs are selected from the bank that meet all of the desired test and content specifications. From the shadow test(s), the item or item set is selected for administration that is most closely matched to the current ability estimate and provides the maximum information. Again, after the next item one or more completely optimal test designs are selected for the examinee and one item or item set is selected for administration and the process continues.

Item selection and ability estimation. In an adaptive test, the first challenge is selecting the initial items in the test. A new examinee can be assumed to belong to a known distribution with an existing mean and standard deviation. The ability estimation can begin at the distribution mean or a random ability location near the distribution mean. Auxiliary information regarding the examinee can be used to make a better initial prediction. For example, if a spatial visualization test is being administered, a male examinee could be estimated to start with a relatively higher ability estimate than a female examinee. Likewise, in a vocabulary or English usage test a female examinee could be estimated to start with a relatively higher ability estimate than a male examinee. Research has shown that males tend

to have higher spatial visualization ability and females tend to have higher vocabulary and English usage ability(Harris, 1978; McGlone, 1978, 1980; Kimura, 1999; Halpern, 2000).

Information from previous tests within a given test battery can be used to select the first initial item(s) on a new test within the battery (e.g., Brown & Weiss, 1977). Given the information that is available, the best estimate of the ability on a new test is based on the information available from a previous test and the correlation structure among the tests within the battery.

CAT Administration Models

There are four widely used CAT administration models. These are the item-by-item CAT, classification CAT, computerized adaptive multistage testing, and testlet response theory. Each of these four models is briefly introduced below. For additional details on CAT models and technical details on computational formulas and procedures, the reader is referred to five primary books on CAT (Wainer, Dorans, Eignor, Flaugher, Green, Mislevy, Steinberg, & Thissen, 2000; Wainer, Bradlow & Wang, 2007; Sands, Waters, & McBride, 1997; van der Linden & Glas, 2000; van der Linden, 2005).

Item-by-Item CAT. This approach to CAT adapts or tailors the difficulty of the test to each examinee, item-by item. Items are selected to maximize test information and minimize the standard error of the examinee's ability/trait estimate.

Classification CAT. This CAT model (Kingsbury & Weiss, 1979) is based on item or testlet selection. However, the final score is a classification score determining whether the examinee score and confidence band are above or below the designed performance standard. The item bank should have a peaked test information function at the passing standard. Testing proceeds until the θ estimate and confidence band based on user-determined standard errors [±1 standard error (68% confidence limit), ±1.96 standard errors (95% confidence limit), ±2.58 standard errors (99% confidence limit)] is completely below (fail status) or completely above (pass status) the designated performance standard.

Structured computer-adaptive multistage tests. Structured computer-adaptive multistage tests are self-administering adaptive tests based on testlet item structures that are selected into item panels. Each item panel may include four to seven testlets. The testlets are assigned to a particular stage of test administration and to a specific route within the panel, such as easier, moderate, or more difficult. Panels are the unit of administration and scoring. The design for the computer-adaptive multistage test is often referenced by the number of testlets available at each stage of the testing process. For example, a 1-3-3 multistage test design would have one testlet for stage 1, 3 testlets for stage 2 and 3 testlets for stage 3.

Testlet response theory. The testlet is a structured sequence or cluster of items that can be used as the unit for CAT item selection and scoring. Test termination criteria are based on the same exit criteria as item-by-item CAT, but the score values are based on testlet scores rather than item scores. Testlets typically have non-overlapping item structures.

Testlets can include items that are scored as dichotomous or polytomous. Polytomous score models for testlet response theory can use the nominal score model, generalized partial credit model, or graded response model. An additional statistical parameter is added to the model to account for any statistical dependencies among items in testlets. A Bayesian test scoring model is typically used to update a prior Bayesian distribution given information from the testlet performance score or pattern. Markov chain monte carlo methods may be used to simulate the calibration of parameters for item difficulty, item discrimination, item guessing, and common or specific testlet effects.

Within the CAT Exam

Within the CAT exam the testing process proceeds by either selecting the item or item set that maximizes the test information function or selecting the item or item set that minimizes the standard error.

Estimate θ ability and standard error. There are four major procedures used in practice for estimating examinee ability and standard error:

1. Bayesian methods update a distribution by estimating a posterior mean and standard deviation. Bayesian estimation methods have been shown to have an inherent bias toward the population mean.
2. Maximum likelihood. Compute the maximum likelihood of the pattern of item scores. Provide scoring procedures at test initiation so there is at least one correct item and at least one incorrect item.
3. Maximum marginal maximum likelihood. Estimates the mode or mean and standard error of θ using a posterior likelihood distribution based on a set of quadrature points, associated quadrature weights, and a prior distribution with the item parameters for items administered. The maximum a posteriori estimate (MAP) is the most likely value of θ for persons with a given response pattern. The expected a posteriori estimate (EAP) is the average θ value of persons with a given response pattern. (Bock and Aitkin, 1981; Bock & Mislevy, 1982)
4. Weighted maximum likelihood. The weighted maximum likelihood estimation method was introduced by Thomas Warm (1989). This ability estimate maximizes the product of the likelihood function and the square root of the information function. In essence, the likelihood of the ability estimate is weighted by the standard error for the item. The weighted maximum likelihood estimator has been shown to be a relatively unbiased estimator of θ up to a factor of $1/n$ where n is the number of items and to have advantages over the maximum likelihood and Bayesian ability estimation methods in some situations.

Check test termination rule(s). After each test item or item set, the CAT system should check the test termination rules to determine if the CAT should continue or terminate. There are typically four termination rules used in practice:

1. Minimum posterior standard deviation or standard error (<.20 or <.25). The test terminates when a minimum posterior standard deviation or standard error has been met. Typical criteria for this value are standard errors less than or equal to 0.20 or standard errors less than or equal to 0.25.
2. Minimum test information. This termination rule is met when the additional test information from the remaining items near the ability estimate is at a minimum. This means that if additional test items are administered from the expected ability estimate region there is little new information that can be added from each additional item.
3. Fixed number of items. This termination criterion is established when there are policy concerns regarding the administration of differing number of items to different examinees.
4. Combinations of rules. Combinations of the rules defined above or others are possible.

Various test termination criteria can be used depending on psychometric and practical application issues in administering the test. A test can be terminated when a minimum value is reached for the test standard error such as 0.25 or 0.20. The fixed test length termination is reached when a required number of test items is administered. The minimum supplemental test information rule terminates the test when administration of additional test items will not result in differing values for the test information. A minimum change in ability or standard error can also be computed by determining the differences between each successive ability estimate and between the successive standard error value. Various combinations of the test termination criteria can also be specified, such as a minimum standard error of 0.25 and at least 25 items in the test.

Produce CAT Score Report

After completion of the CAT test a report is produced which includes the computation of the final θ ability estimate using either the Bayesian, maximum likelihood, marginal maximum likelihood, or weighted maximum likelihood procedures. The CAT report should provide either normative or decision-referenced interpretations of the test score, provide relevant score interpretative materials, and provide the estimate the standard error of the final ability or proficiency measure.

Example Performance Test IRT Analyses

This section provides three empirical examples of the use of IRT analyses for calibrating performance tests, illustrating the potential for using performance tasks and item clusters with IRT within a CAT environment. The first performance test IRT example illustrates the use of IRT for scoring simulation tasks that are scored dichotomously. The second performance test example illustrates mixed scoring of simulation tasks employing both dichotomous and polytomous scoring. The second example also illustrates the application of the graded response model for scoring and IRT analysis of simulation tasks that are scored on a continuous or partial-credit score scale. The data shown in the second example illustrate the Samejima graded response model but additional calibrations were computed for the partial-credit model and the nominal model implemented in MULTILOG (Thissen, 1991, 2003). The third performance IRT example shows the use and comparison of item parameter values for one- and two-parameter IRT calibrations with four-alternative performance simulation tests, each consisting of multiple performance tasks and scores.

The authors understand that these are only initial steps in the development of a rigorous psychometric measurement foundation for the measurement of integrated performance and simulation tasks that include a series of logical measurement opportunities that can be administered at benchmark points along a learning progress pathway. In the example performance tasks analyzed below, the authors were not participants in the performance task design but were asked to perform the statistical analysis of the results. As psychometricians initiate closer collaboration with developers of performance tasks, simulations and live application performance environments, there is more opportunity to define meaningful logical scoring opportunities and use these logical scoring opportunities for measurement of status and proficiency changes across the spectrum of performances.

Example 1: Using Simulation Task Data for IRT Analyses

Simulation task performance data was obtained from 116 persons who had been assessed with 57 information technology simulation performance tasks. A score of 1 was given if the simulation performance task was completed successfully. A score of 0 was given if the simulation task was not completed successfully. One- and two-parameter IRT analyses of the

simulation task data was computed using the BILOG-MG (Zimowski, Muraki, Mislevy and Bock, 1996, 2003) calibration program for dichotomous performance outcomes. The one-parameter analysis computed a common performance task (*a*) slope parameter of 1.01. Table 2 summarizes the one- and two-parameter IRT analyses including minimum, maximum, mean, and standard error statistics. Table 2 also shows the empirical pattern reliability of 0.91 and 0.90 respectively for the one- and two-parameter calibration models and 0.96 and 0.97 respectively for the information curve reliability from the one- and two-parameter calibration models.

Table 2. IT Simulation Tasks: One- and Two-Parameter IRT Summary

	1-Parameter *a*	1-Parameter *b*	2-Parameter *a*	2-Parameter *b*
Min	1.01	-2.04	0.48	-2.40
Max	1.01	0.79	3.39	0.51
Mean	1.01	-0.49	1.19	-0.49
Median	1.01	-0.49	1.02	-.34
Std Error	N/A	0.73	0.67	0.66
Empirical Reliability		0.91		0.90
Information Reliability		0.96		0.97
-2 Log Likelihood		4104.045		3905.537
Largest Change		0.009		0.009

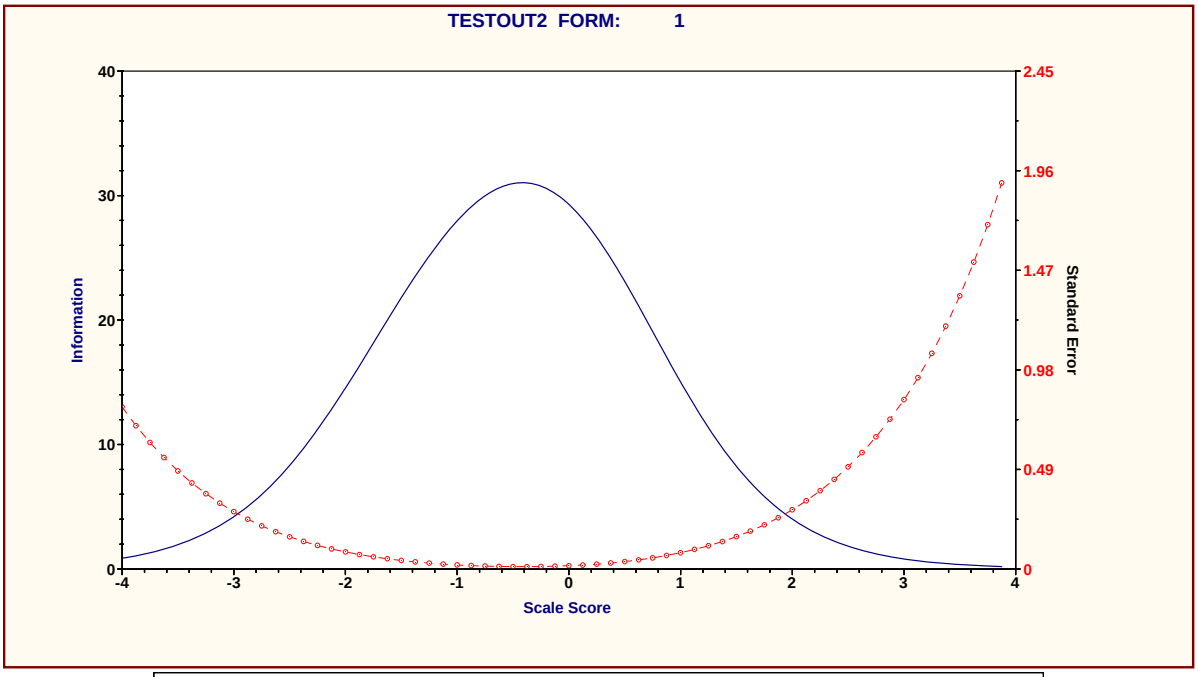
Table 3 shows the performance task ordering of the difficult simulation tasks in columns 1 and 2, the average simulation tasks in columns 3 and 4, and the easier simulation tasks in columns 5 and 6. The Rasch one-parameter model difficulty (*b*) measure was used for ordering of the tasks. Table 3 also illustrates a fairly continuous distribution of simulation task difficulty that could be used as the basis for development and measurement of a continuous learning pathway as discussed above. The simulation task difficulty shown in Table 3 can also be used in a CAT environment to converge on a pass-fail performance standard at a given IRT-defined cut score.

Table 3. Performance Task Ordering for Difficult, Average and Easy Tasks

Difficult Tasks	Measure	Average Tasks	Measure	Easy Tasks	Measure
CHALL02_21	2.30	CISCO05_02	0.24	CHALL02_01	-2.11
CHALL02_18	2.12	PASS2_02	0.15	IPADDR05_2	-2.36
CHALL02_13	1.91	BACK1_04	0.13	CISCO02_04	-2.28
CHALL02_12	1.91	SERTB02__1	0.16	CISCO02_08	-1.88
CHALL02_11	1.91	IPADDR05_3	-0.11	IPADDR5_04	-1.75
CHALL02_10	1.91	IPADDR05_5	-0.11	CISCO02_03	-1.31
CHALL02_09	1.91	PASS2_04	-0.35	CISCO02_6	-1.31
BAN2_03	1.22	BACK1_01	-0.34	CISCO02_7	-1.31
BAN2_02	1.22	CHALL02_14	-0.27	BACK1_02	-0.98
CISCO02_10	1.11	CHALL02_16	-0.27	RIPTRB01_1	-1.04
CHALL02_07	1.00	ADDR2_1	-0.41	CHALL02_02	-0.50
CHALL02_06	1.00	CISCO02_01	-0.51	PASS02_3	-0.75

Figure 3 shows the test information curves and standard error curves for the one-parameter analyses of the IT simulation tasks. Figure 4 shows the test information curve and standard error estimation for the two-parameter analyses. Comparisons of Figures 3 and 4 show a more peaked information curve for the two-parameter IRT model as compared with the one-parameter model. This same result is shown with several performance test examples presented later in this report.

Figure 3. IT Simulation Tasks: One-Parameter Test Information and Standard Error Curves

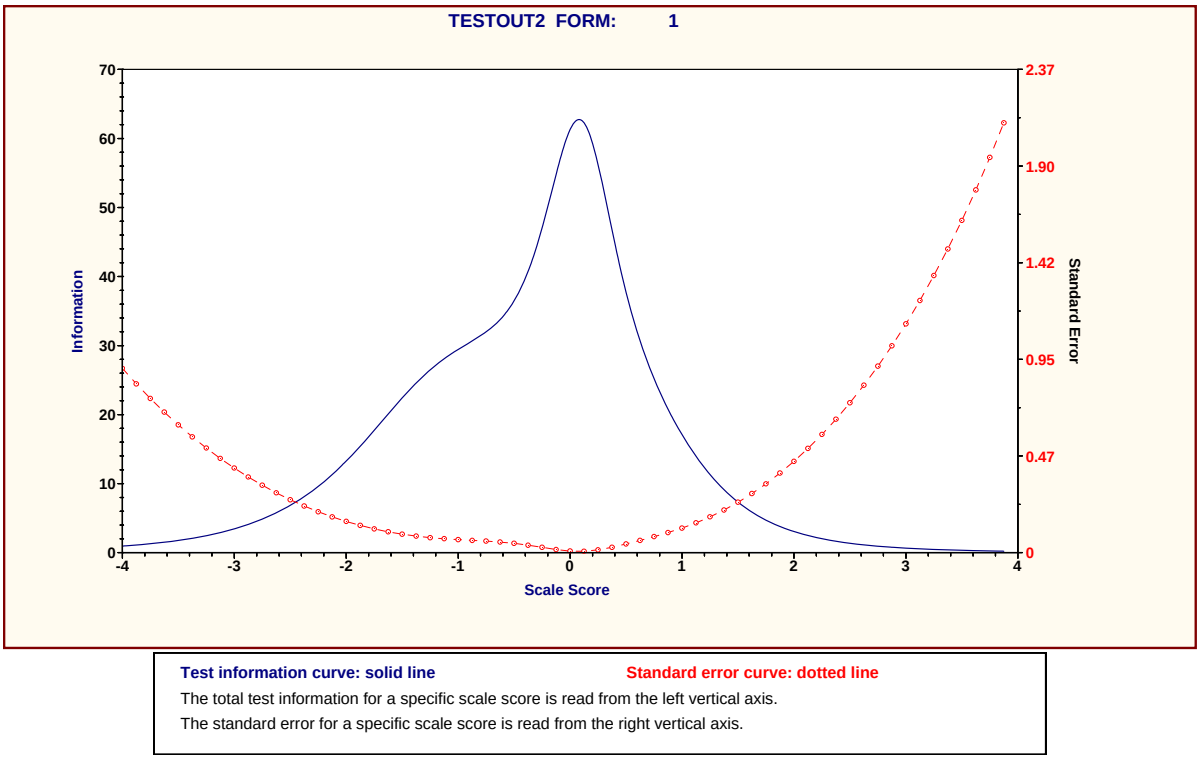


Test information curve: solid line **Standard error curve: dotted line**

The total test information for a specific scale score is read from the left vertical axis.

The standard error for a specific scale score is read from the right vertical axis.

Figure 4. IT Simulation Tasks: Two-Parameter Test Information and Standard Error Curves



Example 2: Office Word-Processing Performance Tasks

The second example includes 18 word-processing performance tasks administered to 1,517 examinees. Five items were scored polytomous with scores ranging from 1 to 3, 13 items were scored as dichotomous. The average weighted proportion correct (*p*) for the performance tasks was 0.70, the average point-biserial correlation of the performance tasks was 0.561, and the average high group – low group discrimination index was 0.205. The alpha reliability for the 18 performance tasks was 0.87. Table 4 presents summary statistics for the 18 Office Word Performance Tasks.

Table 4. Summary Statistics for Office Word Performance Tasks

Statistic	Value
N	1517
Min	1
Max	41
Mean	28.78
Median	30
Mode	35
Std Dev	7.55
Std Error Measurement	2.72
Skewness	-.91
Kurtosis	0.41

A variety of calibration models were computed using MULTILOG (Thissen, 1991, 2003) for the analysis, including the one-parameter and two-parameter logistic models, the nominal

model, and the graded response model. Table 5 presents summary IRT statistics for Samejima's graded response IRT calibration model for the 18 word-processing performance tasks.

Table 5. Samejima Graded Response IRT Calibration

Item	Number of Category Options	Common Item Discrimination Parameter (Std Error)	Item Difficulty Parameter Answer Option 1 or Greater (Std Error)	Item Difficulty Parameter Answer Option 2 or Greater (Std Error)	Item Difficulty Answer Parameter Option 3 or Greater (Std Error)
I01	4	1.09 (0.08)	-3.50 (0.25)	-2.31 (0.16)	0.56 (0.08)
I02	2	1.43 (0.09)	-1.46 (0.09)	1.20 (0.08)	
I03	4	1.32 (0.08)	-2.05 (0.12)	-.05 (0.06)	1.04 (0.08)
I04	2	1.49 (0.09)	-1.64 (0.10)	0.20 (0.06)	
I05	4	1.36 (0.09)	-2.67 (0.16)	-0.83 (0.07)	-0.54 (0.06)
I06	2	1.18 (0.09)	-1.44 (0.10)	-0.27 (0.07)	
I08	2	1.09 (0.08)	-2.57 (0.18)	-0.34 (0.07)	
I09	2	1.27 (0.10)	-2.86 (0.19)	-0.96 (0.08)	
I10	4	1.19 (0.08)	-2.19 (0.14)	-0.47 ((0.07)	1.29 (0.10)
I11	2	1.19 (0.10)	-2.84 (0.20)	-1.20 (0.10)	
I12	2	0.83 (0.08)	-2.20 (0.20)	1.47 (0.15)	
I13	2	0.79 (0.08)	-3.40 (0.31)	0.57 (0.11)	
I14	2	1.18 (0.08)	-1.55 (0.12)	-0.64 (0.08)	
I16	4	1.39 (0.10)	-2.68 (0.16)	-1.86 (0.11)	-0.61 (0.06)
I17	2	1.44 (0.09)	-2.34 (0.14)	-0.56 (0.06)	
I18	2	1.02 (0.08)	-2.39 (0.18)	0.12 (0.08)	
I19	2	1.43 (0.11)	-1.86 (0.11)	-1.03 (0.08)	
I20	2	1.52 (0.10)	-1.46 (0.09)	-0.15 (0.05)	
Ave		1.23			

Table 6 presents an item information table for the 18 word-processing performance tasks. This information table provides the amount of test information for each of the performance tasks at each of nine ability or proficiency levels evenly spaced from -2.0 to +2.0 in increments of .5. For example, performance task I04 provides item information of 0.59 at an ability level of -1.5, item information of 0.58 at an ability level of -1.0, item information of 0.58 at an ability level of -.50, and an information value of 0.59 at an ability value of 0.0.

Table 6. Information Table for Word Processing Performance Tasks: Item Information at Various Ability or Proficiency Levels

Item	-2.0	-1.5	-1.0	-.5	0.0	.50	1.0	1.5	2.0
I01	0.33	0.30	0.28	0.29	0.30	0.31	0.28	0.23	0.17
I02	0.44	0.52	0.49	0.42	0.40	0.46	0.52	0.49	0.38
I03	0.48	0.50	0.51	0.51	0.50	0.50	0.48	0.41	0.30
I04	0.53	0.59	0.58	0.58	0.59	0.54	0.40	0.25	0.13
I05	0.50	0.51	0.53	0.51	0.42	0.29	0.18	0.10	0.06
I06	0.32	0.38	0.41	0.40	0.36	0.29	0.21	0.14	0.08
I08	0.33	0.32	0.33	0.32	0.30	0.25	0.18	0.12	0.08
I09	0.44	0.44	0.43	0.38	0.29	0.19	0.11	0.06	0.04
I10	0.40	0.41	0.41	0.42	0.41	0.41	0.40	0.37	0.30
I11	0.41	0.40	0.37	0.30	0.22	0.15	0.09	0.05	0.03
I12	0.18	0.18	0.17	0.17	0.17	0.17	0.18	0.18	0.17
I13	0.15	0.15	0.15	0.16	0.16	0.16	0.15	0.14	0.12
I14	0.34	0.40	0.41	0.38	0.31	0.23	0.15	0.10	0.06
I16	0.59	0.59	0.57	0.52	0.41	0.28	0.17	0.09	0.05
I17	0.56	0.55	0.56	0.54	0.45	0.31	0.18	0.10	0.05
I18	0.28	0.28	0.28	0.28	0.28	0.26	0.22	0.17	0.12
I19	0.55	0.60	0.57	0.46	0.31	0.19	0.10	0.05	0.03
I20	0.50	0.63	0.67	0.66	0.61	0.47	0.29	0.16	0.08
Total Test Info	8.3	8.7	8.7	8.3	7.3	6.4	5.3	4.2	3.2
Ave. Standard Error	0.35	0.34	0.34	0.35	0.36	0.39	0.43	0.49	0.56

Figure 5 presents a plot of the WINSTEPS (Linacre & Wright, 1991-2000)one-parameter measure and the MULTILOG one-parameter difficulty values. Figure 6 presents a plot of the WINSTEPS measure and the BILOG-MG one-parameter difficulty value. Figures 5 and 6 show that the performance simulation tasks are generally measured consistently, since the majority of the estimation values lie on the diagonal and very few simulation task elements are off diagonal.

**Figure 5. Plot of WINSTEPS Measure (Horizontal)
and MULTILOG One-Parameter Difficulty (Vertical)**

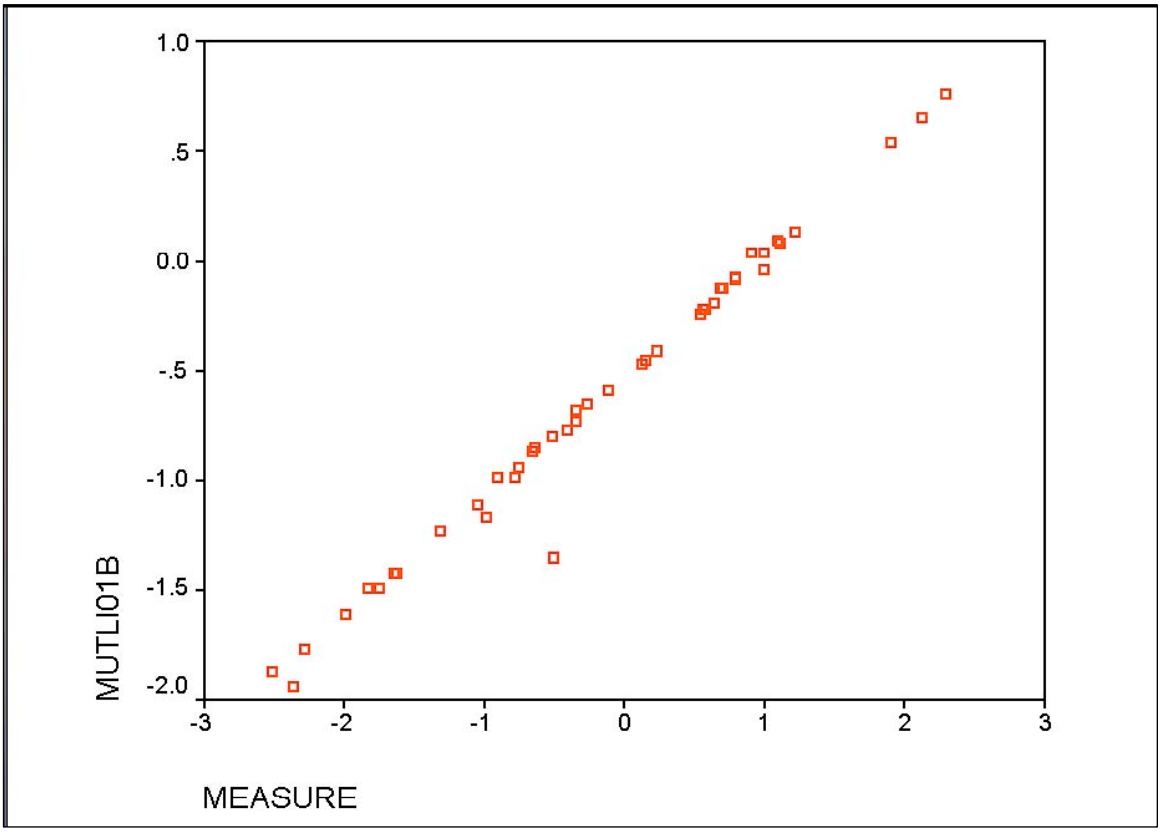
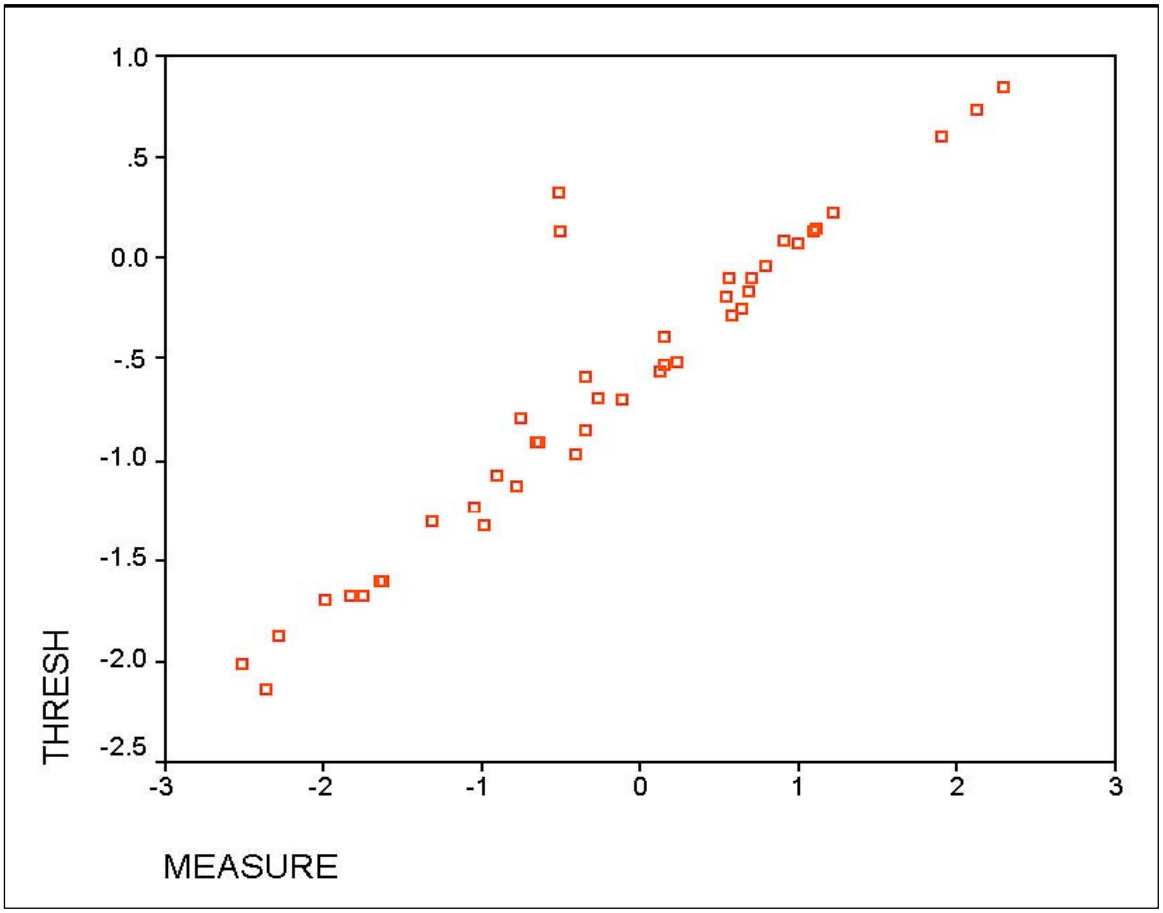


Figure 6. Plot of WINSTEPS Measure (Horizontal) and BILOG-MG One Parameter Difficulty (Vertical)



Dimensionality analysis. Dimensionality analysis was determined using both principal components analysis and nonlinear factor analysis. With the principal components analysis, the first principal component was six times larger than all remaining subsequent components. The first principal component accounted for 32% of the cumulative variance from the eighteen principal components that were extracted, one for each performance task. Figure 7 presents the scree test of the ordered principal components for the data. The scree test shows the presence of one principal or salient factor.

Figure 7. Plot of Eigenvalues from the Components Analysis

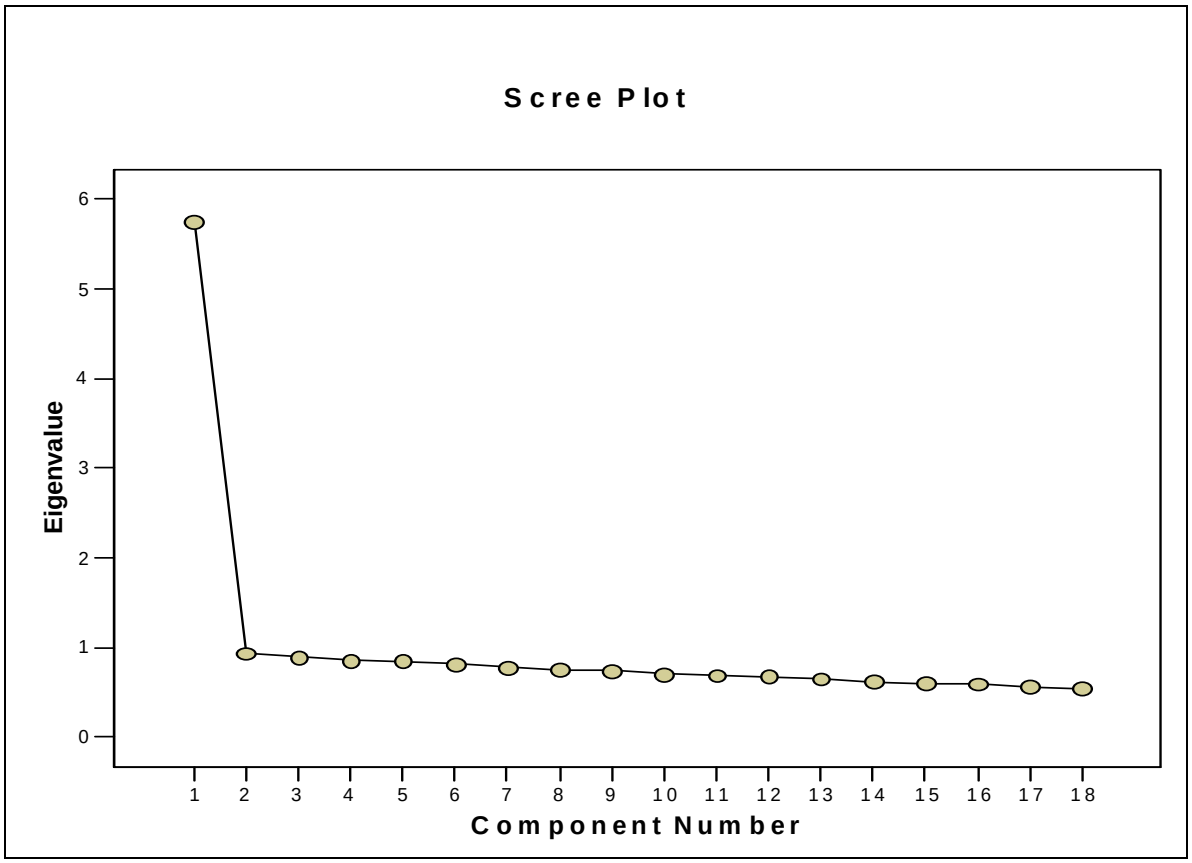


Table 8 provides additional information regarding the dimensionality of the word-processing performance tasks by comparing the factor loadings for the first principal component from SPSS using principal components analysis, principal axis factoring, the full-information factor analysis loadings from TESTFACT (Wood, Wilson, Gibbons, Shilling Muraki, & Bock, 2003) , and the confirmatory factor analysis loadings obtained using the nonlinear factor analysis program NOHARM (Fraser, 1988, Fraser & McDonald, 1988).

Table 8. Dimensionality Loadings for First Factors From Principal Components, Principal Factors, TESTFACT, and NOHARM

Task ID	First Principal Component SPSS	First Principal Axis Factor SPSS	Factor Loadings TESTFACT	Confirmatory Factor Loadings NOHARM
I01	.549	.511	0.553	0.722
I02	.597	.562	0.486	1.000
I03	.604	.569	0.545	0.748
I04	.642	.611	0.607	0.876
I05	.591	.555	0.663	0.751
I06	.550	.512	0.608	0.767
I08	.530	.492	0.539	0.683
I09	.564	.526	0.519	0.693
I10	.570	.533	0.566	0.727
I11	.539	.500	0.572	0.683
I12	.439	.401	0.406	1.000
I13	.414	.376	0.363	0.595
I14	.561	.513	0.574	0.701
I16	.613	.579	0.696	0.799
I17	.629	.596	0.663	0.800
I18	.516	.477	0.513	0.702
I19	.574	.537	0.684	0.824
I20	.648	.618	0.678	0.852
Ave.	.563	.526	0.569	0.774

Table 9 provides the confirmatory factor loadings from NOHARM and the NOHARM computed discrimination (a) and item difficulty parameters. Note the number of discrimination parameters that are over 1.0 for the eighteen performance tasks.

Table 9. Confirmatory Factor Loadings, Discrimination, and Difficulty Parameters

Task ID	Confirmatory Factor Loadings NOHARM	NOHARM IRT Discrimination a Parameter	NOHARM IRT Difficulty b Parameter
I01	0.722	1.044	-1.506
I02	1.000	2.356	1.916
I03	0.748	1.129	-0.124
I04	0.876	1.813	0.557
I05	0.751	1.138	-0.416
I06	0.767	1.197	0.165
I08	0.683	0.936	0.116
I09	0.693	0.962	-0.555
I10	0.727	1.057	-0.031
I11	0.683	0.935	-0.741
I12	1.000	2.020	1.454
I13	0.595	0.740	1.107
I14	0.701	0.984	-0.191
I16	0.799	1.351	-1.262
I17	0.800	1.333	-0.168
I18	0.702	0.985	0.575
I19	0.824	1.455	-0.543
I20	0.852	1.630	0.208
Ave.	0.774	1.282	.0312

Example 3: Network Performance Practicum Calibration and Analyses

The third example includes calibrations from four scenarios—one, three, six and nine—from a network performance practicum examination. The performance tasks were administered with virtual machine technology. The examinee must examine a remote network system and perform the tasks that are designated with each scenario. The performance tasks are scored by running a performance script that compares the examinee results on the performance exam to the subject matter expert decision paths. Scores for each task are dichotomously scored and a weighted average is computed across tasks using scoring weights specified by subject matter experts. The results in Table 10 show summary information for each of the scenarios with number of candidates, number of tasks, mean score, standard deviation, standard error of the mean, alpha reliability, skewness, kurtosis, and standard error of measurement. Tables 11-14 provide classical and IRT analysis statistics for each of the four performance scenarios.

Table 10. Performance Practicum Classical Item Analysis

Test Statistics	Scenario One	Scenario Three	Scenario Six	Scenario Nine
Number of candidates	347	97	222	182
Number of items	68	34	70	54
Mean	49.68	20.124	58.685	45.731
Standard deviation	12.169	8.592	9.402	11.585
SE of mean	0.653	0.872	0.631	0.859
95% confidence interval for mean	1.280	1.709	1.237	1.684
Alpha Reliability	0.947	0.934	0.921	0.969
Skewness	-1.070	-0.687	-1.051	-2.047
Kurtosis	1.566	-0.609	0.806	4.393
SE of Measurement	2.801	2.207	2.643	2.040
95% confidence interval	5.490	4.326	5.180	3.998

Table 11. Descriptive Statistics: Classical and IRT, Scenario One

Statistic	Items	Min	Max	Mean	Std. Deviation
P-value	68	.262	1.000	.731	.227
Point Biserial	68	-.023	.655	.462	.158
High-Low Disc	68	-.006	.618	.273	.186
Weighted Correlation	68	-.018	.631	.448	.154
Rasch *b*	68	-7.300	3.510	-.106	2.266
BILOG: 1-Par *b*	67*	-3.574	.909	-1.193	1.254
BILOG: 2-Par *a*	67*	.269	2.396	1.081	.467
BILOG: 2-Par *b*	67*	-6.807	1.362	-1.388	1.756

*One item was answered correctly by 100 percent of the candidates.

Table 12. Descriptive Statistics: Classical and IRT, Scenario Three

Statistic	Items	Min	Max	Mean	Std. Deviation
P-value	34	.113	.948	.592	.200
Point Biserial	34	-.328	.769	.549	.221
High-Low Disc	34	-.342	.712	.417	.229
Weight Correlation	34	-.441	.669	.482	.202
Rasch *b*	34	-4.080	3.640	-.001	1.751
BILOG: 1-Par *b*	34	-2.464	1.639	-.361	.929
BILOG: 2-Par *a*	33*	.623	5.328	1.746	1.582
BILOG: 2-Par *b*	33*	-2.410	2.101	-.260	.971

* One item did not reach convergence criteria.

Table 13. Descriptive Statistics: Classical and IRT, Scenario Six

Statistic	Items	Min	Max	Mean	Std. Deviation
P-value	70	.518	1.000	.838	.141
Point Biserial	67	-.079	.610	.401	.126
High-Low Disc	70	-.018	.587	.210	.137
Weighted Correlation	67	-.064	.670	.282	.235
Rasch *b*	70	-4.700	3.000	-.325	1.887
BILOG: 1-Par *b*	70	-4.131	-0.095	-1.842	.961
BILOG: 2-Par *a*	70	.001	4.000	1.488	1.451
BILOG: 2-Par *b*	70	-4.042	0.000	-1.694	1.038

Table 14. Descriptive Statistics: Classical and IRT, Scenario Nine

Statistic	Items	Min	Max	Mean	Std. Deviation
P-value	54	.593	.973	.847	.087
Point Biserial	54	-.007	.745	.622	.108
Hi-Lo Disc	54	.035	.564	.284	.126
Weighted Correlation	54	.041	.705	.606	.094
Rasch *b*	54	-3.400	2.700	.000	1.207
BILOG: 1-Par *b*	54	-2.358	-.302	-1.244	.420
BILOG: 2-Par *a*	54	.501	4.326	1.836	1.016
BILOG: 2-Par *b*	54	-4.617	-.295	-1.32	.628

Test information and standard error curves for the one- and two-parameter IRT models were computed for each performance scenario. Comparable test information and standard error curves are provided by both XCALIBRE (Assessment Systems, 1996) and BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996, 2003). The one-parameter curves are presented in Figure 8 and the two-parameter curves are presented in Figure 9 for Scenario One. In a similar manner the one-parameter curves and two-parameter summaries for Scenario Three are presented in Figures 10 and 11. For Scenario Six these results are presented in Figures 12 and 13. For Scenario Nine the results are presented in Figures 14 and 15. The consistent result from these four performance testing IRT analyses shows that the two-parameter analysis provides higher and more peaked test information and lower standard error curves than the one-parameter analysis.

Figure 8. One-Parameter Information and Standard Error Curves for Scenario One

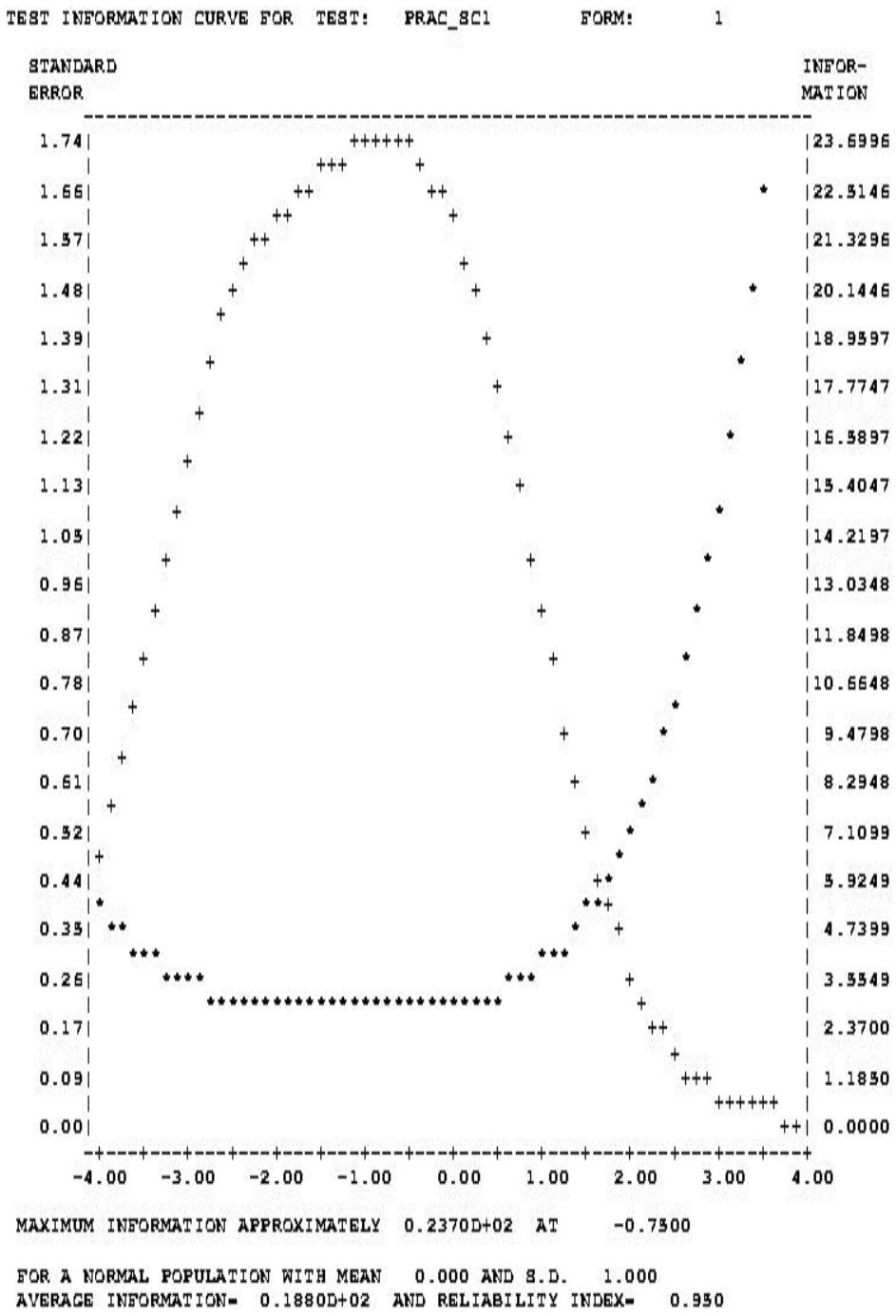


Figure 9. Two-Parameter Information and Standard Error Curves for Scenario One

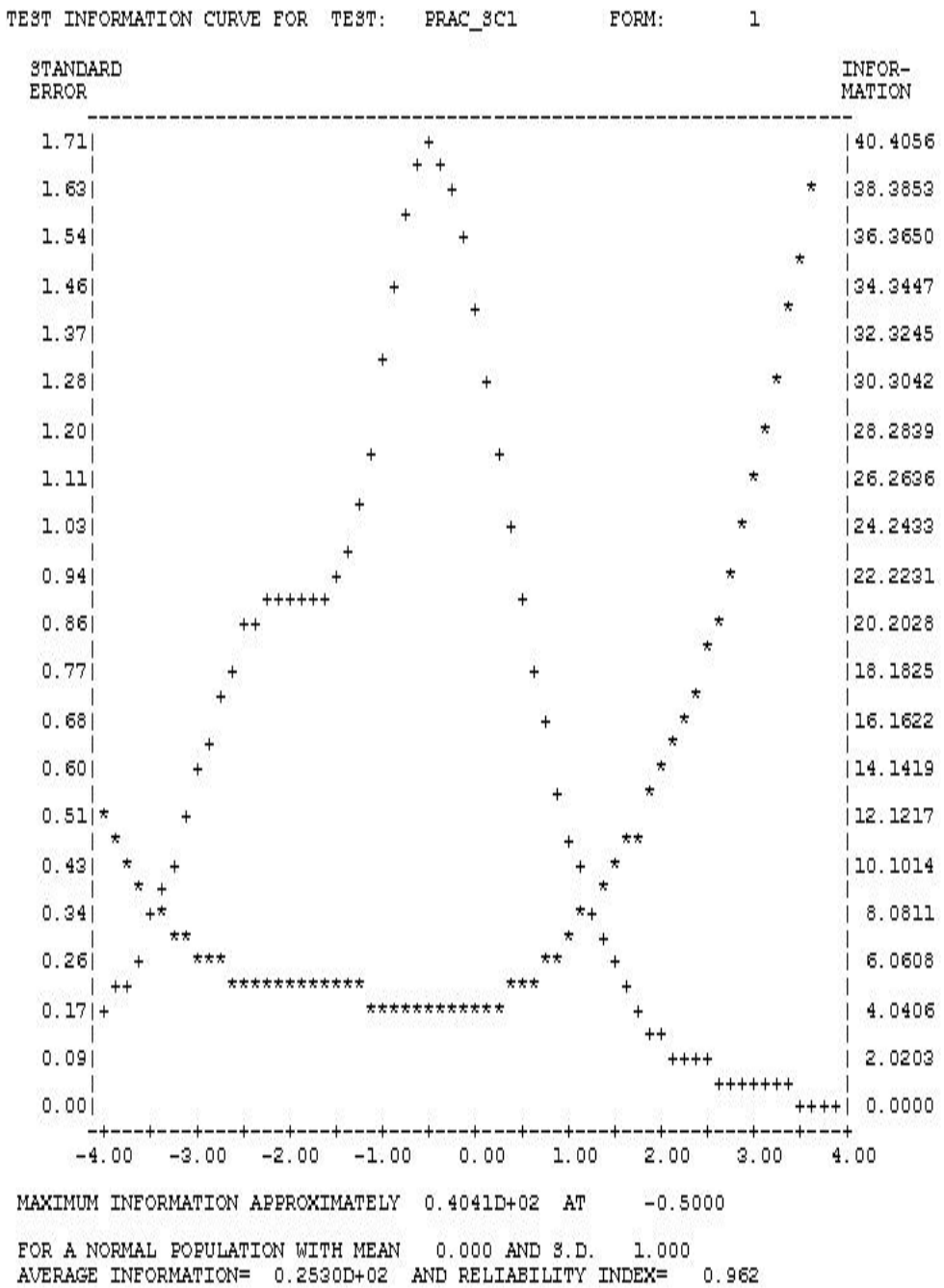


Figure 10. One-Parameter Information and Standard Error Curves for Scenario Three

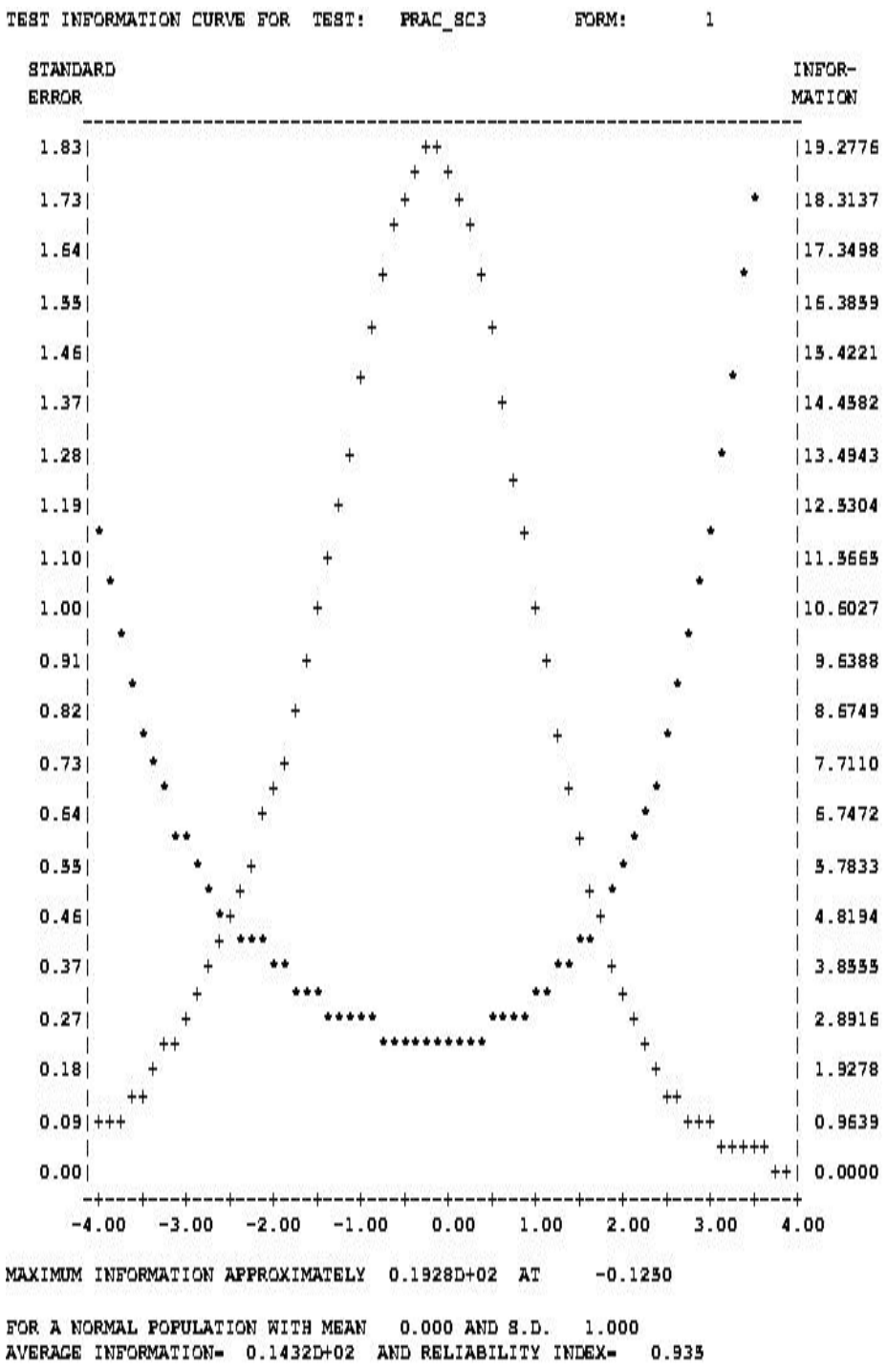


Figure 11. Two-Parameter Information and Standard Error Curves for Scenario Three

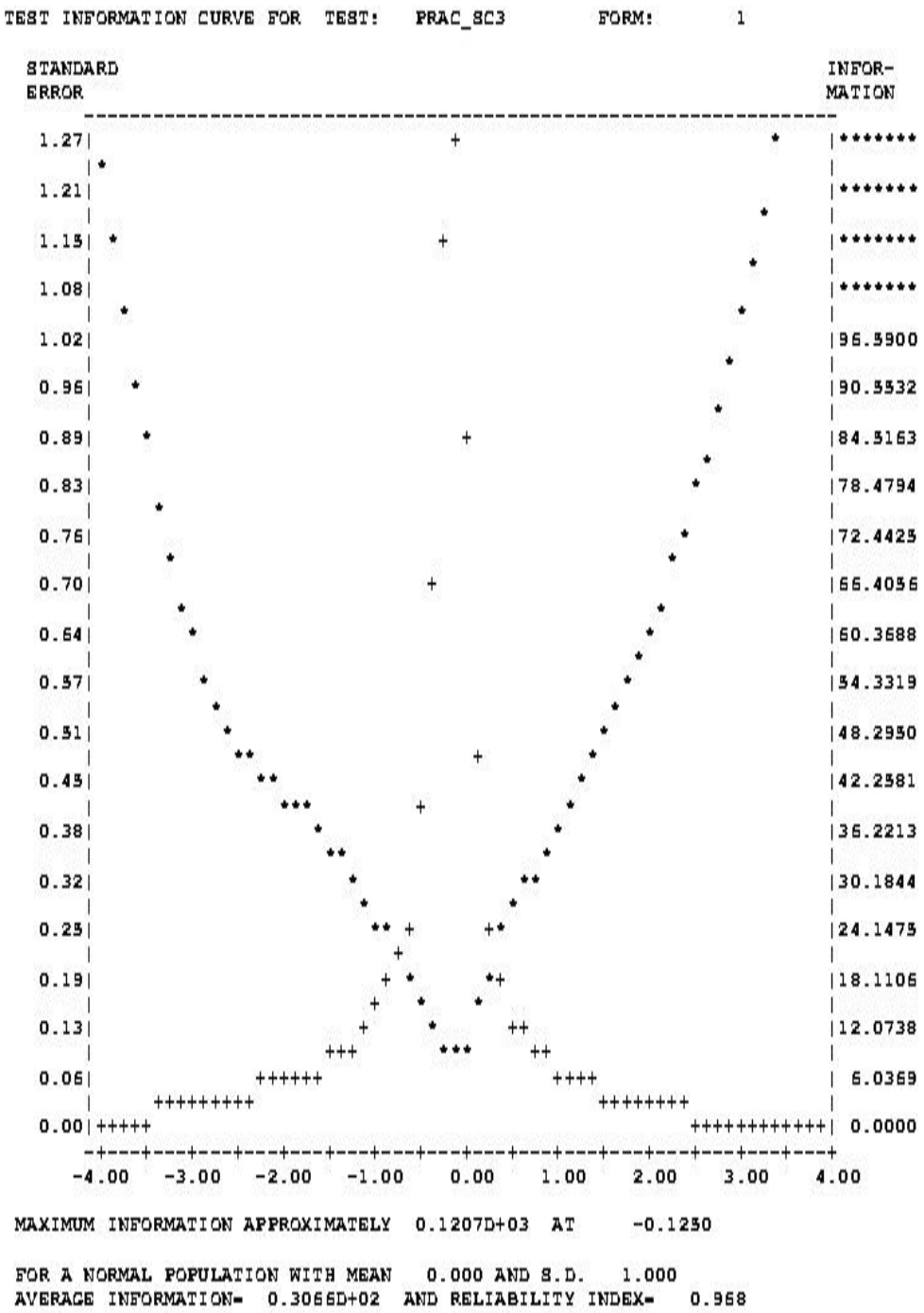


Figure 12. One-Parameter Information and Standard Error Curves for Scenario Six

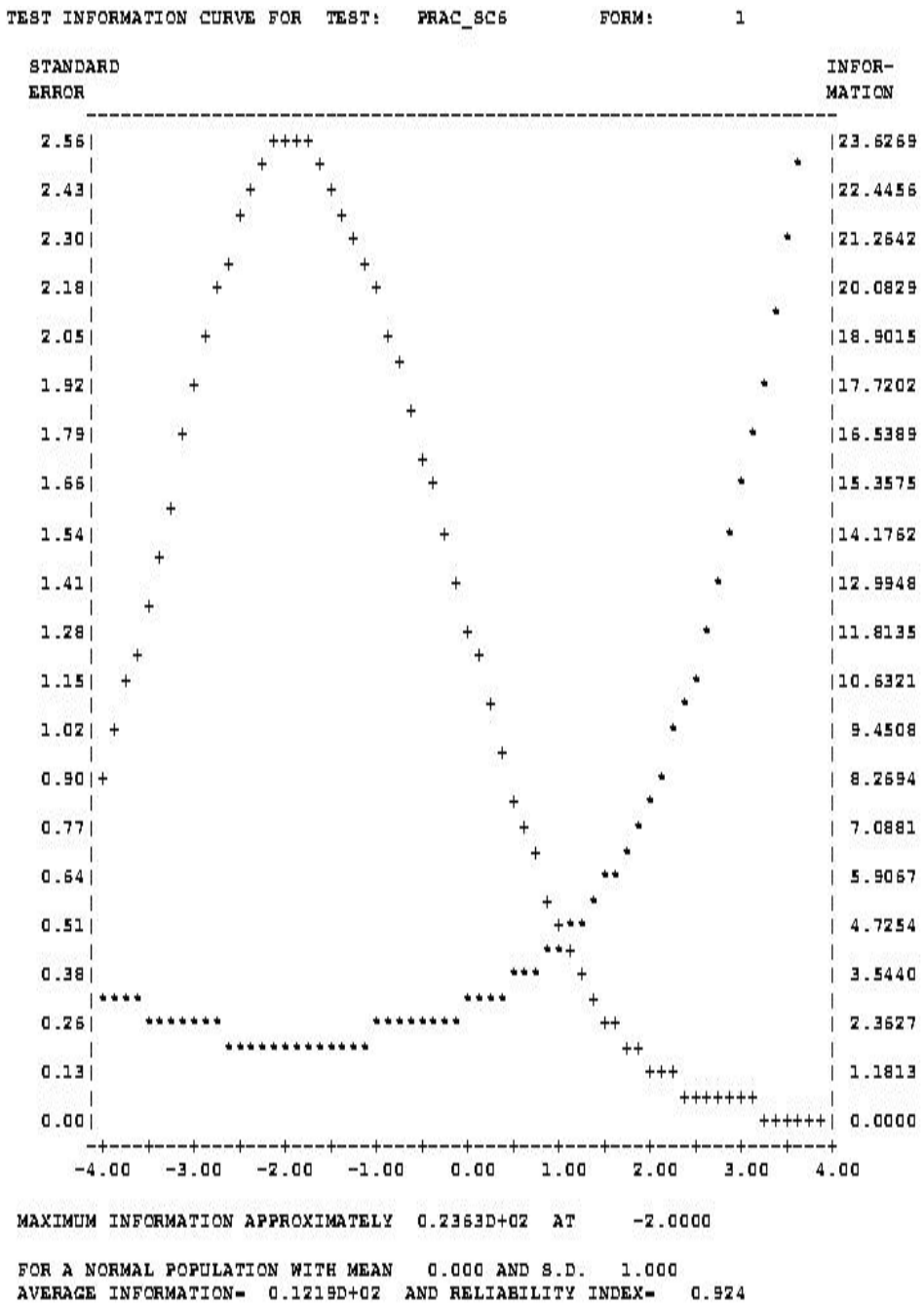


Figure 13. Two-Parameter Information and Standard Error Curves for Scenario Six

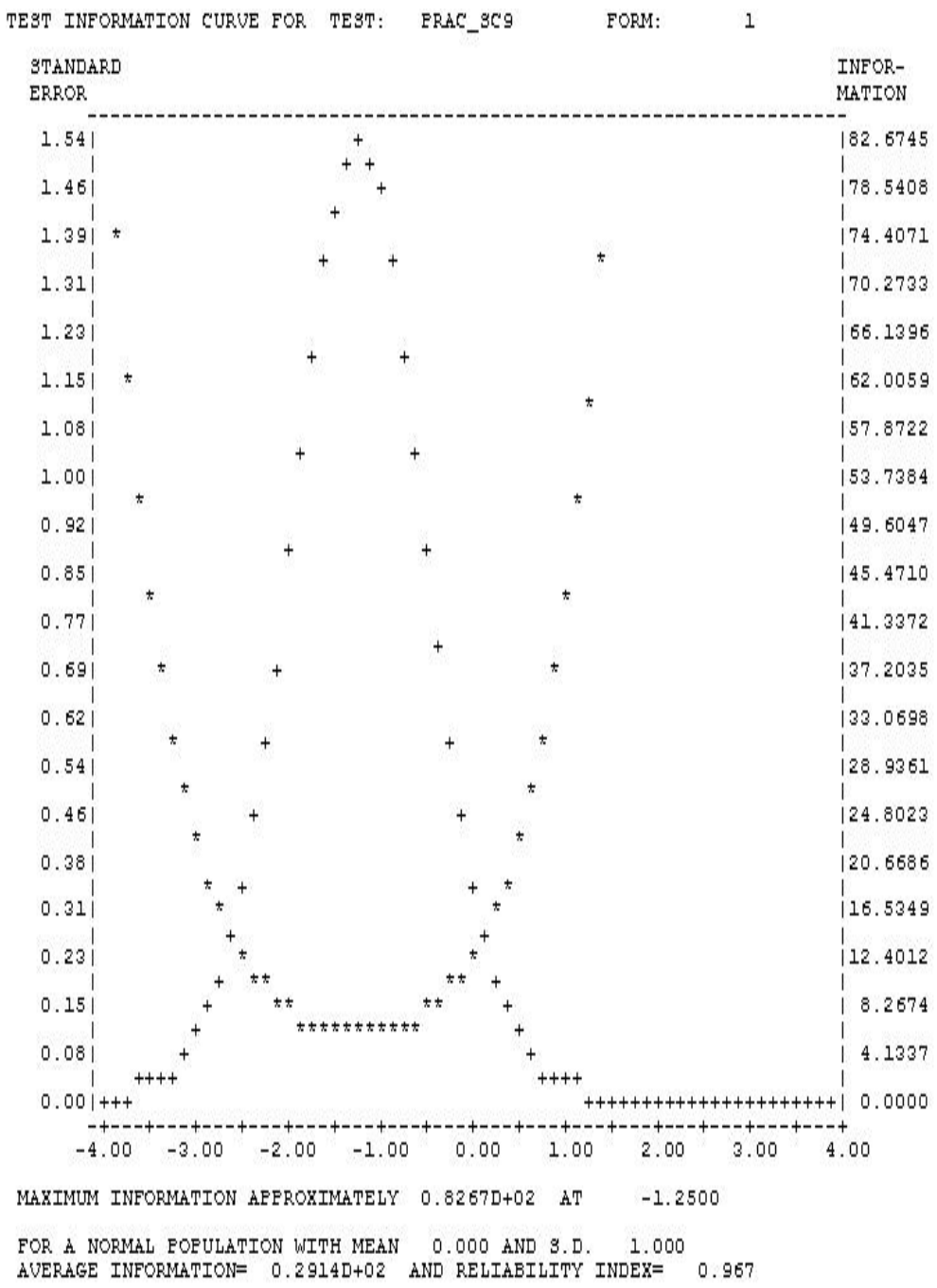
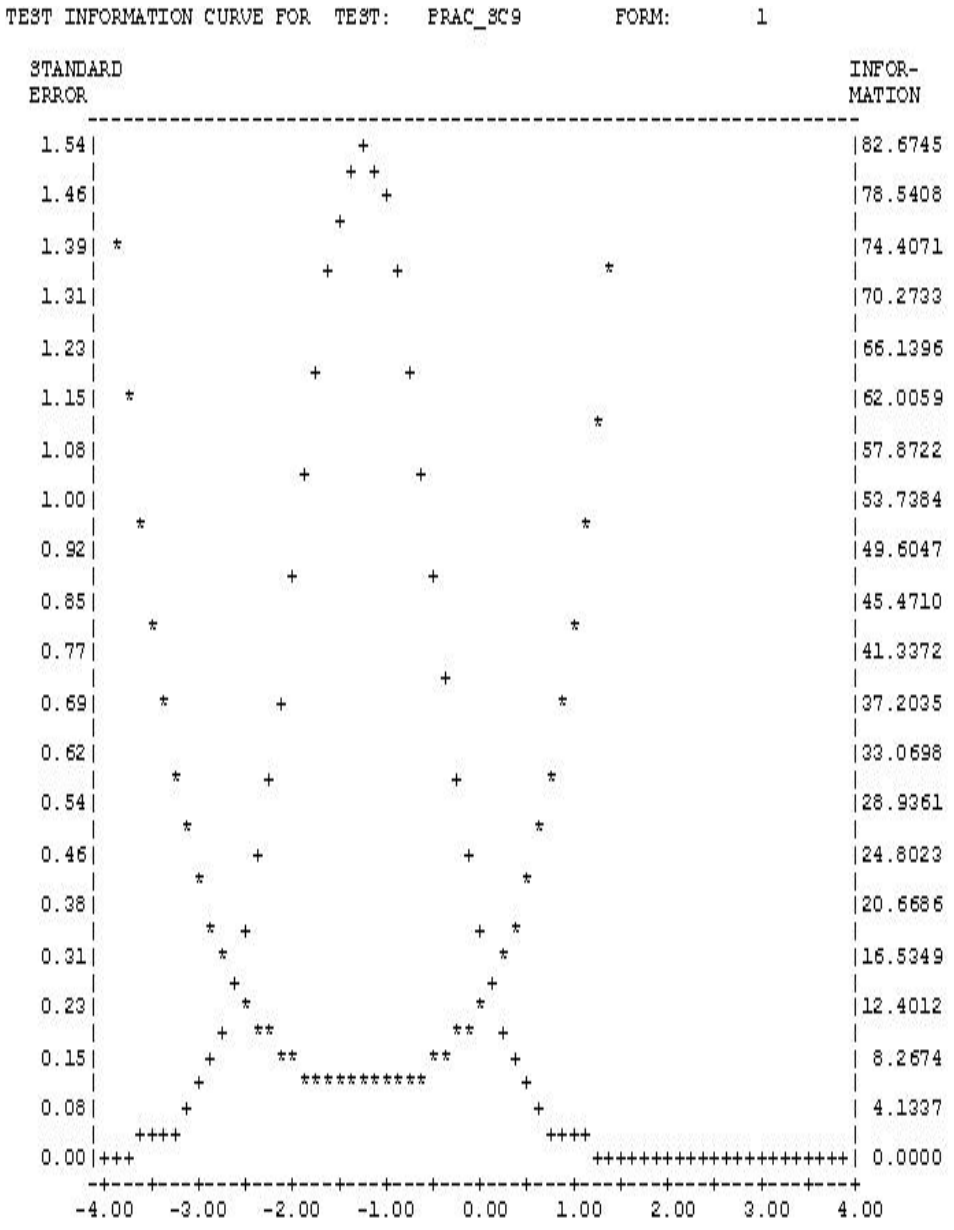


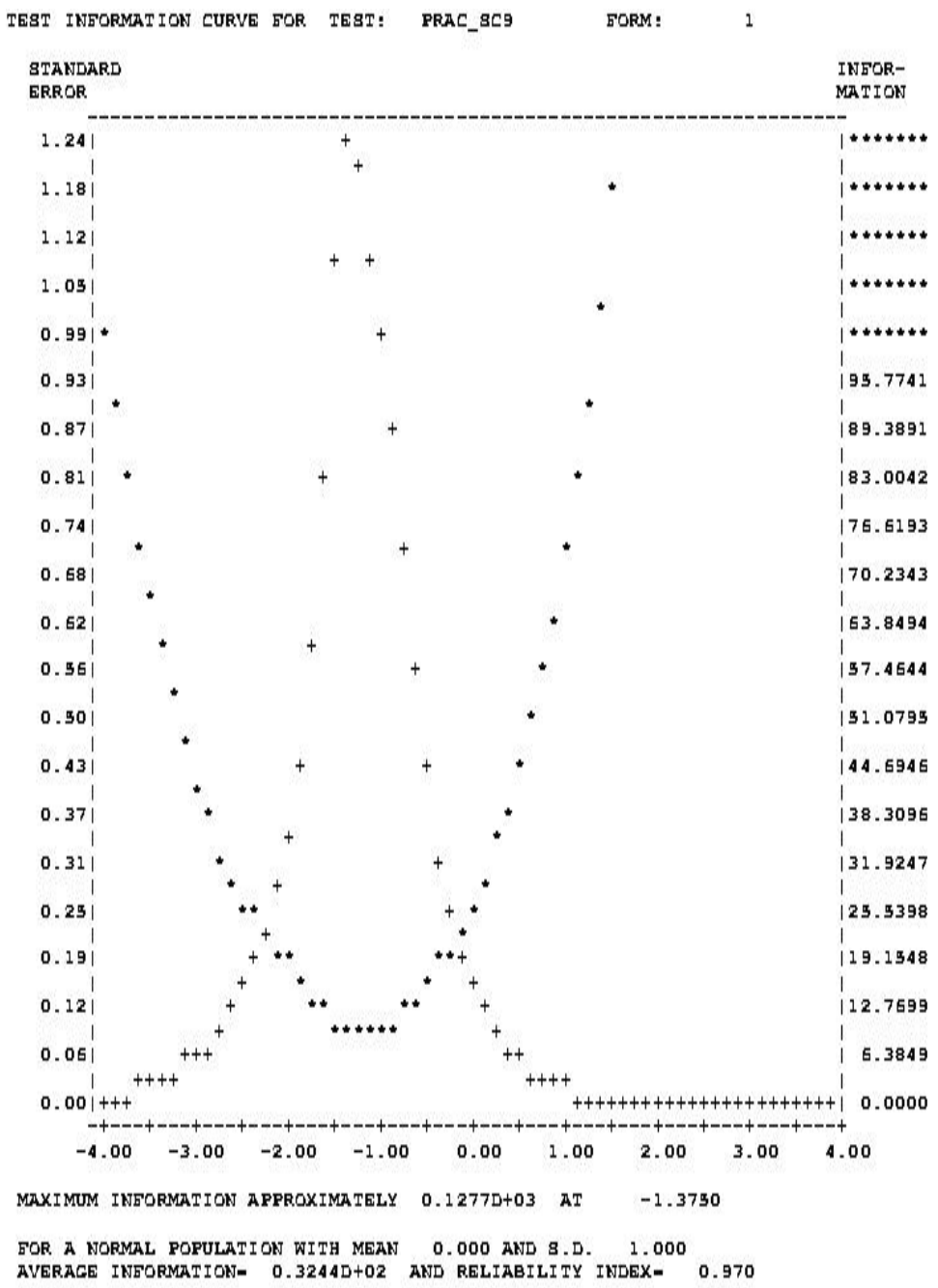
Figure 14. One-Parameter Information and Standard Error Curves for Scenario Nine



MAXIMUM INFORMATION APPROXIMATELY 0.8267D+02 AT -1.2500

FOR A NORMAL POPULATION WITH MEAN 0.000 AND S.D. 1.000
AVERAGE INFORMATION= 0.2914D+02 AND RELIABILITY INDEX= 0.967

Figure 15. Two-Parameter Information and Standard Error Curves for Scenario Nine



Conclusions and Recommendations

This paper highlighted a progressive series of psychometric innovations that have formed the foundation for our work on:

1. Defining and applying psychometric theory to the analysis of integrated, performance work models,
2. The use of logical measurement opportunities to define scores on activities that are not items in the traditional sense,
3. The sequencing of the performance tasks as benchmarks across a learning and performance domain, and
4. The development of continuous learning progress pathways as alternative routes through a learning and performance domain.

The paper illustrated several initial steps in using psychometric models for the analysis and measurement of performance tasks and simulation assessments. IRT calibrations of performance task difficulty, task discrimination, task step parameters, task information, and task standard errors can be used within CAT environments to select the next simulation or performance task given a current ability or proficiency estimate based on performance on previous simulation and performance tasks. The use of logical measurement opportunities within performance and simulation tasks provide a generalization of the testlet concept to refer to sequences of logical measurement opportunities to define the relevant set of test items. We recommend continued research investigation regarding validity-centered design, evidence-centered design and assessment engineering to design item score validity within each stage of CATs (Luecht, 2007; Williamson, Mislevy, and Bejar, 2006).

References

Almond, R. G., Steinberg, L. S., & Mislevy, R. J. (2003). Enhancing the design and delivery of assessments: A four process architecture. *Journal of technology, learning and assessment*, 1(5). Accessed from weblink: http://www.bc.edu/research/intasc/jtla/journal/v1n5.shtml

Assessment Systems Corporation (1996). XCALIBRE marginal maximum likelihood estimation program, Version 1.10 [Computer program]. St. Paul, MN: Author

Bock. D. R. & Aitkin. M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika*, 46, 443-459.

Bock, D. R. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied psychological measurement*, *6*, 431-444.

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. In R. L. Brennan (Ed.) *Educational measurement, 4th Edition*, Westbury, CT: American Council on Education and Preager Publishers., pp. 1-16,

Brown, J. M., & Weiss, D. J. (1977). An adaptive testing strategy for achievement test batteries (Research Rep. No. 77-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Bunderson, C. V. (2003). *On the validity-centered design and continuing validation of learning progress measurement systems*. Unpublished manuscript.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.) *Educational measurement, 3rd Edition*. New York: American Council on Education and Macmillan Publishers, pp 367-407.

Bunderson, C. V. Gibbons, A.S., Olsen, J.B., & Kearsley, G.P. (1981). Work models: Beyond instructional objectives. *Instructional Science*, 10, 205-215.

Chang, H. H., Qian J., & Ying, Z. (2001) *a*-stratified multistage computerized adaptive testing item b-blocking. *Applied Psychological Measurement*, 25, 333-342.

Chang, H. H. & Ying, Z. (1999). *a*-stratified multi-stage computerized adaptive testing. *Applied Psychological Measurement*, 23, 211-222.

Chang, H. H., & Ying, Z. (1997). *Multistage CAT with stratification designs*. Paper presented at the annual meeting of the Psychometric Society, Gatlinburg, TN.

Cohen, A. S. & Wollack, J. A. (2006). Test administration, security, and reporting. In R. L. Brennan (Ed.) *Educational measurement, 4th Edition*, Westbury, CT: American Council on Education and Preager Publishers., pp. 355-386.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In R. L. Brennan (Ed.) *Educational measurement, 4th Edition*, Westbury, CT: American Council on Education and Preager Publishers., pp. 471-515

Fraser, C. (1988). NOHARM: An IBM PC computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory [Computer software]. Armidale, Australia: The University of New England.

Fraser, C. & McDonald, R. P. (1988). NOHARM: Least squares item factor analyses. *Multivariate Behavioral Research*. *23*,267-269.

Gibbons, A. S., Bunderson, C. V., Olsen, J. B., & Robertson, J. (1995). Work models: Still beyond instructional objectives. *Machine-Mediated Learning*, *5*(3 & 4), 221-236.

Gibbons, A. S., & Fairweather, P. G. (1999). Instructional Strategy III: Fragmentation and Integration, Chapter 15 in Gibbons, A. S., & Fairweather, P. G., *Computer-based instruction: Design and Development*. Englewood Cliffs, NJ: Educational Technology Publications, pp. 278-296.

Green, B. F. (1970). Comments on tailored testing. In W. Holtzman, W H. (Ed.). *Computer-assisted instruction, testing and guidance*. New York: Harper & Row, Publishers, pp. 184-197.

Hambleton, R. K. & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff Publishing.

Harris, L. J. (1978), Sex differences in spatial ability: Possible environmental, genetic, and neurological factors. In M. Kinsbourne (Ed.) *Asymmetrical function of the brain*. New York, NY: Cambridge University Press. pp. 405-522,

Halpern, D. F. (2000). *Sex differences in cognitive ability*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Kimura, D. (1999). *Sex and cognition*. Cambridge, MA: MIT Press.

Kingsbury, G. G. & Weiss, D. J. (1979). *An adaptive testing strategy for mastery decisions* (Research Report 79-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Linacre, J. M. & Wright, B. D. (1991-2000). A user's guide to WINSTEPS. [Computer program]. Chicago, IL: MESA Press.

Luecht, R. M. (2007). *Assessment engineering*. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL: April.

McGlone, J. (1978). Sex differences in functional brain asymmetry. *Cortex* , *14,* 122-128.

McGlone, J. (1980). Sex differences in human brain asymmetry: A critical survey. *Behavioral Brain Science*. *3,* 215-263.

Messick, S. (1998). *Consequences of test interpretation and use: The fusion of validity and values in psychological assessment*. Princeton, NJ: Educational Testing Service, RR-989-4.

Messick, S. (1998). Test validity: a matter of consequence. *Social indicators research*, *45*, 35-44.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *32(2)*, 13-23.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18(2)*, 5-11.

Messick, S. (1989b) Validity. In R. L. Linn (Ed.), *Educational measurement*, *3rd Edition*, New York: American Council on Education/Macmillan Publishing, pp.11-103.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer and H. Braun (Eds.), (1988) *Test validity*, Lawrence Erlbaum Associates Publisher, pp. 33-45.

Mislevy, R. J., Steinberg, L. S., & Almond, R. A. (2003). On the structure of educational assessments: *Measurement: Interdisciplinary research and perspectives*, *1*, 3-67.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *Evidence-centered assessment design*. Princeton, NJ: Educational Testing Service.

Olsen, J. B. (2006). Performance testing: Validity issues and design considerations for online testing. In D. Williams, S. L. Howell, and M. Hricko (Eds.) *Online assessment, measurement and evaluation*. Hershey, PA: Information Science Publishing, pp. 259-274.

Sands, W. A., Waters, B. K., & McBride, James R. (1997, 2001). *Computerized adaptive testing: From inquiry to operation*. Washington, D.C.: American Psychological Association.

Thissen, D. (1991). MULTILOG: Multiple category item analysis and test scoring using item response theory . [Computer program]. Chicago, IL: Scientific Software International, Inc.

Thissen, D. (2003). MULTILOG. In M. Du Toit (Ed.) *IRT from SSI,* pp.345-409. Lincolnwood, IL: Scientific Software International, Inc.

van der Linden, W. J. (2005). *Linear models for optimal test design*. New York: NY: Springer Science+Business Media, Inc.

van der Linden, W. J., & Glas, C. A. W. (2000*). Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer Academic Publishers.

van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of modern item response theory*. New York: NY: Springer.

van der Linden, W. J. (2000, 2003). Constrained adaptive testing with shadow tests. In W. J. van der Linden and C. A. W. Glas (Eds.) *Computerized adaptive testing: Theory and practice*. Boston, MA: Kluwer Academic Publishers. pp. 27-52.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory.* New York, NY: Cambridge University Press.

Wainer, H., Dorans, N. J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of educational measurement*. *24*, 185-201.

Warm, Thomas A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

Williamson, D. M., Mislevy, R. J., & Bejar, I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wood, R., Wilson, D. T., Gibbons, R., Schilling, S., Muraki, E., & Bock, R. D. (2003). TESTFACT: Classical item and item factor analysis. [Computer software]. Chicago, IL: Scientific Software International, Inc.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items. [Computer program]. Chicago, IL: Scientific Software International, Inc.

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (2003). BILOG-MG. In M. Du Toit (Ed.) *IRT from SSI* pp.24-256. Lincolnwood, IL: Scientific Software International, Inc.