Measurement at the Extremes of Ability:
A Design for a Computerized Adaptive Intelligence Test

C. David Vale
Assessment Systems Corporation
2233 University Avenue, Suite 310
St. Paul, Minnesota 55114
(612) 647-9220

Paper presented in the Division 33 invited symposium entitled "Computer-Based Testing" at the 92nd Annual Convention of the American Psychological Association at Toronto, Ontario,Canada, August 1984.

Adaptive testing is not a new concept; as applied to psychological test administration, the concept dates back to Binet and his development of the first practical intelligence test. The concept, stripped to its basics, is that the assessment of psychological abilities can be done more efficiently if the difficulties of the items (i.e., questions) administered are tailored to the ability level of the examinee tested. Capable, or more mature, examinees are given more difficult items while less capable, or younger, examinees are given less difficult ones. In the Binet test, this is accomplished by first placing the test items in strata and then administering them sequentially, progressing up and/or down through strata until the examinee can answer all items correctly at a basal level and none correctly at a ceiling level.

Although the concept is nearly 80 years old, the psychometric community has joined the clinical community in applying the concept only in the last 25 years. In that period, however, a number of improvements have been made in the basic concept. First, a family of mathematical models called item response theories (IRT; Lord, 1980) has been developed to provide the groundwork that allows common scores to be efficiently derived from different sets of items. Second, a body of research has shown how the item response theories can be used to efficiently select and score the items (Weiss, 1982). Finally, the computer equipment required to present the items and to perform the computations involved in item selection and scoring has become sufficiently inexpensive that it is feasible to implement the adaptive testing techniques developed in the last 25 years.

I have been involved in adaptive testing for 10 years, on the psychometric side. In recent years, I have been involved in the production of test items for what will probably be the first large-scale application of the new adaptive testing technology: the implementation of a computerized adaptive version of the Armed Services Vocational Aptitude

Battery (ASVAB). More recently, my colleagues and I at Assessment Systems Corporation explored the potential clinical applications of the technology (Vale & Prestwood, Note 1).

Application of Computerized Adaptive Testing to Intelligence Measurement

Computerized adaptive testing (CAT) is most useful in an environment where a test must be used for a wide range of ability, wide enough that a constant set of items contained in a conventional test would include some inappropriate items for everyone tested. Intelligence measurement appeared to be an area that could benefit from an application of the technology. Since most of the individually administered intelligence tests are already, to some degree, adaptive, it seemed plausible that intelligence testing would readily benefit from the recent improvements in the technology.

Several benefits would immediately accrue from the application of the new adaptive technology to intelligence measurement. First, the substantial improvements in scoring and item selection technologies, as compared to those available in current individualized intelligence tests, would provide for more accurate estimates of levels of intelligence. Second, the substitution of the computer for the psychometrist would substantially reduce the cost of administration and allow an individualized (i.e., CAT) test to be used where only group tests had been previously affordable. Furthermore, the computer could maintain a degree of standardization of administration impossible for a human examiner.

There are a number of challenges to the development of a CAT intelligence test. Here I will address only two: the practical problem of item presentation and response acceptance by a computer and the psychometric problem of multidimensionality of intelligence test items.

Item presentation and response acceptance. Computers, as their name implies, were originally designed to do computation rather than to emulate a human. Although giant strides have been made in developing "user friendly" and "artificially intelligent" computer systems, a computer is still very clumsy in interpersonal interaction. A successful computerized test must be designed for administration by computer; the adaptation of an individualized test, previously given by a human examiner, is almost certainly doomed to failure.

Anyone attempting to computerize a current individualized intelligence test would immediately run into at least three problems. First, the computer would have a difficult time establishing rapport with the client. Second, it would have a difficult time presenting pictures with the same degree of resolution as in the original mode. Readily available microcomputers deal reasonably well with item presentation. Graphic capabilities sufficient to present clear figures are available in the popular systems typically available in the schools. (Typical screen resolution is about 300 by 200 dots.) These capabilities are not sufficient to present digitized versions of pictures in current tests without substantial degradation, however. Finally, it would have a difficult time scoring the free, and often verbal, responses made by the examinees. Synthesized voice output of acceptable quality is available for some of these systems. It is rarely standard equipment, however.

Multidimensionality. The second challenge to overcome in the development of a CAT intelligence test is that of multidimensionality in the test items. Almost all of the viable item response theories available today require that all of the items in a subtest measure a single characteristic. This is not as simple as to say that all items must measure intelligence; the characteristic must be factor-pure in that performance on the item is solely a function of the item and the examinee's standing on the underlying characteristic. An example where dimensionality problems may arise is in a reasoning item that consists of an arithmetic story problem. An examinee may fail the item because s/he cannot reason the solution, cannot do the arithmetic, or cannot read. The typical solution to this dimensionality problem is to keep the arithmetic and reading difficulties so low relative to the reasoning difficulty that they are not salient sources of difficulty.

The problem of dimensionality is more complicated in intelligence than in some other areas of testing because of the wide range of ability that must be tested. In the example given above, it may not be possible to make the arithmetic and reading difficulties low enough that they are not salient to the solution. How, for example, can reasoning be assessed before arithmetic is taught? It is also quite possible that difficult story problems assess a different kind of reasoning than do easy ones.

Design of a Computerized Adaptive Intelligence Test

In designing a computerized adaptive intelligence test, we set several design goals. First, the content of the test had to be consistent with the state of the art in intelligence assessment. We did not want to limit the content to that which had been included in tests developed 40 years ago. On the other hand, we did not want to surpass the state of the art, suggesting intriguing new item types that could be administered only on a computer, for example, or basing a design on a new theory that was elegant but untested; we felt it important that the first CAT intelligence test limit its innovations to the previously tested psychometric technologies.

Second, the item types had to be amenable to administration on microcomputer equipment that would likely be in place within the next two years. Thus, the item types included in the initial test could not assume the existence of advanced technologies such as voice recognition or extremely high display resolution.

Third, the characteristics of the items had to be consistent with the psychometric models available. Although breakthroughs in multidimensional IRT models are likely to occur in the next few years, items had to be amenable to unidimensional parameterization. This restriction does not preclude improvements to the items when such models become available; unidimensional models are special cases of more general multidimensional ones.

Our first step in the design of the test was to explicate and integrate the domain of ability that was to be called intelligence, as defined for the new test. We approached this definition by investigating operational, semantic, and theoretical definitions of intelligence. Operational definitions, specifically the content of previous intelligence tests, provided a de facto explication of the domain (i.e., intelligence is that which the tests measure). Semantic definitions, convenient handles for the concept when describing intelligence to laypersons, were of little help in explicating the construct. Definitions of intelligence based on theoretical models did, however, provide some useful insights into cognitive processes as well as factorial interrelations.

We investigated the operational definition of intelligence by examining twenty current intelligence tests included in Buros (1978). By intuitively clustering the subtests, it appeared that the most frequently assessed areas could be called Vocabulary (e.g., synonyms, picture recognition), Verbal Reasoning (e.g., analogies, categories, sentence completion), Nonverbal Reasoning (e.g, analogies or categories using figures or pictures rather than words), and Quantitative Reasoning (e.g., story problems, number series). Less frequent but still important item types included arithmetic, perceptual-motor, spatial perception, information, and memory. The tests and the resulting intuitive clusters are presented in Tables 1 and 2.

We investigated the theoretical definitions of intelligence by reviewing the factor analytic and cognitive literature on intellectual tests. We found as many factor solutions to the structure of intelligence as we found researchers. The cognitive literature, in general, provided insights into the details of some components of intelligence, but did not tie all of them together. Work by Chi (1976), for example, suggested the form of the process underlying tests of short-term memory while that of others (e.g., Merkel & Hall, 1982) suggested the best ways for measuring memory. Sternberg (1980, 1981, 1982) offered an elegant model for integrating the components of intelligence but had not completed the massive task of integrating the details and research findings into it.

Our venture into a definition of the concept provided a loose assembly of psychometric factors, cognitive processes, psychological tests, and eloquent semantics that did not form an obvious, coherent model of the construct. This led us to a practical step of imposing an intuitive model that, while offering no assurance of uniqueness, nevertheless integrated a large body of the data into a coherent framework.

Most of the major intelligence tasks appeared to implicitly relate to facility in dealing with concepts. We thus defined intelligence as the ability to perceive, organize, and manipulate concepts. We further defined a concept as a coherent mental construct that represents a class of objects or other concepts and specifies commonalities and relationships among the elements that it represents.

From that definition, we selected seven subtests in four areas of content for this test. In keeping with our plan to keep the innovations in this design psychometric, the general forms of all the item types we chose had been used previously in other intelligence tests.

Concept Assimilation represents the ability to rapidly recognize a concept and hold it in memory. This area is closely related to traditional tests of short-term memory. However, recent cognitive research suggests that these tests assess the effectiveness of an individual's strategies for locating and activating representations and concepts in long-term memory. Consequently, our design incorporates items that assess these locating and activating abilities. Specifically, our design calls for a test of semantic memory which is a variant of the digit span test that manipulates difficulty by the complexity of the chunks to be memorized and a test of spatial memory that manipulates difficulty by the complexity of the objects in the field.

Concept Manipulation concerns the ability to recognize the commonalities and relationships among concepts represented by words, pictures, or figures and to restructure or manipulate the concepts or relationships to achieve some goal. We propose to assess this area with analogy and category (e.g., oddity), possibly both containing items of verbal, figural, and pictorial formats.

Concept Organization refers to the ability to assimilate and organize a body of information and to manipulate that organization in order to solve a problem. The assimilation and manipulation components are similar to those discussed above; the organization skill will be the difficulty factor of primary salience in these items, however. The most notable difference between concept manipulation and concept organization items is that the concept organization items require the examinee to deal with a large amount of information. This area is closely related to what has been called reasoning or problem solving in previous tests. We plan to assess Concept Organization in both quantitative and analytical environments. The quantitative environment will contain arithmetic story problems of the conventional type. As with all good story problems, the arithmetic and reading requirements will be kept to a minimum. The analytical environment will contain items similar to those found in the Analytical Reasoning section of the Graduate Record Examination. These items present interrelated but unstructured pieces of information. The examinee's task is to structure the information and to draw inferences from the structure.

The final category, Concept Repertoire, does not flow directly from the semantic definition of intelligence offered above but rather from the implicit need for examinees to possess a set or repertoire of concepts that can be manipulated, organized, or used in assimilating new concepts. Two tests appeared as candidates for assessing this part of the construct: vocabulary and general information. Because the vocabulary format was more commonly encountered in previous tests and because difficult items in general-information tests tend to contain esoteric information outside of the experiential realm of many examinees, the vocabulary format was chosen.

Problems Remaining

Few of the subtests chosen are likely to be unidimensional over an unlimited ability range. Since they must be forced into a unidimensional model for the first version of the test, some limitations must be applied to this initial form. First, sufficiently unidimensional items can be written in each of the areas if the range of ability over which the test is used is somewhat restricted; any dramatic multidimensionality problems are likely to occur if the developmental (i.e., age) range over which the test is used is pushed too far. Thus, we expect to restrict the age range for the first version to cover adults down to an age above that where the dimensionality becomes problematic. This may seem a contradiction of the implicit design goal to measure ability at the extremes. It is a compromise in that the extremes of ability assessed by the test are not as extreme as they might be. The extremes assessable by a test such as this are, nevertheless, substantially more extreme than those possible with a conventional test. Assessment of the far extremes may have to wait for further psychometric developments, however.

Restricting the initial age range also ameliorates potential equipment problems. While a keyboard would be impractical for a 6-year old, a limited keyboard should be of no particular difficulty to a 10-year old. Similarly, although voice-given instructions might still be desirable at 10 years, a well-designed text-and-graphics instruction sequence with proctor support can do the job adequately.

The development of a complete test in line with this design is a massive undertaking. One of the challenges in developing an adaptive test that I did not discuss above is the requirement for a large number of subjects to calibrate the items (i.e., estimate their parameters). Our initial estimates called for in excess of 47,000 subject hours to develop and calibrate the entire battery. While this is a feasible requirement, it mandates a good deal of confidence in the infallibility of the design. Since our organization is relatively small, we cannot underwrite that much confidence.

Consequently, we are planning techniques to develop the overall test in parts. This will allow us to re-evaluate our design and to make corrections before we are so committed to the design that we cannot make a change. It will also permit us to use the first parts to support the development of the latter parts. We hope to have the first part finished within two years. Ultimately, we expect that the complete test will provide better measurement precision than the best of the current individually administered intelligence tests in about half the time and at a fraction of the cost.

Reference Note

1. Vale, C. D., & Prestwood, J. S. *Preliminary considerations in the development of an adaptive intelligence test*. St. Paul: Assessment Systems Corporation, 1984.

References

Buros, O. K. (Ed.). (1978). *The eighth mental measurements yearbook* (Vol. 1). Highland Park, NJ: Gryphon Press.

Chi, M. T. H. (1976). Short-term memory limitations in children: Capacity or processing deficits? *Memory & Cognition*, *4*(5), 559-572.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Merkel, S. P., & Hall, V. C. (1982). The relationship between memory for order and other cognitive tasks. *Intelligence*, *6*, 427-441.

Sternberg, R. J. (1980). Sketch of a componential subtheory of human intelligence. *Behavioral and Brain Sciences*, *3*, 573-614.

Sternberg, R. J. (1981). Testing and cognitive psychology. *American Psychologist*, *36*(10), 1181-1189.

Sternberg, R. J. (1982). Reasoning, problem solving, and intelligence. In R. J. Sternberg (Ed.), *Handbook of human intelligence*. New York: Cambridge University Press.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, *6*, 473-492.

Table 1. Tests Included in Content Review
(From Vale and Prestwood, Note 1)

Title	Publisher
California Test of Mental Maturity, 1963 Revision	CTB/McGraw-Hill
Cattell Culture Fair Intelligence Test	Institute for Personality and Ability Testing
Cognitive Abilities Test	Houghton Mifflin
Columbia Mental Maturity Scale	Harcourt Brace Jovanovich
Differential Aptitude Tests	Psychological Corporation
Full Range Picture Vocabulary Test	Psychological Test Specialists
General Aptitude Test Battery	U.S. Employment Service
Kaufman Assessment Battery for Children	American Guidance Service
McCarthy Scales of Children's Abilities	Psychological Corporation
Miller Analogies Test	Psychological Corporation
Multidimensional Aptitude Battery	Research Psychologists Press
Otis-Lennon Mental Ability Test	Harcourt Brace Jovanovich
Peabody Picture Vocabulary Test	American Guidance Service
Porteus Maze Test	Psychological Corporation
Progressive Matrices	H. K. Lewis and Company, Ltd.
School and College Ability Tests--Series II	Addison-Wesley Testing Service
SRA Primary Mental Abilities, 1962 Edition	Science Research Associates
Stanford-Binet Intelligence Scale	Houghton Mifflin
Wechsler Adult Intelligence Scale--Revised	Psychological Corporation
Wechsler Intelligence Scale for Children--Revised	Psychological Corporation

Table 2. Types of Items Included in Tests Reviewed
(After Vale and Prestwood, Note 1)

Skill (Number of Tests)	Item Type
Vocabulary (14)	Vocabulary (Synonym or Phrase Recognition)
	Picture Recognition (Point to..../What is this?)
	Synonym or Antonym Pairs
	Verbal Production
	Rhyming Words
Reasoning--Verbal (13)	Analogies
	Categories (Which is different?/Which belongs with?)
	Categories (How are these alike?)
	Verbal Production within Categories
	Sentence Completion
	Meanings of Proverbs
	Opposite Analogies
	Word Matrices
	Inference of Correct Word from Clues
	Paragraph Comprehension
	Nonsense Syllogisms
	(What should be done if...?)
	(Why do we use...?)
	(What is foolish about...?)
Reasoning--Non-Verbal (13)	Analogies using Figures or Pictures
	Arrangement of Pictures
	Categories (Which is different?/Which belongs with?)
	Categories (Figure Grouping)
	Matrix Completion
	Incomplete Pictures or Figures
	Figure Series
	Mazes
	Mechanical Comprehension
	(Insert dot to duplicate conditions.)
Reasoning--Quantitative (9)	Story Problems
	Concepts
	Number Series
	Letter Series
	Picture Marking
	Equation Building
Arithmetic (9)	Speeded Computation

Table 2. Types of Items Included in Tests Reviewed (Continued)

Skill (Number of Tests)	Item Type
Perceptual Motor (7)	Block Design Fine Motor Object Assembly Gross Motor Motor Imitation Drawing (Objects, Figures, People)
Spatial Perception (6)	Matching Figures or Pictures Left-Right Orientation North-East-South-West Orientation Box Completion Block Counting
Information (6)	General Information Recognition of Pictures of People and Places Materials
Memory (6)	Forward Digit Span Backward Digit Span Word Order Page Position of Figure Shown 4-key Xylophone Recognition of Objects Shown Repetition of Hand Movements Verbatim Repetition of Words or Sentences Story Recall Delayed Paragraph Comprehension Paired Associates Recall of Object Placement Memory for Designs
Clerical Speed (5)	Digit Symbol Names Comparison Mark-Making Matching Pairs of Letters
Perception (1)	"Magic Window" Gestalt Closure Recognition of Faces