Item Banking, Test Development, and Test Delivery

David J. Weiss

University of Minnesota

N660 Elliott Hall
Department of Psychology
University of Minnesota
Minneapolis MN 55455-0344

Email Address: djweiss@umn.edu

In Press, 2011
The APA Handbook on Testing and Assessment
Kurt F. Geisinger, Editor-in-Chief
Washington DC: American Psychological Association

Item Banking, Test Development, and Test Delivery

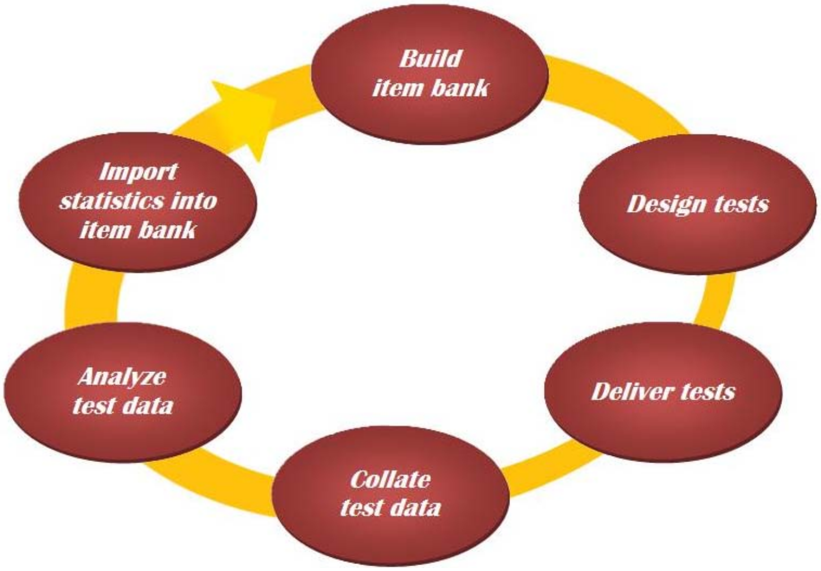
The paper-and-pencil test that dominated psychological, educational, personnel and other applications of testing for the majority of the twentieth century was born in the second decade of the 1900s in response to the personnel needs of World War I (Dubois, 1970). With the need to screen and classify large numbers of recruits rapidly and efficiently, the then predominant mode of testing by individual psychologists was not able to meet the demands of the U.S. military. The multiple-choice test question was invented, tests were written and printed and given to groups of recruits –the first major implementation of group, rather than individual, testing.

Because of its efficiency, paper-and-pencil (P&P) testing spread rapidly into other fields that previously had relied on individually administered tests – education, intelligence testing, and other personnel testing applications. P&P testing also began to be used for measuring attitudes, interests, and other personality variables, thus permitting the recently born field of psychology to generate data on a wide variety of variables quickly and efficiently.

Although data acquisition using P&P tests was efficient, the process of test development – especially for larger testing programs – was anything but efficient. Figure 1 provides an overview of the major components of the test development process. For at least the first 50 or 60 years of P&P testing, maintaining a collection of items for any continuing testing program, including classroom testing, was a tedious process fraught with numerous opportunities for error. Test questions (items) were frequently written, or perhaps typed, on index cards. The cards were kept in file drawers, sometimes separated into content classifications. When item statistics were available on items, they were frequently written on the backs of the cards, identified by test form and date. To create a test, the test developer would manually search the file drawer, review the content and statistics for an item, and put it aside if selected for use in the test. When a sufficient number of cards had been selected, they might be reviewed by others, and some replaced from the file drawer while returning the rejected items. There were obviously many opportunities for item cards to get lost, misplaced, or misfiled.

Once an appropriate set of items had been selected, the cards would be manually ordered in the desired order, and then typed onto a duplicating master. If an alternate form was needed, the order of the cards would be modified and a new test again typed from the cards. Of course, the typed test forms had to be completely proofread each time to ensure that the test items had not been inadvertently changed from the text on the index cards. The next steps, which are still necessary today for P&P tests, were duplication, collation, shipping (if required), distribution to the examinees, collection of answer sheets and booklets after administration, and scoring. Between the mid-1930s and through the end of World War II, only the largest testing programs had access to machines that could scan the answer sheets and provide scores. The alternative for the vast majority of testing programs was hand scoring of the answer sheets: a template was placed over each answer sheet and the number of marked answers was manually counted – a procedure also fraught with the potential for error; moreover, the entire paper-based process was obviously inefficient and time-intensive.

Figure 1. Major Components of the Test Development Cycle



The Impact of Technology on Testing

As with many aspects of our lives after World War II, technology began to impact testing. The first impact was on the error-prone and labor intensive test scoring process. Less expensive optical scanners began to appear in the early 1950s that were capable of reading answer sheets, comparing the scanned answers to a set of correct or “keyed” answers and producing a score (and in some cases multiple scores) for each answer sheet scanned. The early machines were very large and expensive devices that were not true computers, but provided some basic functions that were similar to computers – they effectively were “business machines” designed for a specific purpose. They were slow and temperamental and sometimes had reliability problems, but were considerably more efficient, less expensive, and likely more accurate than the hand scoring process using templates. Because of their expense and temperament, the machines were maintained by specialized staff, and answer sheets had to be mailed to the scoring organization and results mailed back to the test user. Thus, this process eliminated the labor necessary to hand score answer sheets, but created delays both in the transport of the answer sheets and in their processing at the busy scanning centers.

Initially, these machines did not provide any group summaries of results. If an “item analysis” was desired, the item responses could be output on punch cards and the cards could be run through other “business machines” to obtain basic frequency counts that could be used to hand-compute classical item difficulties. Other statistics, such as point-biserial correlations, would have to computed by hand using calculators. Of course, the test development cycle would have to be completed by hand-entering the item statistics onto the backs of the index cards for review by the test developers so that poorly functioning items could be identified for exclusion or revision prior to use in future tests.

Computers began to be used in testing as they became more generally available in the early 1960s. Their first application was to replace the early scanners, providing somewhat more reliable scanning, faster scanning, and the capability to be programmed to incorporate item analysis results for a defined set of answer sheets. Computers also began to be used in that decade in some larger organizations (particularly universities) for more complete test analysis, including validity analyses and factor analysis. Although the early vacuum tube computers were somewhat unreliable and used rudimentary input-output devices (e.g., punched paper tape output) the introduction of the solid-state computer and more reliable input-output equipment improved their performance considerably. For the first time, psychometric analysis could be done without hand calculations, but the rest of the test development cycle still remained the same in 1970 as it had been for the last 50 years since the inception of the P&P test.

Minicomputers came on the scene in the mid-1970s as solid state computers began to shrink in size. These computers were one-tenth or less the size of the original solid state computers of the previous decade and extended computing power to many organizations and projects that did not have them previously. Their impact on testing was relatively minimal, with one exception noted below, except for making scanning and basic item analysis more widely available and less expensive.

THE PERSONAL COMPUTER IMPACTS TESTING

Major changes in the way tests were developed, analyzed, and delivered began to occur with the introduction of the personal computer (PC) in the mid-1980s. This impact can be divided into three phases – storing items, banking items and assembling tests, and delivering tests.

Storing Test Items

As a labor-saving device for the production of manuscripts and other documents, the PC came with word processing software that could be adapted for other purposes. Thus, one of the first uses of the PC in testing was to allow test developers to store test items in word processing software. Word processing software allowed test developers to type their items only once, then select them as needed to create a test. A new document would be opened and the text of stored items copied and pasted into the new document in the order in which the items were desired in the test. This process effectively removed the necessity to completely proofread a new test assembled from the master test item files, and made it easier to assemble alternate forms of tests when needed. All that was required from the test developer after a test was assembled was to check to see that all of the items were in the correct place, and that when the test was formatted that items did not break across pages – a considerable time saving from completely proofreading and correcting one or more forms of a test.

Some test developers adapted other standard PC software – notably spreadsheets – as item storage mechanisms. Again, the advantage was that item text could be stored, copied, and pasted to eliminate retyping. An added advantage was that the different pages of the spreadsheet could be used to separate items into subsets, perhaps representing the structure of a domain to be tested. A final advantage was that different cells in the spreadsheet could be used to store other data on the items and that information could be physically associated with the items, thus allowing both sides of an item “index card” to be stored together. Although word processors also allowed storage of the full “index card” for an item, spreadsheets were more flexible in their layout options and could more easily hold a wider variety of information on an item. Both word processors and spreadsheets allowed users to search for items that had specific values of an item statistic, but were generally limited to simple searches.

Item Banking and Test Assembly

The test development process improved dramatically as special purpose software was developed for item banking and test assembly (Vale, 2006). In the development of this type of software, an item bank was conceptualized as a database, and database software was programmed to perform the special functions necessary to maintain a testing program. Item text became one or more fields in a database, additional fields were defined for item statistics and other information, and the development of hierarchical structures to represent bank structures was facilitated by the item banking software.

DOS Item Banking Software

Specialized item banking software first began to appear in the mid-1980s using the DOS operating system on PCs. These were generally text-based bankers that had very limited graphics capability, since computer displays and printers of that era were primarily text oriented. Thus, with the exception of the MicroCAT Testing System (Assessment Systems Corporation, 1987) which had relatively advanced graphics for its era, item bankers in the late 1980s were limited to tests that used items almost entirely comprised of text. They also had very limited formatting capabilities, in terms of fonts and special effects, since these were not supported by the line printers available at that time.

Nevertheless, the DOS-based item bankers greatly improved the efficiency of test development. Some had search capabilities that allowed the test developer to search through stored item information to identify items that had specified sets of characteristics – frequently permitting searching on multiple variables simultaneously – to identify candidate items for a specific test. Some permitted limited test formatting and page numbering, and many permitted the easy creation of multiple forms of a test.

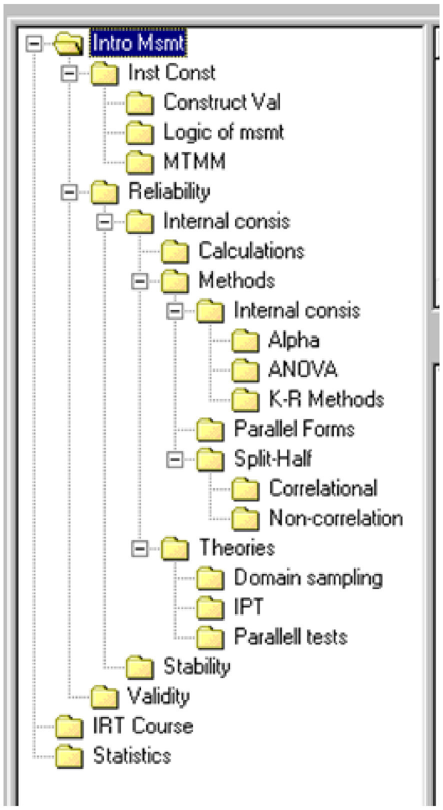
At the same time that item bankers were simplifying and streamlining the storage of items and item statistics, and permitting test developers to assemble tests with specific characteristics, some special-purpose software extended test assembly capabilities even further. Most notable among these was ConTEST (Timminga, van der Linden, & Schweizer, 1996) which was designed to solve complex test assembly problems that involved a large number of constraints. ConTEST used linear programming methods to create one or more test that satisfied all the constraints imposed. It operated from item statistics, however, and the resulting tests had to be manually assembled from separate bankers or databases.

Windows Item Bankers

While the DOS-based bankers began to change the way items were stored and tests were assembled, they were quite rudimentary compared to the Windows item bankers of the 21st century [e.g., PARTEST (http://www.scantron.com/parsystem/), LXRTEST (http://www.lxrtest.com/site/home.aspx), and FastTEST(Assessment Systems Corporation, 2010b)] . Windows item bankers usually incorporate a complete point-and-click interface to allow the test developer to interact with a database structure designed specifically for purposes of item banking and test assembly, and can incorporate a range of types of graphic displays in items. The capability of printing tests with these bankers was greatly enhanced by the widespread availability of PC-compatible laser printers, beginning around 1990.

The most useful item bankers allow item banks to be designed to reflect the structure of the domain to be tested, which is frequently operationalized in a test “blueprint.” The blueprint is usually an outline or a hierarchical structure that delineates the structure and subdomains of the primary domain, frequently with additional levels of specificity. The number of levels in a bank hierarchy is determined by the structure of the domain, sometimes combined with characteristics of the test items. Figure 2 shows an item bank structure from the FastTEST Test Development System (Assessment Systems Corporation, 2010b) for an introduction to psychological measurement item bank.

Figure 2. Bank Structure for an Introduction to Psychological Measurement Item Bank



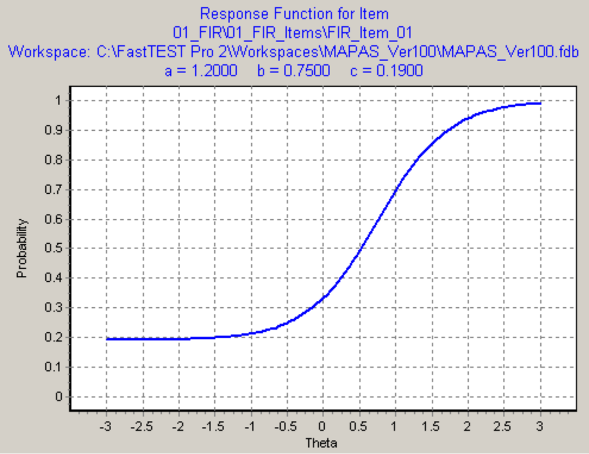
- Intro Msmt
  - Inst Const
    - Construct Val
    - Logic of msmt
    - MTMM
  - Reliability
    - Internal consis
      - Calculations
      - Methods
        - Internal consis
          - Alpha
          - ANOVA
          - K-R Methods
        - Parallel Forms
        - Split-Half
          - Correlational
          - Non-correlation
      - Theories
        - Domain sampling
        - IPT
        - Parallell tests
    - Stability
  - Validity
- IRT Course
- Statistics

In addition to storing item text and any related graphics, item bankers allow storage of other information associated with each item. These, of course, will include item statistics. Some bankers are designed only for use with classical test theory statistics – item difficulty (proportion correct) and item discrimination (biserial or point-biserial correlation) – while others allow storage of item parameters from item response theory (IRT; see chapter xxx) and display of IRT item functions (e.g., Figure 3).

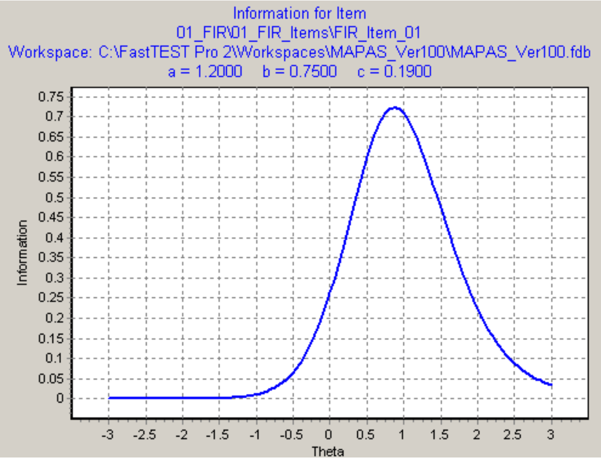
Other information stored on items includes the correct or keyed answer to the item, tests in which it has been used, name of the item writer and date created, special user-supplied statistics (e.g., Angoff rating), keywords that characterize the items, and other notes concerning the item. The information associated with each item is typically organized in a set of tabs for easy access. For example, FastTEST has five tabs: Item Identifier (including keywords and description), Item Text (using a full-featured built-in word processor), Item Information (item type, keyed responses, author, source), Statistics (both IRT and classical), and Notes.

Figure 3. FastTEST IRT Item Response Function and Item Information Function for an Item

a. Item Response Function



b. b. Item Information Function



Thus, using multi-part and multi-field records in a database system, Windows item bankers replicated, automated, extended, and greatly improved the efficiency of the functions of the index cards originally used for test item storage and retrieval. In addition, however, computer-based database systems also have the capability of permitting highly efficient and virtually error-free search and retrieval. Windows item bankers capitalize on these capabilities to permit efficient and effective test assembly.

For simple test assembly, a test of a specific number of items can be randomly selected from an item bank, or portions of an item bank. The latter approach would be used in successive cumulative searches to create a test that has a specific content structure with proportional representation of a larger content domain.

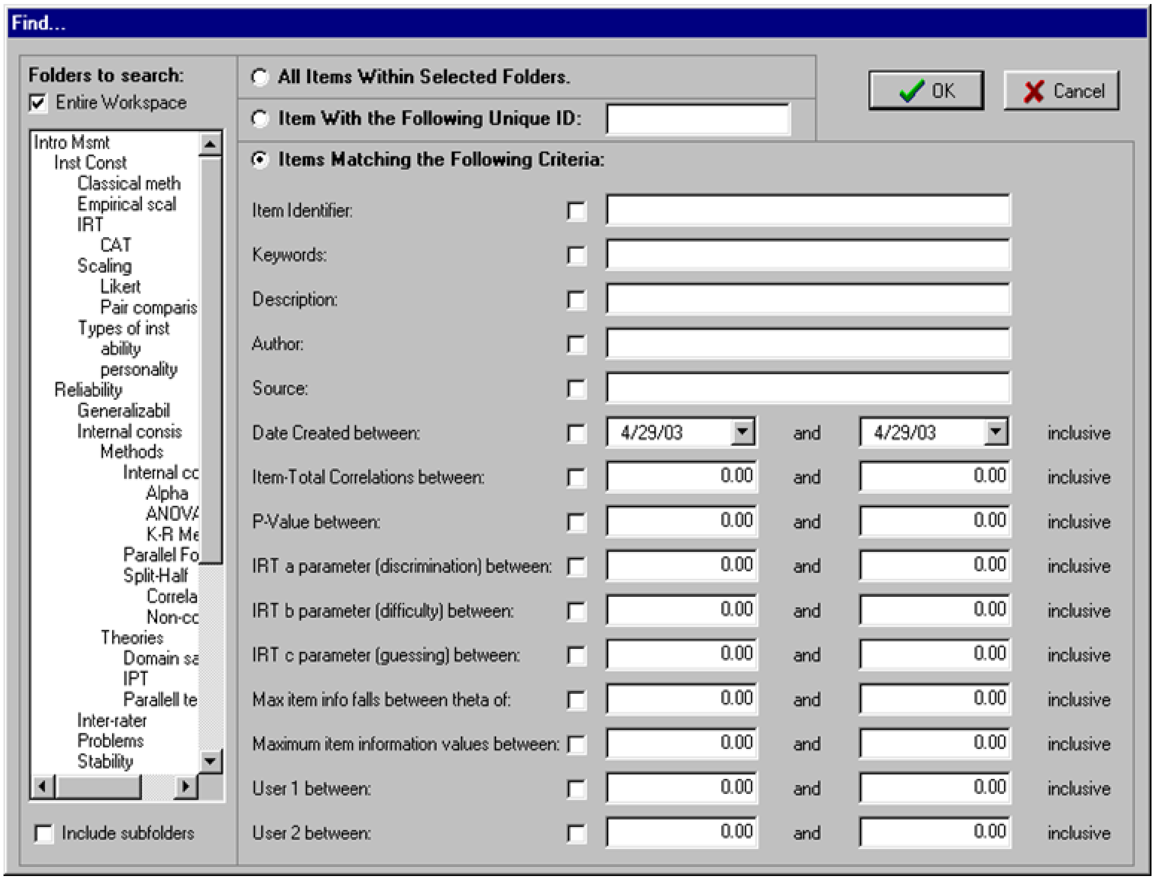
For constructing tests with deliberate non-random item selection, item bankers allow intelligent searching of information on items. Figure 4 show the item search window from FastTEST. As the figure shows, items in a bank, multiple banks, or portions of a bank can be specified to be searched. Searches can be implemented within most of the fields in the item record. Item identifiers, keywords, and item descriptions frequently include content or item type information that is not included in the item bank structure, allowing item subsets to be identified that have specific content or structural characteristics (e.g., all free-response items, if that information is included in any of these fields). In addition, Figure 4 shows that separate or simultaneous searches can be made on all the psychometric data stored for each item. For classical test assembly, item-total correlations (discriminations) in a given range can be searched for, while at the same time searching for items that have *p*-values (difficulties) in a desired range. Searches of this type can be combined with content searches either by restricting the portion of the item bank searched to a particular content subsection of the item bank or by simultaneously limiting the statistics search within item subsets that match content search criteria.

When IRT item statistics/parameters are available in the item records, item banks or portions thereof can be searched for various combinations and ranges of the IRT discrimination, difficulty, and pseudoguessing parameters. For more sophisticated IRT test assembly, FastTEST allows searches on item information, thereby helping the test developer create tests with a desired test information function. For example, a test developer might implement successive cumulative searches for items that have their maximum information values within specified ranges of the trait (θ) and for which the maximum information values are contained within a designated range. The result would be a set of items (if they existed in the bank) that had high maximum information throughout the θ range of the combined searches. In all cases, item bank searches frequently occur in a second or two, with slightly longer times for very large banks.

The result of an item search in an item banker is typically a list of items that meet the search criteria. Given that subset of items (akin to looking through the card file drawer and selecting a tentative set of items) the test developer then will usually select the items to include in the final test. This can be done in several ways. One approach is simply to randomly select a subset of items among the items that meet the search criteria. A second is to browse through the item text and other information on the items and manually select items from the searched pool of item candidates. In either case, items are added to the test with a simple click of a mouse.

Once the items that will comprise the test are selected, the next step frequently is to reorder the items in the test as desired. In a Windows item banker, this can be done by drag-and-drop within the item list that comprises the test, if a defined order is desired, or by randomly scrambling the items. If alternate forms of the test are needed, with a click of the mouse the test constructor can create any number of randomly scrambled alternate forms.

Figure 4. Item Search Options in FastTEST



Find...

Folders to search:

☑ Entire Workspace

Intro Msmt
  Inst Const
    Classical meth
    Empirical scal
    IRT
      CAT
    Scaling
      Likert
      Pair comparis
    Types of inst
      ability
      personality
  Reliability
    Generalizabil
    Internal consis
      Methods
        Internal cc
          Alpha
          ANOV/
          K-R Me
        Parallel Fo
        Split-Half
          Correla
          Non-cc
      Theories
        Domain sa
        IPT
        Parallell te
    Inter-rater
    Problems
    Stability

☐ Include subfolders

○ **All Items Within Selected Folders.**

○ **Item With the Following Unique ID:**

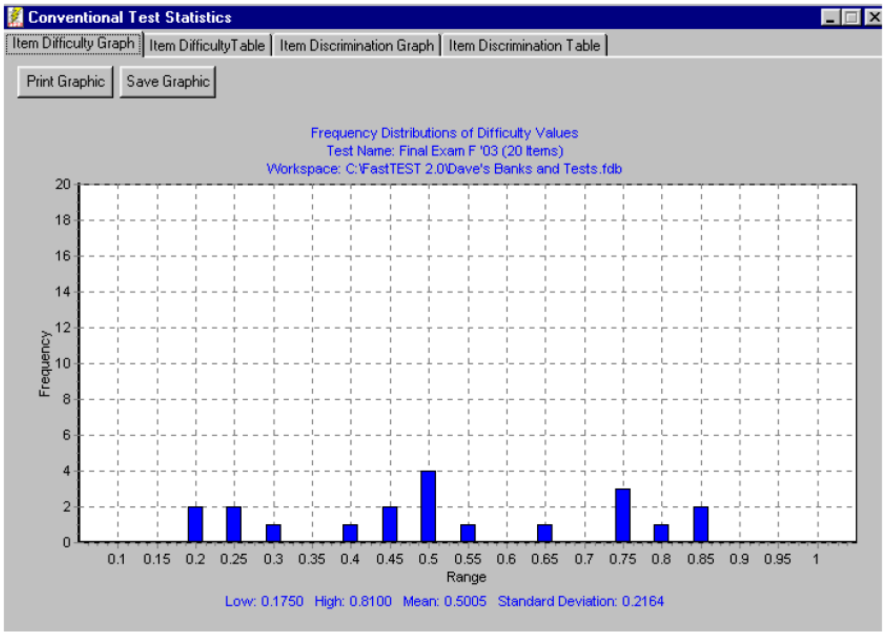
● **Items Matching the Following Criteria:**

Item Identifier:	☐				
Keywords:	☐				
Description:	☐				
Author:	☐				
Source:	☐				
Date Created between:	☐	4/29/03	and	4/29/03	inclusive
Item-Total Correlations between:	☐	0.00	and	0.00	inclusive
P-Value between:	☐	0.00	and	0.00	inclusive
IRT a parameter (discrimination) between:	☐	0.00	and	0.00	inclusive
IRT b parameter (difficulty) between:	☐	0.00	and	0.00	inclusive
IRT c parameter (guessing) between:	☐	0.00	and	0.00	inclusive
Max item info falls between theta of:	☐	0.00	and	0.00	inclusive
Maximum item information values between:	☐	0.00	and	0.00	inclusive
User 1 between:	☐	0.00	and	0.00	inclusive
User 2 between:	☐	0.00	and	0.00	inclusive

Before a test is finalized, a test developer might want to examine the statistical characteristics of the test based on item statistics in the item bank. A few mouse clicks will make this information available in either graphical or tabular form. Figure 5 shows a frequency distribution of classical item difficulties in a test assembled with FastTEST; a similar graphic is available for item discriminations. If the test developer is not satisfied with the statistical characteristics of the test (before it is administered) she/he can drop and add items and instantly reexamine the revised test's statistical characteristics.

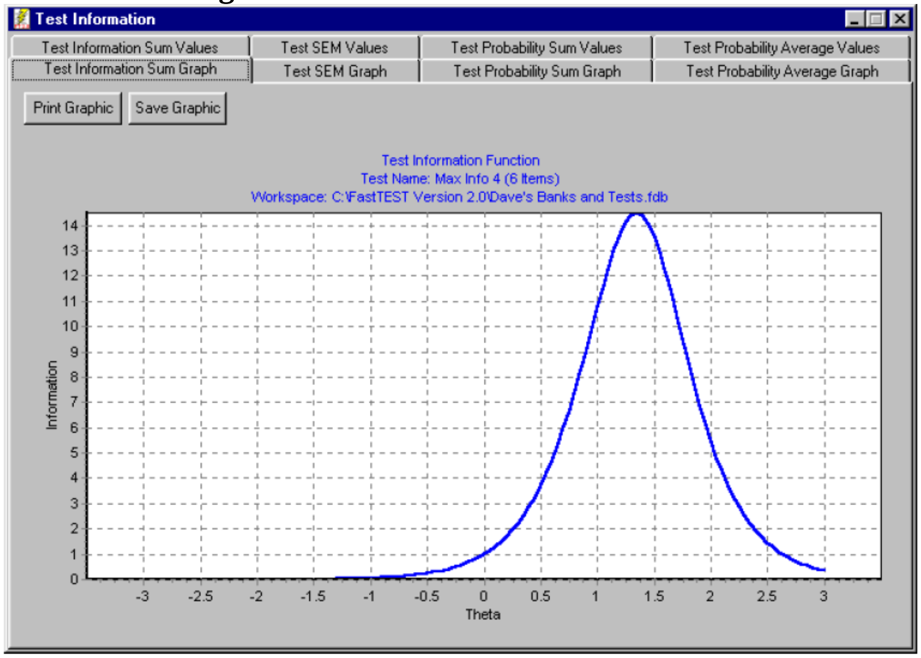
IRT item parameters are available, the banker can display a test information function (e.g., Figure 6), a test response function, and/or a test standard error function. The test information in Figure 6 shows that the test being assembled provides a considerable amount of information around a θ value of +1.3 and very little information elsewhere along the θ scale. It is, as designed, a good test for differentiating individuals who are below or above $\theta = 1.3$, but has little measurement precision outside a range of about ±1 standard deviation around that point. This test provides virtually no precision for θ values below average (0.0). Depending on the purpose for which the test was being built, revision of this test might be in order before it is used.

Figure 5. Frequency Distribution of Proportion Correct for a 20-Item Test


Conventional Test Statistics
Item Difficulty Graph
Item DifficultyTable
Item Discrimination Graph
Item Discrimination Table
Print Graphic
Save Graphic
Frequency Distributions of Difficulty Values
Test Name: Final Exam F '03 (20 Items)
Workspace: C:\FastTEST 2.0\Dave's Banks and Tests.fdb
Frequency
0
2
4
6
8
10
12
14
16
18
20
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
Range

Low: 0.1750 High: 0.8100 Mean: 0.5005 Standard Deviation: 0.2164

Figure 6. A Test Information Function


Test Information
Test Information Sum Values
Test SEM Values
Test Probability Sum Values
Test Probability Average Values
Test Information Sum Graph
Test SEM Graph
Test Probability Sum Graph
Test Probability Average Graph
Print Graphic
Save Graphic
Test Information Function
Test Name: Max Info 4 (6 Items)
Workspace: C:\FastTEST Version 2.0\Dave's Banks and Tests.fdb
Information
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
Theta

The final phase of item banking and test assembly is frequently one or more printed tests. Typically, item bankers will permit the insertion of instructions into the test document before it is printed. Most item bankers will also output a printed test with final or near final formatting. Some will output the test as an RTF file that then can be further formatted in a word processor prior to printing. They also will typically output a scoring key for each form of the test that they print. Of course, if no changes have been made in the items when they are formatted as the final test or when they are printed by the banker, no proofreading of item text is required.

The final component of the test development cycle is updating the item bank with item statistics from item analyses of the data from the P&P administered test. This can be done manually, with appropriate item statistics typed into the item record for each item in the bank. Or, more efficiently and accurately, item statistics output from item and test analysis software can automatically be imported into the item banker. For example, FastTEST includes a "wizard" that will import item statistics output from any item analysis software, thus completing the test development cycle.

Thus, the marriage of computer technology and database software designed specifically for testing has, in a short period in the history of testing, radically changed the way that tests can be developed. The key element is banking software that allows the user to create structured banks, search the banks on a wide range of criteria, and assemble tests based on both psychometric and content considerations. The process of creating a test has transitioned from a tedious and error-prone process that consumed many person hours to a simple process that can occur in a matter of minutes, once one or more properly constructed banks of test items have been entered into a well-developed item banking system.

Electronic Test Delivery

The major change in how tests are delivered was also a result of the introduction of the PC and, as for item bankers, with additional impetus from the availability of Windows software. Electronic testing, or computer-based testing (CBT), began in the early 1970s. CBT eliminates both printed tests and answer sheets – test questions are stored in the computer, displayed on a monitor, and answers are generally entered by keyboard and more recently by mouse. Early CBTs were delivered on mainframe time-shared computers (De Witt & Weiss, 1974). These were typically connected by dialup telephone modems operating at 10 or 30 characters per second connected to "dumb" character-based displays. It quickly became apparent, though, that this computer configuration was inadequate for test delivery. In addition to being limited to test items that were entirely character based, transmission and display time was far too slow and system response time of these early systems was far too unpredictable – processing of a single item response and transmission of the next item sometimes took 30 seconds or more.

In the mid-1970s minicomputers became available for testing research and were used for early delivery of adaptive tests by the author (e.g., De Witt & Weiss, 1976). Because these computers were dedicated to the single task of testing, and monitors were hardwired to the computer, system response time and display time were virtually instantaneous. They were, however, also limited to solely character-based test items. These systems, however, foreshadowed the primary improvements to be realized from CBT: (1) a fixed set of items could be administered in different orders to different examinees (2) different subsets of items could be administered to different examinees to achieve certain measurement objectives, (3) item response data were instantly captured and cumulated across examinees, and easily prepared for analysis, and (4) tests were immediately scored and individual reports could be prepared and available in seconds.

Randomized Tests

Randomized P&P test forms have been sometimes used in large testing programs to minimize copying among examinees in adjacent seats. In this application, two or three versions of a test are created with the base form randomized once or twice to create alternate forms. In CBT, the process of whole-test randomization can be extended to separate randomizations of item order for each examinee. This can be useful in CBT environments in which a number of examinees are taking the same test in the same computer lab, to minimize answer copying by students whose visual field might include another examinee's monitor.

A second form of CBT individualized randomization involves randomly selecting a subset of items from a larger domain of items. For example, an item bank might contain 200 items that define a specific content domain and any given examinee might receive 50 items randomly selected from that domain. This process results in a relatively unique set of items administered to each examinee (there will, of course, be random item overlap among examinees) and a random sequence of items administered to each examinee. A variation of random item selection uses a stratified approach to randomly select items from a domain that has been subdivided into subdomains. For example, a mathematics domain might be stratified by type of operation – addition, subtraction, multiplication, and division. A randomized CBT might be designed to administer ten items randomly selected from each subdomain to each examinee, for a test consisting of forty items. Both whole test randomization and subdomain randomization can be implemented with most PC-based testing systems [e.g., the FastTEST Professional Testing System (Assessment Systems Corporation, 2008) as well as Web-based testing systems [e.g., *FastTEST Web*™ (www.fasttestweb.com)].

Random item selection thus explicitly implements the concept of "domain sampling" commonly articulated as the basis for reliability theory using classical test development methods as well as minimizes answer copying in a CBT environment. The process is, however, contradictory to the classical process of constructing some tests. In the first 60 or so years of P&P testing, some tests were (and still are) built with items in increasing order of difficulty, on the assumption that examinees performed better when they had a sufficient number of easy items at the start of the test to reduce test anxiety. Obviously, either strictly or stratified randomized CBTs cannot easily accommodate this rationale. Little to no research has addressed the effects of item randomization on examinees and their test scores as compared to tests built to accommodate warmup effects.

Intelligent Item Selection

Contrasting with randomized item selection in CBT are tests that use intelligent item selection. There are three major types of these CBTs – linear-on-the-fly tests (LOFTs), sequential tests, and adaptive tests, each designed to implement different measurement objectives.

LOFTs. LOFTs are essentially fixed-length randomly selected tests with constraints (Thompson, 2008). They operate from a large item bank with IRT parameters available for each item. Items are pseudo-randomly selected, but the IRT parameters are used as the test is delivered to each examinee to monitor the psychometric characteristics of the test in real time, and the results are compared to psychometric targets defined in advance. As a result, tests for each examinee will have similar psychometric characteristics, but achieved using different subsets of items for each examinee. A major advantage is that of equalizing item exposure to increase the security of an item bank across tests that are administered over time to a large group of examinees.

Sequential tests. Sequential tests are typically designed to make classifications. These tests might be used in as school to make pass-pail decisions, in an employment context to make a decision to hire or not to hire, or in a professional certification program to determine whether an individual meets specified certification criteria. Although some sequential tests use random item selection, the more effective tests use intelligent item selection to the extent that psychometric information on test items is used to order items prior to item delivery. Then, given the fixed item order, items are administered and scored one at a time. After each item is administered a classification algorithm, such as the sequential probability ratio test (Eggen, 1999; Reckase, 1983), is used to attempt to make a classification of the examinee. If a classification can be made within prespecified error tolerances, test administration is terminated for that examinee. If a high-confidence classification cannot be made, the next item is administered and the decision criteria again re-evaluated. The result is tests that can make accurate classifications very efficiently, with a minimum number of items for each examinee.

Adaptive tests. Computerized adaptive tests (CATs) implement "fully intelligent" item selection (Wainer, 2000; Weiss, 1985, 2004). Unlike sequential tests that use a fixed order of items and allow only test length to vary, the more advanced versions of CATs also allow each examinee to start their tests with different items and to receive quite different sets of test items.

There are a number of varieties of CATs; to some degree they all dynamically select items to be administered to each examinee based on their answers to previous items in the test. Some use prestructured item banks in which an examinee's next item is determined by a branching tree structure in which a correct answer to a given item results in a particular next item and an incorrect answer leads to a different item. Others divide items into subsets or "testlets" (Mead, 2006; Wainer, Bradlow, &Du, 2000; Weiss, 1974). In this approach, each "testlet" or mini-test, is scored and based on that score a decision is made as to which testlet is to be administered next. In yet another approach, test items are stratified by item difficulty (Weiss, 1973) or discrimination (Chang & van der Linden, 2003) and administered sequentially within or between strata.

The most flexible and, therefore, efficient CATs are the fully adaptive CATs based on IRT. These CATs are based on IRT item information functions (e.g., Figure 3b), which are transformations of the IRT item parameters. The use of information functions allows each examinee to start their test with a different item if valid prior information is available. Then, based on their answer to that item, a score is computed for that examinee, expressed on the IRT trait scale (θ) using estimation methods that take into account which answer the examinee gave to the item (correct/incorrect, keyed/not keyed, or which rating scale alternative was selected) and the item parameters for that item. The updated score is then used to select the one unadministered item out of an entire bank that provides the most information for that examinee, which is also the item that maximally reduces the uncertainly associated with the θ estimate (as expressed in the individualized standard error of measurement associated with the θ estimate). One or more termination criteria are them consulted – these are typically a specified minimum value of the standard error and/or some maximum number of items. If the examinee has not met one of the termination criteria, the current θ estimate is used to select the next best item and the process continues. When a termination criterion is met, the test is ended and the final θ estimate and its standard error are recorded for that examinee.

CAT was first implemented primarily in the ability/achievement domain (Weiss & Betz, 1973). In recent years it has begun to be used in personality measurement (Reise & Henson, 2000) and in medical research by measuring patient-reported outcomes of medical processes and procedures (Reeve et al. 2007). Early CAT research in the ability/achievement domain (e.g., Kingsbury & Weiss, 1983; McBride & Martin, 1983) indicated that CATs could measure with equal precision to that of conventional tests using at least 50% fewer items; these findings have been supported and extended in numerous applications (e.g., Mardberg & Carlstedt, 1998; Moreno & Segall, 1997). More recent research in the personality and mental health domains indicates that reductions in test length as high as 95% can be obtained on a general impairment scale, and 85% for measuring four subscales, with little or no reduction in measurement accuracy from full-length tests that use an entire large item bank (Gibbons et al., 2008).

The major advantage of CAT is the capability of designing and delivering efficient tests that measure all examinees with an equal level of precision. This means that in a CAT properly designed for this measurement objective, all examinees will be measured with the same standard error and minimum test length, an objective not easily achieved with any other kind of test. Obviously, because of the extensive real-time calculations necessary to implement CATs, they cannot be delivered by any other means than computers. The FastTEST Professional Testing System (Assessment Systems Corporation, 2008), in conjunction with CATSim (Assessment Systems Corporation, 2010a) permit the design and delivery of fully adaptive tests using IRT given an item bank of items with estimated IRT parameters.

Other Advantages of CBT

Data capture. Because all forms of CBT involve electronic item delivery and the immediate electronic capture of item responses, all of the problems associated with printing and distributing test booklets and answer sheets, as well as the unreliability of the scanning process, have disappeared. Item response data in CBT are stored as each examinee answers each item and can be accumulated across examinees with ease. If the test are randomized, sequential, or adaptive, responses are automatically reordered into a common order to allow analysis. Depending on the software system, cumulated item responses can be immediately analyzed and the results available at any time, even on a real-time basis if desired.

An additional potential advantage of CBT is the availability of item response times. The PC can record the time from when the item is presented to the examinee to when she/he clicks the mouse to select an answer and/or clicks the “next” button to move to the next question. Although such item response times can be somewhat unreliable, careful analysis of them might result in additional information, beyond the correctness/incorrectness of an examinee’s answers to assist in obtaining better measurements of the examinee’s ability, attitude, or personality variables (Ferrando & Lorenzo-Seva, 2007).

Instant reporting. In addition to instant capture of item responses and instant scoring, CBT provides the capability of generating a wide variety of reports that can be displayed immediately to the examinee on completion of the test session (which can include multiple tests); to a testing room supervisor, proctor, or teacher; or can be printed or saved in electronic files for later use. These reports can be as simple as a certificate of completion of the test with a passing score, or as complex as a graphic plot of multiple test scores followed by a multi-page interpretation of test results. Obviously the combination of instant data capture and instant reporting permit test data and test results to be used for applied purposes far more quickly than was possible with P&P tests.

New types of measurements. A final major advantage of CBT is the capability of measuring variables that cannot be easily measured with P&P. This capability includes the use of detailed color graphics in test items, audio and video, and animation. The use of audio and video is especially useful in certain types of language testing in which language segments are spoken through headsets and presented to examinees for translation and other processing. Other language-related applications include presenting test items with an audio or video option to examinees who have reading limitations. More recently, innovative item types have expanded upon simple multimedia to include interactive simulations in an attempt to provide more fidelity to the measurement of complex processes.

CBT also allows the measurement of some abilities and other variables for which P&P tests are not optimal. For example, although memory is an important ability for success in academic environments and many jobs, there have been no major P&P tests of memory ability because measuring memory requires an interactive, individualized, and controlled process that would be very labor intensive. Such tests, however, could be easily developed in a CBT environment (e.g., Letz, 2003) where it is possible to control the period of time that material is displayed, to use a wide variety of material – words, phrases, audio clips, and video clips – and test for recall after specified time intervals. The process can also be made adaptive in that display and recall times could be individualized for an examinee based on their performance on earlier tasks.

Another type of new item that is uniquely computer-administered is the so-called "scenario" or problem-solving item. In this kind of test, a situation is described and the examinee is given a choice of various elements of information that pertain to the situation . After consulting the selected information, questions are posed to the examinee that then lead to other information sources. The process continues until some resolution is reached which, depending on the sequence of choices made by the examinee, could result in an adequate solution to the original problem, or solutions that are inadequate to various degrees (and, therefore, result in lower scores). This kind of interactive problem-solving test is most notable in the medical training (e.g., Dieckmann, Lippert, Glavin, &Rall, 2010) and licensing (e.g., www.nmbe.org) environment where the "patient" presented in the original scenario is either "cured" or dies, or the sequence of choices made by an examinee results in some intermediate suboptimal state.

THE INTERNET IMPACTS TESTING

As with many functions performed by PCs, the rise of the Internet and the World Wide Web began to affect the test development and delivery process beginning in the late 1990s, as it has impacted many other areas of psychological research (e.g., Gosling & Johnson, 2010). Test development is frequently a process that draws on the expertise of a variety of personnel. In a large testing program, such as that of a school system or a licensing or certification organization, test items are written by a number of people with specific expertise, some of whom are geographically dispersed throughout a country or even different countries. Although it is possible to collect test items from remote experts by sending email files to a central location for item bank development, the Internet presented an opportunity to allow test items to be entered into item banks from any computer with access to it. Thus, once an item bank is developed, software systems such as *FastTEST Web*™ (www.fasttestweb.com) allow item writers (with appropriate security safeguards) to access designated portions of item banks and to directly add new items to the banks.

A second stage of item bank development, also available for remote access in systems like *FastTEST Web*™, is item review and editing. Item reviewers and editors have different skill sets than item writers and are, therefore, likely to be different personnel and located at different places. *FastTEST Web*™ defines a different “role” for item editors and allows the test development supervisor to limit their activities to only items that are appropriate for review. Other roles include test assembly and test (versus item) review; each activity can be done remotely by any number of appropriate personnel at any location without the need for sending any material to a central location for processing. The result is an item and test development process that is even more efficient than that possible by the use of PC item bankers. Of course, Internet item bankers such as *FastTEST Web*™ include all the functionality in item banking and test assembly as the PC-based item bankers, plus the capability of running a wide range of reports on banks and tests from any location. In addition, the tests developed through Web-based bankers can be printed or delivered directly through the Web to examinees at any location.

A major characteristic of PC-based CBT is that the test and test data are stored on individual computers on which tests are being administered, or on a network server that is hardwired to the testing computers. When tests are delivered on standalone computers, test data must be collected from each computer and aggregated for further storage and analysis. Although this process is easily automated to a degree, it still requires physical transmission of test data by some means. In addition, when the tests themselves are stored on independent testing stations, they must be individually installed and their existence on testing station hard drives can create potential item security problems unless the tests are well encrypted.

Internet test delivery solves these problems, although not without creating some others. In Internet testing, which has become very popular in recent years, tests are stored centrally, along with all the information necessary to score them (e.g., IRT item parameters or classical item option score weights). Items are sent through the Web, one or more at a time, presented to the examinee, and the response is accepted and transmitted back to the server. The next item, or set of items, is selected and presented and the process is continued until the test is completed. At the end of the test, as in PC-based testing, an assortment of reports is available for presentation to the examinee and/or other appropriate personnel. The advantages are, of course, that tests can be delivered to any computer that has Internet access, test items are not stored on the testing computer but rather appear only on the monitor screen, and all test data are instantly stored in a central database and are available for analysis at any time.

A number of new problems are raised by Internet testing, however (Naglieri et al., 2004). In PC-based testing, which typically has been implemented in testing centers or testing labs, monitors and other associated equipment can easily be standardized. Standardization involves specifying a defined set of conditions under which the measurements are obtained that are designed to control extraneous influences which might affect the measurements (and add error to them). Internet-delivered tests, however, frequently are administered to individuals using their own computers, which can be desktops, laptops, or notepads. These computers might have different display resolutions and different display sizes. As a consequence, the same test item might render differently on different computers. In the P&P testing era, standardization of the test material was heavily emphasized – a given test was always formatted and printed exactly the same way. With Internet testing, because of the various displays possible under certain remote test administration conditions, there is a danger that the characteristics of the display will degrade the standardization required for adequate measurement of some variables.

A second threat to standardization in Internet test delivery is that of the testing environment. P&P testing and the majority of applications of PC-based testing emphasized that test delivery should occur in a quiet well-controlled environment in which examinees could concentrate on the tasks and questions posed in the test with minimal distraction. Test instructions were standardized, lighting and room temperature were controlled, and other outside influences were eliminated or minimized. Unless an Internet-delivered test is administered in a space devoted to testing or a location that is under the supervision of a test administrator, there is no control over the testing environment. To the extent that test scores can be influenced by non-standardized testing conditions – noise, other people present, variations in temperature and lighting, and a host of other factors that might exist in non-standardized testing environments – scores from such tests cannot be relied upon to be as precise and valid as tests taken under controlled conditions.

When Internet-based tests are delivered in unsupervised environments, it is also frequently not possible to know exactly who is taking the test or what they are doing during test delivery. There might be other people available to the examinee who are being consulted during test administration, or in an extreme case someone other than the presumed examinee might complete the test or a portion of it. In addition, unless the test delivery software explicitly locks the examinee's computer from accessing its hard drive and simultaneously locks the Internet browser from accessing other Web sites, an examinee might access other electronic sources during the test in order to answer the test questions. Even under complete electronic lockout, an unsupervised examinee can access printed sources without the knowledge of the organization providing the test. It, thus, should be clear that unsupervised Internet test delivery is not appropriate for "high stakes" tests that are being used to make important decision about an examinee.

A final source of lack of standardization of Internet-delivered tests lies in the nature of the Internet itself. The Internet is basically an extremely sophisticated time-sharing system, but one that involves a great number of loosely networked computers. As such, there are always delays between when information is sent to when the sending computer receives the information that it requests. These delays can be minimal – a second or two – or quite a bit longer. But they always are unpredictable. They result from a combination of many factors, including the amount of traffic on the Internet, the speed of transmission over the various components of the system used for a message to reach its destination server and return, the speed of the server and the load on it when the message is received, and server processing time.

For many testing purposes, these delays might be relatively inconsequential, especially since many people have become accustomed to them. But the delays can accumulate in testing applications where items are delivered one item at a time, such as sequential testing and adaptive testing. In these applications, in addition to the system delays there are computations that must be done between each item delivery, thus potentially exacerbating the delays. Delays of several seconds between items can result in a testing experience that is less than optimal for many examinees, and their unpredictability might be a source of test anxiety for some examinees.

CONCLUSIONS

The first 60 or so years of psychological, educational, and personnel testing were dominated by the paper-and-pencil test. Item banking, test assembly, and test scoring were entirely manual procedures that were labor intensive, tedious, and prone to errors. Tests were highly standardized, as were conditions of administration. Changes began to occur with the introduction of electronic optical mark readers which reduced test scoring to a relatively accurate partially automated procedure that dominated standardized testing for many decades. But the introduction of the personal computer in the mid-1980s began a major evolution of testing away from the traditional way of building tests and delivering them.

The years since 1985 have seen computers automate the processes of item banking, test assembly, test analysis, and test delivery. The personal computer allowed the development of new modes of test delivery – random, sequential, and adaptive – and new kinds of test items, The advent of the Internet extended test delivery to any computer that could connect to it, albeit not without some problems.

The result of this evolution of testing is a set of processes that are considerably less labor intensive, more accurate, and more efficient. In the process of this ongoing conversion, a number of questions have arisen, some of which have not yet been satisfactorily studied or even addressed. There are few research questions surrounding computerized item banking and test assembly. The major questions have risen in the context of CBT. In the early days of CBT, it was natural to address the question as to whether CBTs functioned the same as P&P tests. Generally, it was found that they did (Mead & Drasgow, 1993), although early research indicated some differences on reading comprehension tests (e.g., Kiely, Zara, & Weiss, 1986; Mazzeo and Harvey, 1988). However, as CBTs begin to be used to measure constructs that cannot be measured by P&P comparability is no longer an issues and CBTs will have to be validated on their own merits.

There are obviously a host of questions about how to best implement CATs and sequential tests that have resulted in substantial research over the last 20 or 30 years and will continue to do so (for an extensive bibliography of CAT research see http://www.psych.umn.edu/psylabs/catcentral/). CBTs also raise a number of questions about the psychological environment of testing that generally have not been addressed. In the process of creating a test of appropriate difficulty for each examinee, CATs create a different psychological environment than do P&P tests. Does that difference affect examinee performance? PC-delivered tests have virtually no delays between items in comparison to Internet-delivered tests. Do the unpredictable delays in Internet-based testing affect examinees' test anxiety and thereby influence test performance? Do the random variations in the testing environment that occur for unsupervised Internet-based testing affect test scores?

Finally, as the Internet continues to pervade our activities through various electronic devices, some have suggested that certain kinds of psychological measurements (e.g., attitudes, personality variables) can be delivered by portable electronic devices such as PDAs and cellular phones. If these modes of test delivery are implemented, the usefulness of the resulting measurements will have to be carefully scrutinized because of the extremely variable testing conditions under which such measurements will be obtained. As the APA Task Force on Psychological Testing on the Internet (Naglieri, et al., 2004) concluded,

> Despite the flash and sparkle of Internet testing, critical questions of the validity of the inferences made from test scores must be demonstrated. This is a fundamental issue of test validity that must be weighed in relation to the ease of availability, cost, and convenience of Internet testing. (p.161)

REFERENCES

Assessment Systems Corporation (1987). *User's manual for the MicroCAT Testing System (*2nd ed.*).* St. Paul MN: Author.

Assessment Systems Corporation (2010a). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul MN: Author. Available from http://www.assess.com.

Assessment Systems Corporation (2010b). *User's manual for the FastTEST 2.0 Item Banking and Test Development System.* St. Paul MN: Author. Available from http://www.assess.com.

Assessment Systems Corporation (2008). *Manual for the FastTEST Professional Testing System, Version 2.* St. Paul MN: Author. Available from http://www.assess.com.

Chang, H.-H. & van der Linden, W. J. (2003) Optimal stratification of item pools in *a*-stratified computerized adaptive testing. *Applied Psychological Measurement, 27,* 262-274.

De Witt, J. J. & Weiss, D. J. (1974). *A computer software system for adaptive ability measurement* (Research Report 74-1). Minneapolis MN: University of Minnesota, Department of Psychology, Computerized Adaptive Testing Laboratory.

DeWitt, L. J. & Weiss, D. J. (1976). Hardware and software evolution of an adaptive ability measurement system. *Behavior Research Methods and Instrumentation, 8,* 104-107.

Dieckmann, P., Lippert, A., Glavin, R., & Rall, M. When things do not go as expected: Scenario life savers. *Simulation in Healthcare, 5,* 219-225.

DuBois, P. H. (1970). *A history of psychological testing.* Boston: Allyn and Bacon.

Gibbons, R. D., Weiss, D. J., Kupfer, D. J., Frank, E., Fagiolini, A., Grochocinski, V. J., Bhaumik, D., K., Stover, A., Bock, R. D., Immekus, J. C. (2008). Using computerized adaptive testing to reduce the burden of mental health assessment. *Psychiatric Services, 59(4),* 49-58.

Gosling, S. D. & Johnson, J. A. (Eds.) (2010). *Advanced methods for conducting online research.* Washington D.C.: American Psychological Association.

Kiely, G. L., Zara, A. R., & Weiss, D. J. (1986). *Equivalence of computer and paper-and-pencil Armed Services Vocational Aptitude Battery tests* (AFHRL-TP-86-13). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Kingsbury, G. G. & Weiss, D. J. (1983). A comparison of IRT-based mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Eggen, T. J. H. M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement, 23*, 249-261

Ferrando, P. J. & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31,* 525–543.

Letz, R. (2003). Continuing challenges for computer-based neuropsychological tests. NeuroToxicology, *24*, 479–489.

Mardberg, B. & Carlstedt, B. (1998). Swedish Enlistment Battery: Construct validity and latent variable estimation of cognitive abilities by the CAT-SEB. *International Journal of Selection and Assessment, 6,* 107-114.

Mazzeo, J., & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature.* College Board Report No. 88-8. ETS RR No. 88-21.

McBride, J. R. & Martin, J. T. (1983) . Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Mead, A.D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19,*185-187

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114,* 449-458.

Moreno, K. E., & Segall, O. D. (1997). Reliability and construct validity of CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.). *Computerized adaptive testing: From inquiry to operation* (pp. 169-179). Washington DC: American Psychological Association.

Naglieri, J. A., Drasgow, F., Schmit, M., Handler, L., Prifitera, A., Margolis, A., & Velasquez, R. (2004). Psychological testing on the Internet: New problems, old issues. *American Psychologist, 59,* 150-162.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait theory and computerized adaptive testing* (pp. 237-254). New York: Academic Press.

Reeve, B. B., Hays, R. D., Bjorner, J. B., Cook, K. F., Crane, P. K., Teresi, J.A.., Thissen, D., Revicki, D. A., Weiss, D. J., Hambleton, R. K., Liu, H., Gershon, R., Reise, S. P., Lai, J., Cella, D. (2007). Psychometric evaluation and calibration of health-related quality of life item banks: Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). In The Patient-Reported Outcomes Measurement Information System (PROMIS) Overview and Developmental Work, 2004-2006. *Medical Care,45(5)* Suppl 1:S22-S31.

Reise, S. P., & Henson, J. M. (2000). Computerization and adaptive administration of the NEO PI-R. *Assessment, 7,* 347-364.

Thompson, N. A. (2008). A proposed framework of test administration methods. Journal of Applied Testing Technology, *9(5).* Available at http://www.testpublishers.org/mc/page.do?sitePageId=112031&orgId=atpu

Timminga, E., van der Linden, W. J., & Schweizer, D. A. (1996). *ConTEST 2.0: A decision support system for item banking and optimal test assembly* [Computer program and manual]. Groningen, The Netherlands: iec ProGAMMA.

Vale, C. D. (2006). Computerized item banking. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development.* Mahwah NJ: Erlbaum.

Wainer, H. (2000) (Ed.). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale NJ: Erlbaum.

Wainer, H., Bradlow, E. T., and Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245-270). Norwell MA: Kluwer.

Weiss, D. J. (1973). *The stratified adaptive computerized ability test* (Research Report 73-3). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Weiss, D. J. (1974). *Strategies of adaptive ability measurement* (Research Report 74-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program

Weiss, D. J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*, 774-789

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37(2), 70-84.*

Weiss, D. J. & Betz, N. E. (1973). *Ability measurement: Conventional or adaptive?* (Research Report 73-1). University of Minnesota, Department of Psychology, Psychometric Methods Program.