Development of General Aptitude Test Battery (GATB) Forms E and F

Steven J. Mellon Jr., Michelle Daggett, Vince MacManus
and Brian Moritsch

Submitted To:

Division of Skills Assessment and Analysis
Office of Policy and Research
Employment and Training Administration
U.S. Department of Labor

Submitted By:

Pacific Assessment Research and Development Center
191 Lathrop Way, Suite A
Sacramento, CA 95815

1996

Addendum

Please note that the General Aptitude Test Battery (Forms E & F) referred to within this report has been renamed the Ability Profiler (Forms 1 & 2). The name of the assessment was changed to reflect: 1) the focus on reporting a profile of score results from the instrument for career exploration purposes; 2) the technical improvements made to the assessment compared to previous forms of the instrument; and 3) the capacity to use the Ability Profiler in conjunction with other instruments to promote whole person assessment for career exploration.

This material is included as Chapter 2 in the unpublished report: R. A. McCloy, T. L. Russell, L. L. Wise (Eds.), *GATB improvement project final report*. Washington, DC: U.S. Department of Labor.

Table of Contents

List of Tables

List of Figures

CHAPTER 2
DEVELOPMENT OF GATB FORMS E AND F

Steven J. Mellon, Jr., Michelle Daggett, Vince MacManus, and Brian Moritsch
Pacific Assessment Research and Development Center (PARDC)

The primary purpose of the GATB Forms E and F Development Project was to develop alternate forms of the cognitive portion of the GATB (Parts 1-7) following procedures that fulfill the highest professional standards. The project was initiated prior to the National Academy of Sciences (NAS) review. After the NAS review, the project expanded to include other objectives explicitly recommended by the NAS or otherwise implicit in its findings. Those objectives included reducing test speededness and susceptibility to coaching, investigating scoring procedures, developing items free from bias, assembling tests as parallel to each other as possible, improving the aesthetics of the tests, and revising answer sheets and other materials. The ARDP met the expanded objectives for the new GATB forms through a series of research steps. First, to deal with the speededness and coaching issues, the ARDP made changes to specifications for test length and format. Once new items were written, the ARDP conducted item reviews, an item tryout, and statistical screening of items. Based on the data, the ARDP developed new final forms. An equating study linked Forms E and F to base Form A.

In reviewing the General Aptitude Test Battery (GATB), the National Academy of Sciences (NAS) Committee identified several problems relating to test security and the speededness of the GATB tests (Hartigan & Wigdor, 1989). Stated recommendations for alleviating the problems included the following:

1. There are currently two alternate forms of the GATB operationally available and two under development. This is far too few for a nationwide testing program. Alternate forms need to be developed with the same care as the initial forms, and on a regular basis. Form-to-form equating will be necessary. This requires the attention to procedures and normative groups described in the preceding chapter.
2. Access to operational test forms must be severely limited to only those Department of Labor and Employment Service personnel involved in the testing program and to those providing technical review. Strict test access procedures must be implemented.
3. Separate but parallel forms of the GATB should be made available for counseling and guidance purposes.
4. A research and development project should be put in place to reduce the speededness of the GATB. A highly speeded test, one that no one can hope to complete, is eminently coachable. For example, scores can be improved by teaching test takers to fill in all remaining blanks in the last minute of the test period. If this characteristic of the GATB is not altered, the test will not retain its validity when given a widely recognized gatekeeping function (Hartigan & Wigdor, 1989, p. 116).

The primary purpose of the GATB Forms E and F Development Project was to develop alternate forms of the cognitive portion of the GATB (Parts 1-7) following procedures that fulfill the highest professional standards. The project was initiated prior to the NAS review and included a review of test lengths and scoring procedures. Subsequent to the NAS review, the focus of the project was expanded to include other objectives explicitly recommended by the NAS or otherwise implicit in its findings.

The expanded objectives were

- Develop new forms of the GATB that are less speeded and less susceptible to coaching by reducing the number of items and investigating the feasibility of increasing test time limits.
- Investigate and incorporate into the test the most appropriate scoring procedures, and develop instructions to examinees that clearly describe those scoring procedures.
- Develop test items free from bias, in terms of both the ethnic and gender sensitivity of the language and the statistical functioning of the items for different groups.
- Assemble test forms as parallel to each other as possible and link scores on these forms to scores from earlier forms.
- Improve the aesthetics of the test booklets and test items.
- Revise answer sheets and other related GATB materials to be consistent with changes in the test format and to provide the opportunity for examinees to maximize their test scores.

The ARDP met the expanded objectives for the new GATB forms through a series of research steps. First, to address the speededness and coaching issues, the ARDP made changes to specifications for test length and format and to supporting materials. Once new items were written, the ARDP conducted item reviews, an item tryout, and statistical screening of items. Based on the data, the ARDP developed new final forms. Finally, a study to link new Forms E and F to base Form A was undertaken. More detailed information about this research is contained in the Technical Report on the Development of GATB Forms E and F (Mellon, Daggett, MacManus, & Moritsch, 1996).

Changes to Specifications for Test Length and Format and to Supporting Materials

Concerns and Objectives

In its report, Fairness in Employment Testing, the NAS concluded that the seven paper-and-pencil tests of the GATB contained many more items than most examinees could possibly complete in the amount of time allotted for each test. Scores on items at the end of the test were more likely to be an indication of whether the examinee was coached on a rapid responding strategy than of the aptitude that the test is intended to measure. Consequently, the inclusion of items that few, if any, examinees reach if they seriously attempt to answer each question detracts from the validity of the test. Options for reducing the speededness of most of the GATB tests included increasing the time allotted and/or reducing the number of items for each test. An analysis of these options was the first step taken in revising the test specifications. Additional steps to reduce the impact of test-taking strategy, including changes to test instructions and the scoring procedures used with the tests that remain speeded, were also considered.

The NAS Committee and GATB users also expressed concerns related to the test's format, the overall aesthetic appeal of test items, and the format of the answer sheets. Actions taken to improve appearance and format were addressed in detail in the Test Aesthetics Project (Daggett, 1995), which constituted the second major step in revising the test specifications.

Reducing Speededness

The ARDP addressed issues pertaining to the GATB's speededness in three steps. First, ARDP staff analyzed score distributions for the current forms and developed initial recommendations for reducing the number of items in each test. The second step involved new research on test speededness conducted by the American Institutes for Research (AIR) under contract to DOL (Sager, Peterson, & Oppler, 1994). The final step was a review of the above work by an expert panel, who developed revised recommendations concerning both the number of items and the time to be used for each test. Each step is described briefly here, with more complete information provided in the technical report (Mellon et al., 1996) and in Sager et al. (1994).

Initial Recommendations. SARDC staff performed an initial analysis of possible test length reduction (SARDC, 1992). They used normative data in the GATB development manual and score conversion tables in the administration manuals for Forms A through D to determine the number of correct responses required to achieve a 99th percentile raw score for each test. The intent was to approximate a practical limit to the number of items used to differentiate among nearly all of the current examinees. Column four of Table 2-1 contains ranges across the four forms in the 99th percentile scores.

Table 2-1
Recommended Changes in GATB Test Order, Length, and Time Limits

Revised Test Order (Current Order)	Forms A-D		Form A-D 99% ile Ranges	Initial Proposal	Revised Proposal	
	Length	Time			Length	Time
Arithmetic Reasoning (Part 6)	25 Items	7 Min.	19-21 Items	24 Items	18 Items	20 Min.
Vocabulary (Part 4)	60 Items	6 Min.	43-52 Items	50 Items	14* Items	6* Min.
Three-Dimensional Space (Part 3)	40 Items	6 Min.	29-32 Items	35 Items	20 Items	8 Min.
Computation (Part 2)	50 Items	6 Min.	37-40 Items	40 Items	40 Items	6 Min.
Name Comparison (Part 1)	150 Items	6 Min.	70-82 Items	90 Items	90 Items	6 Min.
Tool Matching (Part 5)	49 Items	5 Min.	39-44 Items	42 Items	42 Items	5 Min.
Form Matching [Dropped] (Part 7)	60 Items	6 Min.	29-39 Items	50 Items	0 Items	0**Min.
Total, Tests 1-7 (Parts 1-7)	434 Items	42 Min.	281-295 Items	331 Items	224 Items	51 Min.

* Specifications for the final Forms E and F versions of the Vocabulary test were subsequently changed to 19 items with 8 minutes of testing time.

** Elimination of the Form Matching test also saved roughly 5 minutes in instruction and practice (sample item) time.

Because the calculations for estimating the number of items shown in column 4 of Table 2-1 do not take into account the total number of items attempted (correct plus incorrect), the test lengths suggested by these estimates could result in more than one percent of examinees completing all the items. This, in turn, could artificially constrain test scores. PARDC staff confirmed this in an examination of Form D operational data (N = 18,072) (California Test Development Field Center, 1992). For example, on the Three-Dimensional Space test, more than 19 percent of the 18,072 examinees attempted 29 or more items (the number shown in column 4). PARDC proposed new test lengths based on the total number of items attempted by fewer than one percent of the examinees. These results, combined with operational considerations, led to a small increase in the number of items originally recommended by the SARDC. Recommendations based on the items attempted are shown in column 5 of Table 2-1. A complete description of the two studies is presented in Southern Test Development Field Center (1992) and California Test Development Field Center (1992).

New Research on Test Lengths. AIR reviewed prior literature on issues of test speededness (Peterson, 1993) and designed and executed a study to provide further data on the most relevant issues found in the literature review. The review covered (1) methods of assessing test speededness, (2) relative merits of power and speeded power tests, (3) relationships between speeded and power tests of similar items, (4) differential effects of speededness, and (5) adverse psychological reactions.

A key finding from this review was that several of the constructs measured by GATB tests (particularly Arithmetic Reasoning, Vocabulary, Three-Dimensional Space, and perhaps Computation) were measured by much less speeded instruments in other programs. Key concerns identified included the extent to which speeded and unspeeded tests in these areas measured the same constructs and whether there might be greater score differences for applicant groups defined by race, ethnicity, age, and gender on the speeded versus the non-speeded tests. According to Peterson (1993), "available research provides mixed, at best, support for the expectation that reducing the speededness of the GATB power tests will reduce adverse impact or have other beneficial effects for African Americans or females, but there are some potentially beneficial, practical consequences for reducing speededness (scores would be less susceptible to changes in administration conditions, whether intended or not and adapting tests for disabled examinees is more easily accomplished)" (p. 45).

After completing the literature review, AIR staff designed a study to address the issues judged most salient—specifically, whether speeded and non-speeded versions of four of the GATB tests measured the same constructs; the extent to which examinee subgroups defined by race, ethnicity, age, or gender showed greater differences in one form or the other; and the extent to which correlations among the tests in each form were the same for different examinee groups. An additional issue addressed was whether changes in instructions, item formats, and answer sheets had a significant impact on test scores. A non-speeded test battery including Computation, Three-Dimensional Space, Vocabulary, and Arithmetic Reasoning was constructed by reducing the number of items in Form D of each of these tests and increasing the time for Arithmetic Reasoning from 7 to 11 minutes. A speeded battery comprising Computation, Three-Dimensional Space, Vocabulary, Arithmetic Reasoning, and Name Comparison was used for comparison.

Participants for the study were recruited at six state employment offices in Maryland, Tennessee, and Texas, yielding a total sample of 1,681 participants. Roughly one half of the participants (867) completed both the speeded and non-speeded batteries, using existing instructions and format. The other half (814) completed both batteries, using revised instructions and format. In both cases, the order of the speeded and non-speeded batteries was counter-balanced across test sessions. The sample included 596 African Americans, 580 Hispanics, 445 participants who were at least 40 years old, and 600 females.

In general, the nonspeeded versions of the GATB power tests measured constructs that were similar to the constructs measured by the speeded versions (i.e., all estimated true score correlations but one were greater than .80). However, the constructs were not identical (i.e., all true score correlations were less than 1.00). The difference was that the speeded versions included a speed factor. Evidence for the speed factor was provided by the higher correlations between Name Comparison and speeded versions of the GATB power tests than between Name Comparison and the unspeeded versions. Confirmatory factor analyses also provided evidence of a speed factor in the speeded versions. Further, the confirmatory factor analyses showed similar constructs across the two instructions/format conditions and across the age, gender, race, and ethnicity subgroups.

Sager et al. (1994) drew six general conclusions from this new research on GATB test speededness issues:

- Nonspeeded versions of the Three-Dimensional Space, Arithmetic Reasoning, and Vocabulary tests can be developed without large increases in the operational time limits. A nonspeeded version of the Computation test would require an increased time limit.
- For the subgroups participating in the study, nonspeeded versions of the GATB power tests would not greatly increase mean subgroup differences and could reduce mean subgroup differences between whites and African Americans on three power tests.
- Speeded and nonspeeded versions of the GATB power tests measure the same constructs as the operational versions of these tests except for a speed factor that appeared in the speeded versions. The speed factor also influences scores on the operational versions of the GATB power tests.
- The speeded and nonspeeded versions of the GATB power tests show the same relationships across the subgroups included in the study.
- The study demonstrated that speededness does not lead to differential construct validity across subgroups. There were, however, somewhat smaller mean subgroup differences between whites and African Americans on the power versions of three tests than on speeded versions of the same test. These smaller differences could not be accounted for entirely by differences in reliability between the speeded and power versions.
- The two instructions/format conditions have no effect on nonspeeded GATB power test scores and little or no effect on speeded GATB power test scores. Furthermore, instructions/format does not affect mean subgroup score differences for speeded and nonspeeded GATB power tests. Finally, the relationships between the speeded and nonspeeded GATB power tests do not change across the two instructions/format conditions.

Revised Recommendations. After initial results from the speededness research were available, DOL convened a panel of experts to review the results and develop revised recommendations on specifications for content and length of the tests to be included in the Form E and Form F batteries. The panel, which included Dr. Lloyd Humphreys, Dr. Renee Dawis, Dr. Lauress Wise, and Dr. Neal Kingston, met with ARDP representatives and AIR staff responsible for the speededness study.

The panel concluded that the current range of verbal, quantitative, spatial, and perceptual constructs should continue to be measured but that three of the tests (i.e., Vocabulary, Arithmetic Reasoning, and Three-Dimensional Space) should not be speeded. The relationship between test length and test reliability was considered in developing the recommendations for revised test lengths and times shown in the last two columns of Table 2-1. To the extent that it was not feasible to increase overall testing time significantly, the panel recommended dropping the Form Matching test to allow more time for other tests. Tool Matching was judged to be an adequate measure of the perceptual speed and accuracy construct, and this construct was judged to be somewhat less important than measures of verbal, quantitative, and spatial skills. One other recommendation—that the power tests should be administered as a group before the speeded tests—led to the proposed reordering of the remaining tests shown in Table 2-1.[1]

[1] Although Form Matching was eventually dropped, the final decision came late in the project. Consequently, it was included in many of the development steps described in this chapter. Also, Tool Matching was eventually renamed Object Matching.

Scoring Procedures

The ARDP altered scoring procedures to reduce the effects of test-taking strategies. Previous GATB forms used number-correct scoring, in which the final score is simply the total number of questions answered correctly, with no penalties for incorrect answers. Examinees who were willing to guess, even to the point of responding randomly (but rapidly) to items in the more speeded tests, were able to increase their total scores. Efforts to reduce the speededness of the power tests were designed to reduce the influence of this type of test-taking strategy.

After reviewing alternative approaches, the ARDP selected a conventional formula scoring procedure (i.e., one inflicting a penalty for each incorrect response) for use with the three remaining speeded tests. The penalty for incorrect responses is based on the number of response alternatives for each item. The concept is that, if there are k alternatives, an examinee who responds randomly will have k-1 incorrect responses for every correct response. Responding randomly does not require any knowledge of the construct being measured. The conventional formula introduces a penalty for incorrect responses that will cancel out the number of correct responses expected by chance through random responding. The general form of the formula is R-W/(k-1), where R is the number of correct (right) responses, W is the number of incorrect (wrong) responses, and k is the number of options for each item. The specific scoring formulas for the three GATB speeded tests in Forms E and F are as follows:

- Computation: R - W/4 (a reduction of 1/4th point for incorrect responses);
- Object Matching: R - W/3 (a reduction of 1/3rd point for incorrect responses);
- Name Comparison: R - W (a reduction of one point for incorrect responses).

The ARDP considered alternative approaches that included a larger penalty for guessing. Most speeded tests contain items that all examinees should answer correctly given sufficient time. More severe penalties for incorrect responses are sometimes introduced in an attempt to force examinees to take enough time with each item to answer it correctly. This approach was not recommended for use with GATB Forms E and F for two reasons. First, it introduces effects of test-taking strategy, albeit in the opposite direction. Guessing would lead to lower scores in comparison to omitting. Second, placing a greater emphasis on accuracy relative to speed changes the construct being measured to some extent, making generalizations from prior validity studies more tenuous.

The ARDP decided that use of number-correct scoring should be continued for the power tests. To the extent that examinees answer all items in a test, the number correct and formula scores are linearly related. Examinees are ordered in exactly the same way in both cases. The number-correct score is simpler to explain to examinees. Instructions to examinees on how to maximize their scores—by attempting to answer every item—are also simpler when using number-correct scoring.

Instructions to Examinees

For GATB Forms A through D, both the general instructions and the test-specific instructions provide limited information regarding test-taking strategies, and neither discusses scoring procedures. Test standards developed by the AERA/APA/NCME joint committee on test standards (1985) require that examinees be told how tests will be scored and given specific instructions that allow them to maximize their scores. For this reason, changes to test scoring procedures were accompanied by consideration and revision of examinee instructions.

Item Pretest. During the item pretest, both general and test-specific instructions were modified to improve the information provided to examinees. The relevant portion of the general instructions used during the pretest stated

> *You probably will not be able to finish all the questions in the first three parts. Each part has so many questions that very few people can finish in the time allowed. However, answer as many as you can.*

For each of the speeded tests, the instructions given after completion of the practice items included the following:

> *Work as FAST and as CAREFULLY as you can. On this exercise SPEED is very important. If you have some idea of the answer to a question, even if you are not absolutely positive, it is to your advantage to take your BEST GUESS. For example, if you can eliminate one or more of the choices to a question, take your BEST GUESS. However, if you have no idea what the correct answer is, don't spend time guessing. Move on to the next question.*

For the power tests, examinees were simply instructed to "work as ACCURATELY and FAST as you can."

Test Tryout. During the test tryout, instructions were modified to provide information on scoring procedures as well as advice on test taking strategy. At this time, the tests were reordered so that all of the power tests were administered first, followed by the three speeded tests. Separate sets of general instructions were provided for the power and speeded tests. The general instructions for the power tests (Parts 1, 2, and 3) stated

> *On the next three parts work CAREFULLY. You should have enough time to answer each question. It is to your advantage to ANSWER EVERY QUESTION. Even if you're not sure of an answer, make your BEST GUESS, fill in your answer, then go to the next question. Your score for each part will be the number of questions you answer correctly. There is no penalty for answering incorrectly.*

This information was repeated in the specific instructions following the practice items for each of the power tests.

After the power tests were completed, general instructions for the speeded tests were provided. These instructions stated:

> *The next three parts are different from the parts you've already taken. On these parts, SPEED is VERY IMPORTANT. You won't have time to answer every question. You must work as FAST as you can but don't be careless.*
>
> *If you have even the slightest idea of the answer, it is to your advantage to make your BEST GUESS. If you can eliminate one or more wrong choices to the question, then make your BEST GUESS*

> *from the remaining choices. However, if you have no idea of the correct answer, don't spend time guessing; go to the next question.*
>
> *You will receive one point for each correct answer. You'll be penalized for wrong answers. Points will not be subtracted for questions you don't answer.*

This information was also repeated in the specific instructions following the practice items for each of the speeded tests. At that point, examinees were told the specific penalty for incorrect responses on that test. For the Computation test, for example, examinees were told:

> *You will receive one point for each correct answer. You'll lose one quarter (1/4) of a point for each wrong answer. Points will not be subtracted for problems you don't answer.*

Operational Use. DOL determined that the instructions used in the test tryout should also be used operationally with Forms E and F. No significant problems with these instructions were discovered in the test tryout. Further, changes in test instructions might jeopardize the generalizability of linking results from the test tryout study.

Research on Test Aesthetics

The goal of the Test Aesthetics Project was to improve the physical appearance and user friendliness of the GATB and supporting administration materials. The procedures used and resulting improvements are summarized briefly here; a more complete description of the project is presented in Daggett (1995).

The project involved three major activities:

1. interviews with testing professionals to identify specific areas to be addressed; literature searches of recent aptitude tests, personality inventories, vocational interest inventories, and supporting administration materials to define specific aspects of these areas;

2. focus groups with military testing professionals, employment counseling professionals, and representatives from the five Assessment Research and Development Centers and the National Office to discuss user issues, editorial styles, and recommended practices;
3. a survey of organizations that contract with the state of California to administer the GATB.

The project resulted in a number of recommendations for format revision which were, in turn, incorporated into the revisions of the GATB test booklets, answer sheets, and administration manual. Major format revisions are presented below for each document.

Test Booklets and Instructions. Revisions made to the test booklets and accompanying examinee instructions include increasing white space, changing the type font to contemporary 12-point Palatino, and increasing the use of italics, bolded print, and underlining to emphasize specific words or phrases. In addition, cueing graphics were added to the bottom of pages when appropriate and higher grade paper and print quality were used.

Revisions for the final versions of the test booklets included (a) combining the two booklets used in previous GATB forms into a single booklet, (b) revising the instructions to reduce redundancy and the number of words in each sentence, and (c) adjusting space provided for the

Computation and Arithmetic Reasoning items to conform to the amount of space needed by each item.

Test Items. In general, a number of changes were made to improve the appearance and user friendliness of the test items themselves. Specific changes for the individual items are discussed below for each test.

Arithmetic Reasoning – Individual items were placed in cells containing one double vertical line (i.e., two vertical lines adjacent to each other) and one single horizontal line. Although the two-column format was maintained, the number of items on each page was reduced. Only Arabic numerals (instead of words) were used to express numbers. Except for monetary values, a zero was used as the first digit for decimal values less than one.

Computation – The same item format developed for the Arithmetic Reasoning items was used for the Computation items. Also, there were no more than eight Computation items on each page. Consistency in punctuation, response alignment, and use of monetary symbols was maintained throughout the test. Finally, the arithmetic operation symbol for each item was placed within the item. (In prior forms, the operation symbol was placed above the item.)

Form Matching – The second item block was reduced from 35 to 25 items, and the number of response options was reduced from 10 to 5.

Name Comparison – The number of items on each page was reduced from 50 to 30. The horizontal line after every fifth item was replaced by a blank line, and a blank space preceded and followed the dash that separated the two names within each item.

Object (Tool) Matching – The print quality was improved, and the test title was changed from Tool Matching to Object Matching.

Three-Dimensional Space – The print quality and resolution were improved through the use of the CorelDRAW! 4 graphics package (Corel Corporation, 1993) to develop the individual items.

Vocabulary – The item format was changed from horizontal to vertical. The number of items on each page was reduced from to 30 to no more than 10. The 19 items were arranged in three columns of five items and one four-item column, each separated by a double vertical line and one horizontal line.

Answer Sheet. Several changes were made to the answer sheets used in the Forms E and F development research. In general, response formats were revised to conform to those made in the test items. Response bubbles were reduced in size by eliminating the top and bottom portions of each bubble and adding small horizontal lines to form ovals. This change was made in response to a concern raised in the NAS review that difficulty in filling in large circles might impede performance on the speeded tests. Specific changes incorporated at each stage in the development research are described below.

Item Pretest Phase. Four scannable answer sheet formats were used during this phase for data collection. The answer sheets were modified to: (1) contain four sections; (2) support the experimental test configuration; (3) accommodate format changes made to Form Matching and Vocabulary tests; (4) present all demographic information on one page; and (5) use more contemporary and legally appropriate wording. In addition, each answer sheet was further adapted to be used specifically with one of the four power tests. The four answer sheets and 16 test booklets were color-coded to facilitate the appropriate test booklet-answer sheet combination.

Test Tryout Phase. The test tryout answer sheet was used with both Forms E and F. It was printed in purple to differentiate it visually from Form A and Form B answer sheets. The number of sections was increased from four to seven, and they were rearranged to coincide with the new test arrangement (Part 7-Form Matching was later eliminated). The shaded practice boxes were placed at the top left of each section. The correct number of item responses was placed in each section and reformatted into columns of equal numbers with the tops of each column starting on the same line. With the elimination of the wraparound columns, the phrase “Begin Here” became unnecessary and was eliminated. The demographic section was modified to collect research-specific information. The most notable modification, however, was replacing the large response circles with ovals.

Operational Phase. The operational answer sheet differs from the test tryout answer sheet in the following ways:

- Research-specific demographic information was eliminated.

- The Form Matching section (Part 7) was eliminated.
- The response identification letter was placed inside each response oval.
- Alternate blocks of five item responses were shaded.

Administration Manual. The specific revisions made to date include

- reformatting as a technical manual;
- incorporating a conversational business tone (i.e., employing simple, concrete, clear language);
- writing in a manner to address reader’s needs first;
- consolidating procedures and instructions into distinct subject areas (i.e., eliminating unnecessary language, inconsistencies, repetition, and redundancy);
- subtitling each subject area;
- updating and rewriting information so that the reading level and detail are appropriate for all users;
- formatting manual into a logical sequence for better understanding of administration procedures and instructions;
- adding cueing graphics and bullets;
- using colored paper as a cueing technique;
- increasing white space and reducing line length;
- changing to 12-point serif typeface with increased use of headers;
- using higher grade paper with improved print quality;
- developing a standardized introduction script; and
- binding the manual into an 8 1/2 in. by 11 in. hardback binder (because of ongoing modifications).

The following changes have been proposed for the final version of the administration manual:

- different colors of print to serve as cueing devices for the test administrator (e.g., phrases to be read aloud) at different locations in the manual;

- reducing the size of the pages to 7 in. by 10 in. with wire spiral binding; and
- including demonstration models for the Three-Dimensional Space practice items.

The final administration manual for GATB Forms E and F will include scoring procedures and conversion tables. The use of administration aids such as color, checklists, and forms will be increased. Additional aesthetic and format changes will be made to further enhance readability and usability, such as reducing the size of the manual, tabbing the sections, adding more cueing graphics, and incorporating a two-column format with shorter lines.

Development and Review of New Items

An extensive effort was undertaken to develop new items for GATB Forms E and F. Note that many more items were developed than used in the final forms. Efforts to specify content and difficulty categories, write items for each of these categories, and then review items for editorial and sensitivity considerations are described here. Item tryouts, statistical screening, and calibration procedures are described in the following section.

Item Writing

A brief description of the development of experimental items for Parts 1-7 of Forms E and F is given below. For each test, items from previous forms were analyzed and sorted into categories potentially related to item difficulty. Sources for item material also were identified. More detailed information on specifications and item types/content categories for each test are presented in the Forms E and F project technical report (Mellon et al., 1996). Brief summaries of the item development procedures used for each test are presented here.

Name Comparison. The 400 Name Comparison items were developed to be parallel to Form A items and representative in terms of gender and ethnicity. The number of items with names that were the same was equal to the number of items with different names. Item sources included directories, dictionaries, and item developer creativity. Analyses were then performed to develop preliminary estimates of item difficulty. Based on these analyses, the number of characters in the left-hand column of the two-column format used for this test was selected as the item difficulty measure. The 200 items for each form were divided into four 50-item quarters of approximately equal estimated overall difficulty. The item order was then randomized within each quarter.

Computation. The 136 Computation items were developed to be parallel to Forms A-D. The original items were developed and reviewed to evaluate difficulty. The number of digits across numbers within each type of operation was used as the item difficulty measure. The 68 items for each form were divided into four 17-item quarters of equal estimated overall difficulty. Type of arithmetic operation and response options were balanced within each quarter. A low-difficulty item was assigned to the first position within each quarter with the remaining items ordered randomly.

Three-Dimensional Space. The 130 Three-Dimensional Space items were developed to be similar in content to prior forms. The number of folds was used as a measure of item difficulty; it had six levels. Newly developed items were grouped according to the number of folds so that an equal number of items would be developed for each of the six difficulty levels. Items were then drawn on a computer, using the CADD-3 software package. Items were continually reviewed for clarity and correctness, and shading was added. Completed items were transferred to Mylar paper and reduced in size photographically, then plates were made for printing. Items were reviewed again and revised when necessary. Items were then assigned to forms on the basis of difficulty, and response options were checked and tallied. Option positions were changed as necessary. The items were rephotographed and printed.

The ARDP used the CorelDRAW! 4 graphics package (Corel Corporation, 1993) to redraw all of the items to make them consistent in appearance. Camera-ready copies of the reformatted items were prepared and sent to a graphic artist for proofing. Some of the items were later revised to correct the problems identified by the graphic artist. Three difficulty levels were identified based on the number of folds and/or rolls made in each item. These difficulty values were then used to form three 16-item quartiles and one 17-item quartile of approximately equal estimated overall difficulty within each form. Within each quartile a low-difficulty item was assigned to the first position with the order of the remaining items randomized. The correct response option frequencies were balanced within each quartile.

Vocabulary. The 160 Vocabulary items were developed to be parallel to Form B. Item review also focused on word difficulty but used a different approach from previous GATB development efforts. Specifically, The Living Word Vocabulary (Dale & O'Rourke, 1981) provided estimates of item difficulty. This reference assigns a grade level to each word meaning. The assigned grade level is based on the responses of students who completed vocabulary tests during the period of 1954-1979. When multiple word meanings were reported for a given word, the average grade level was used. Higher grade levels indicated greater difficulty. The mean of the reported grade levels for the four words that made up each item was used to estimate item difficulty. Four difficulty level categories were formed. These categories were used to prepare four 20-item quartiles of equal estimated overall difficulty for each form. For each quartile, the two items with the lowest estimated difficulty appeared in the first two positions with the order of the 18 remaining items randomized. The correct response option frequency distributions were balanced within quartiles and forms.

Object Matching. The 163 original Object Matching items were developed to be parallel to Forms A-D. The ARDP used the number of shaded areas in the four response alternatives for each item to estimate difficulty level. Difficulty level, content considerations, and location of the correct response were used to form four 20-item quartiles of similar overall difficulty for each form. (Three items were deleted.) The item order was randomized within each quartile. A surplus item was then added to each quartile to form three seven-item pages that could be shifted to meet the requirements of the research design.

Arithmetic Reasoning. The 66 Arithmetic Reasoning items were developed to be parallel to Form A. New situations, contemporary monetary values, gender representation, exclusion of extraneous information, and a sixth-grade reading level were additional considerations in item development. The ARDP reviewed and revised the items so they conformed more closely to the guidelines for development. Item difficulty was estimated by the number of operations needed to solve the problem, the type(s) of operations, and the number of digits included in the terms used in the operation(s). One of the two least difficult items was assigned to the first item position in Form E and the other item assigned to Form F. The remaining 64 items were then assigned to four eight-item quartiles for each form on the basis of difficulty, type (s) of operation(s), correct response key, and content. The items in each quartile were ordered from least to most difficult with the item order then randomized within each quartile.

Form Matching. The 200 Form Matching items were developed to be parallel to Forms A-D items in terms of content and parallel to Form A item size and arrangement. Eight 25-item blocks were developed by modifying each of the eight blocks of items in Forms A-D. The number of response options for each item was reduced from 10 to five.

Editorial Review and Screening

Literature Review. The development of the item review procedures began with a literature review focusing on the process for conducting item reviews and selecting the participants in the review process. Two types of review procedures were identified: (1) procedures used for research purposes, and (2) procedures used in ongoing testing programs.

Item Review Instruments. Through the literature review, it was determined that (1) most previous item review procedures were designed for use with educational achievement tests, and (2) review procedures used in most previous studies were not highly structured and appeared to be developed independently with limited guidance from the educational measurement literature. However, the literature review did uncover eight references (Boldt, 1983; Hambleton & Rogers, 1988; Harms, 1978; Lockheed-Katz, 1974; Madaus, Airasian, Hambleton, Consalvo, & Orlandi, 1979; Olson & Smoyer, 1988; Schratz & Wellens, 1981; Tittle, 1982) that provided the foundation for the instruments and procedures that were used in the GATB Forms E and F item review. These eight references provided information in three areas: bias guidelines, procedural issues, and rating questions.

Preliminary Review. Draft versions of item sensitivity review questions, instructions, and an answer form were sent to ARD centers for review. Based on the comments, ARDP staff revised draft versions of the sensitivity review materials and sent them to ARDCs for further review. The only revision was a minor change in the answer form.

Pilot Test. A pilot test was conducted in-house with three Cooperative Personnel Services (CPS) staff members, enabling individuals who were not involved in the ARDP test research program to provide input to the review process. The results led to a number of modifications in procedures, instructions, and documents that would be used for the item review.

Item Review Materials. Nine documents were used in the item review process: (1) a list of the criteria to select panel members, (2) a confidentiality agreement, (3) a description of the GATB tests and aptitudes, (4) written instructions for panel members, (5) the administrator's version of the written instructions for panel members, (6) a list of characteristics of unbiased test items, (7) a list of the review questions with explanations, (8) an answer form, and (9) an answer form supplement.

Panel Member Characteristics. Seven panel members participated in the review. The panel included two African Americans, three Hispanics, and two whites. Three members were male and four female. Three members were personnel analysts, two were university professors in counselor education, one was a personnel consultant, and one was a postdoctoral fellow in economics.

Procedures. At an orientation meeting held at each of the three participating ARDCs, confidentiality agreements were signed, GATB items and instructions were given to panel members, and several items in each test were reviewed and discussed. Panel members reviewed the remaining items at their convenience. After all items were reviewed, a follow-up meeting was held at each center to resolve any problems and to discuss the review process.

Summary of Results. The answer forms of each panel member were reviewed. Summaries of the comments for each test are presented below.

> *Name Comparison* – Comments focused on racial, ethnic, and gender stereotyping and representation. Specific concerns included the lack of female and minority businesses, and the need for more females in nontraditional professions, jobs, and businesses.
>
> *Computation* – Comments primarily dealt with item characteristics. Specific concerns included difficult and time-consuming problems that might be skipped by testwise applicants, poor distractors, and unclear instructions.
>
> *Three-Dimensional Space* – Comments concerned possible gender bias and item characteristics. Comments included the presence of male-oriented items and abstract items

that might be unfamiliar to females, difficult and time-consuming items that could be skipped by testwise applicants, gender-biased instructions, and overly complicated items.

Vocabulary – Comments concerned high reading grade level, overly difficult words; words with different meanings for different groups; and inclusion of foreign-language words and technical, biological, and scientific terms.

Tool Matching – Comments focused mainly on possible gender bias due to differences in familiarity and the presence of male-oriented items. However, concerns were also expressed that items with electrical and mechanical components might cause problems for minorities due to lack of familiarity and opportunity to learn. Other comments concerned clarity of instructions and positioning of the response letters for the item alternatives.

Arithmetic Reasoning – Most comments were directed toward two areas: (1) racial, ethnic, and gender representation, and (2) gender occupational and activity stereotyping. Other comments concerned time-consuming items that might be skipped by testwise applicants, confusing and incomplete instructions, the presence of items that were overly complicated or involved too many steps, and some groups not having the opportunity to learn how to perform the operations needed to answer the complex items.

Form Matching – Comments included a possible practice effect for the test and unclear instructions because of reading level. Comments that were directed toward specific items included linear illustrations being perceived as “hostile,” minute differences among shapes, and possible confusion due to shape similarity and location.

Item Content Revision. Based on results from the panel evaluation, the content of specific items was revised. A summary of the types of changes introduced for each test is presented here.

Name Comparison – The revisions addressed the racial, ethnic, and gender stereotyping and representation criticisms. Guidelines based on the 1990 U.S. Census were used to increase racial/ethnic and gender representation. Stereotyping was addressed by including items with minorities and females in nontraditional occupations and businesses; more professional occupations and businesses were included. Fewer items with Germanic names were used. Format changes included separating the items into blocks of five, eliminating horizontal lines, and increasing the horizontal and vertical space within and between items. Finally, the instructions were reworded to increase clarity; bold and italicized types were used for emphasis.

Computation – Distractors were revised to make them more plausible based on five error types. Minor format changes included adding commas to numbers with at least four digits and placing the operation sign within the item. Finally, the instructions were reworded slightly to increase clarity, and bold and italicized types were used for emphasis.

Three-Dimensional Space – Individual items were revised when needed to increase clarity. Revisions were reviewed by a graphics expert familiar with the test format and the drawing software to ensure that the items were free of errors. Instructions were reworded slightly to increase clarity and eliminate possible gender bias; bold and italicized types were used for emphasis.

Vocabulary – Words were replaced on the basis of the item review panel member comments and on an analysis of word difficulty in Dale and O’Rourke (1981). Items were modified as needed to ensure that each item’s level of word difficulty was appropriate, word forms within items were identical, and the same type of correct response (i.e., synonym or antonym) was maintained within each item. The item format was changed from horizontal to vertical ordering of words. Finally, the instructions were reworked (e.g., bold and italicized types were used to emphasize important points, and a statement was added stressing that all choices should be considered before selecting an answer).

Tool Matching – Item revisions included eliminating inconsequential differences among item responses, eliminating duplicate responses, and refining responses (e.g., removing extraneous matter, drawing sharper lines, eliminating broken lines). Finally, the instructions were reworded slightly to increase clarity; bold and italicized type were used for emphasis, and the test name was changed from Tool Matching to Object Matching. (Future forms will include more generic items even though the results from item analyses indicated that female scores are slightly higher than male scores on the current items.)

Arithmetic Reasoning – Revisions involved four areas: making minor item format modifications, eliminating gender stereotyping, making the distractors more plausible, and increasing racial, ethnic, and gender representation. The instructions were reworded slightly to increase clarity; bold and italicized types were used for emphasis.

Form Matching – Changes included enlarging figures to increase clarity, repositioning items to equalize space among items in the lower blocks, and revising an item family to make it less similar to another item family. The number of response options was reduced from 10 to 5. Finally the instructions were reworded slightly to increase clarity; bold and italicized types were used for emphasis.

Item Tryout and Statistical Screening

The ARDP conducted a tryout once new items were written and screened, administering the new items to a sample of examinees. The statistical information gathered served two purposes: item screening and item calibration. Items were dropped from further consideration (screened out) if they were too easy or too difficult, if they failed to discriminate between higher and lower ability examinees, or if they showed significant differential functioning by gender or race or ethnic group. For the remaining items, the tryout data were used to estimate item statistics for use in constructing parallel forms. The rest of this section summarizes the results of the item pretest data collection, item analysis, and item selection.

Item Tryout Booklet Design

There were many more items than any one examinee could reasonably be expected to complete, so the ARDP organized the items into separate booklets and assigned different booklets to different examinees. The design of the tryout booklet addressed a number of issues. First, to allow selection of the best items, the ARDP needed to try out more items for each test than would eventually be used operationally. Further, time-per-item was increased to improve the quality and quantity of the item data (e.g., fewer omits and unreached items). Together, these considerations meant that it was not reasonable for each examinee to complete all of the tests. Instead, the tryout booklets were designed so that examinees would complete some of the items for each of the three speeded tests and all of the items for one of the four power tests (i.e., Arithmetic Reasoning, Three-Dimensional Space, Vocabulary, and Computation[2]).

For the speeded tests, the ARDP grouped new items within each test into four sets and rotated the order of the four sets across four different booklet pairs for each form. Not all of the

[2] The Computation test was originally included with the power tests but was subsequently treated as a speeded test (cf. Sager et al., 1994).

items were printed in each booklet. Specifically, for Form Matching, two sets of 25 items each were printed in each booklet; for Tool (Object) Matching, two sets of 20 items each were printed in each booklet; and for Name Comparison, three sets of 50 items each were printed in each booklet. The intention was that the items within each block would be analyzed for the booklet where the block appeared in the first position. In this way, enough examinees would complete (reach) the item to allow assessment of item difficulty, item-total correlation, and subgroup differences in item performance.

For the speeded tests, there was no attempt to equate across booklets at the item level. Item results were expected to vary widely according to each item's position in the test, so equating would be handled at the test level in the form calibration study. For the power tests, an attempt was made to equate item difficulties and IRT parameter estimates not only between the Forms E and F item sets, but also with the item parameter estimates from an operational form, Form A. Consequently, all items from the Form A version of a given power test were included along with half of the new items for that test (i.e., all items for either Form E or F) in Part 4 of a tryout booklet.

The ARDP determined the order of items within each power test by dividing Form A (anchor) items into discrete blocks and then spacing these blocks throughout the tryout booklet. The remaining item positions were filled with new items. For power tests, an item's position within a form should not affect its difficulty (or discrimination). The tryout booklet design included provision for testing this assumption. Specifically, two versions of each power test were created with the order of the new items reversed in the even-numbered tryout booklets relative to their positions in the odd-numbered booklets. The (Form A) anchor items were printed in the same position in each of these booklet pairs. Table 2-2 summarizes the design of the 16 booklets developed for the item tryout study.

Table 2-2
Item Tryout Booklet Design

Tryout Booklet (Booklet pairs are identical except for reversed order of Part 4 new items)		Part 1 Form Match 50 items, 10 min.	Part 2 Tool Match 49 items, 6 min.	Part 3 Name Comp. 150 items, 11 Min	Part 4 One of the Power Tests (Form A anchor items were grouped in fixed blocks; new items filled remaining positions.)			
Form E	Form F	Item #s	Item #s	Item #s	Test	Anchor	New	Time
1,2	9,10	1-50	1-49	1-150	Computation	50 Items	68 Items	67 min.
3,4	11,12	26-75	21-69	51-200	3D Space	40 Items	65 Items	50 min.
5,6	13,14	51-100	41-80, 1-9	101-200, 1-50	Vocab	60 Items	80 Items	70 min.
7,8	15,16	76-100, 1-25	61-80, 1-29	151-200, 1-100	Arith. Reasoning	25 Items	33 Items	73 min.

Item Tryout Sample

The primary target sample was Employment Service local office applicants. Local offices used for data collection were representative of offices serving the working population in terms of gender, ethnicity, age, and educational level. Supplemental sources for study participants were also identified. These sources included employed workers, community groups or associations, high school seniors and junior college students, and vocational training centers. The sample members were not to have taken any form of the GATB within the 12-month period immediately prior to testing. Study participants were reimbursed for travel expenses.

Sample size targets were set to support accurate estimation of item statistics and analyses of differential item functioning (DIF) across gender and race/ethnic groups. An overall target of 1,000 examinees per item was set to support item response theory (IRT) and classical item analyses. Within the overall target, a roughly equal split of males and females, and a minimum of 200 members from each of the three race/ethnic groups to be compared (Whites, African Americans, and Hispanics) were desired. Sampling at each site was conducted to assure an equal gender split within each race/gender group insofar as possible. These target sample sizes applied to pairs of booklets that differed only in the ordering of the Part 4 (power test) items. Except for analyses of item position effects in Part 4, the data for each pair of booklets were pooled in the item analyses. A target sample size of 500 examinees per booklet was set to achieve samples of 1,000 for each pair of booklets.

The subjects for the Item Pretest phase were applicants of Employment Service local offices in the five ARDP geographic regions. Each ARDC tested approximately the same number of examinees, and to the extent allowed by regional demographic characteristics, each center obtained the same minimum subsample sizes for the required ethnic and gender groups as specified in the research design. The total number of individuals tested for this study was 9,237. The sample was approximately 45 percent female and 55 percent male. Sample ethnic composition in percentages was as follows: African American, 40 percent; Asian, 3 percent; Hispanic, 18 percent; white, 35 percent; Native American, 2 percent; and subjects choosing the “other” category, 2 percent. The mean age and education of the total sample were 34.91 and 12.62 years, respectively. The standard deviation (years) was 12.27 for age and 2.72 for education. Table 2-3 shows the sample composition by booklet, gender, and race for subjects with usable data on the Form Matching test (Part 1).

Data Collection Procedures

Each subject completed 1 of 16 item tryout booklets. Time limits were set to enable the subjects to complete all items in one power test, and 25 percent of the items in one form of the three speeded tests. Although the time limits for the speeded tests were increased slightly above operational time limits, the speeded nature of these tests was preserved. One half of the subjects completed Form E, and the other half completed Form F. All subjects completed one of the three speeded tests (Parts 1, 5, and 7), one of the four power tests (Parts 2, 3, 4, or 6), and an anchor test (GATB Form A) for the power test. The overall time limit ranged from 77 to 100 minutes for each of the eight test booklets for each test form as shown in Table 2-2.

Four answer sheets were prepared to conform to the four tests within each booklet. All data were recorded on an optical scanner and read to diskettes. The answer sheets and diskettes were submitted to PARDC for analysis. PARDC also prepared data collection and submission instructions.

Table 2-3
Item Tryout Sample Size for Each Test Booklet

Booklet	Total	Male	Fem.	White	Black	Hisp.
1+2	1,233	666	567	548	393	211
1	574	299	275	244	185	110
2	659	367	292	304	208	101
3+4	1,169	644	525	386	473	216
3	604	319	285	194	241	125
4	565	325	240	192	232	91
5+6	1,201	668	533	319	589	200
5	555	312	243	145	258	111
6	646	356	290	174	331	89
7+8	1,193	661	532	479	446	204
7	604	339	265	244	218	107
8	589	322	267	235	228	97
9+10	1,102	605	497	403	403	207
9	577	326	251	223	195	112
10	525	279	246	180	208	95
11+12	1,112	623	489	375	457	207
11	558	300	258	197	219	104
12	554	323	231	178	238	103
13+14	1,016	576	440	277	464	196
13	553	314	239	156	244	107
14	463	262	201	121	220	89
15+16	1,140	602	538	413	435	214
15	561	302	259	211	205	104
16	579	300	279	202	230	110

Test Administrator Training. Training for collecting item pretest data was accomplished in two stages. First, ARDP staff conducted a two-day train-the-trainer session to provide project lead staff from each of the five ARDCs with a comprehensive overview of the project and detailed instructions for training staff who would collect the data. The lead staff were then responsible for conducting similar training sessions in their respective geographical regions. Training was divided into eight modules:

Module 1. An Overview of the E and F Development Project

Module 2. E and F Pretest Data Collection Overview

Module 3. Administration Overview

Module 4. Administering the GATB

Module 5. Before, During, and After the Testing Session

Module 6. Specific Instructions for GATB Tests

Module 7. Administering the GATB - A Practical Exercise

Module 8. Checking in with the Data Collection Coordinator.

Materials needed for completing the training modules included: *Trainer's Guide, GATB Item Pretest Administration Manual*, sample test booklets, and sample answer sheets. The Administration Manual was designed to contain all information needed by data collection staff for completing training and actual data collection. Copies of the *Trainer's Guide and the GATB Item Pretest Administration Manual* are included in Appendix E of the *Technical Report on the Development of GATB Forms E and F* (Mellon et al., 1996).

Sixteen test booklets and four separate scannable answer sheet formats were required to collect data for the 16 test versions. Instructions for matching each of the four answer sheets to the appropriate test booklet were provided in the GATB Item Pretest Administration Manual. The Administration Manual instructions also included a checklist of all supplies and materials needed for data collection. PARDC provided all materials except scrap paper, pencils, and stop watches to each ARDC. NARDC developed and distributed scanning software and instructions.

Data were collected sequentially for each of the 16 test booklets, beginning with Booklet No. 1. Where possible, within each region, targeted sample sizes for the six subgroups (African American female, African American male, Hispanic male, Hispanic female, white female, and white male) and the total sample size for each booklet were obtained before beginning data collection on each succeeding booklet. This was done to allow for preliminary reviews of data before the entire data collection effort was completed. Specific instructions for collecting, recording, and submitting data were provided in the Administration Manual.

Data Analysis

Analyses of the power and speeded test items proceeded somewhat independently due to differences in the nature of the data to be analyzed and the information requirements for screening and form construction. The analysis of items for GATB power tests was done under separate contracts with Measured Progress, HRStrategies, and Young and Associates. Winnie Young performed preliminary edits and produced classical item analysis statistics. Dr. Neal Kingston of Measured Progress evaluated the dimensionality of the power tests and performed computer analyses to estimate Item Response Theory (IRT) parameters, in addition to conducting a preliminary selection of items. HRStrategies used this information, in conjunction with their own analyses for assessing differential item functioning (DIF), to select the final items for the new forms (HRStrategies, 1994). Dr. Fritz Drasgow provided technical advice to HRStrategies staff. The Human Resources Research Organization (HumRRO) analyzed speeded test data and selected items for the four speeded tests (McCloy, Russell, Brown, DiFazio, & Green, 1994). Dr. Bert Green provided technical advice to HumRRO staff. Note that the Computation test items were included in both sets of analyses. As noted earlier, this test was originally included with the power tests but subsequently treated as a speeded test.

The procedures used to estimate item statistics for (i.e., calibrate) the power test items and to screen them for fairness to different examinee groups are described in the next section. This is followed by a description of the procedures used with the speeded test items and a discussion of how item statistics were used to select items for the final version of each Form E and Form F test.

Calibration and Screening of Power Test Items

After the raw data were edited, analysis of the tryout data for the power test items began with attention to two methodological issues. The first issue was whether there were significant differences in apparent item difficulty (and to a lesser degree, item discrimination) as a function of the item's position in the tryout booklet. Recall that each new item was tried out in two booklets, with the ordering of the new items in the even-numbered booklets being reversed from their ordering in the next lower odd-numbered booklet. The question was whether it was reasonable to combine the data from the two booklets into a single analysis of item statistics. The second issue was whether all of the items for a test measured a single underlying construct. Because IRT models (which are commonly used in item screening and in selecting items for inclusion in a test) assume unidimensionality, it was important to check this assumption before proceeding.

Once these two issues were addressed, PARDC, Measured Progress, and HRStrategies estimated item statistics and analyzed DIF across ethnic and gender groups. At the conclusion of these steps, a few items had been dropped because of DIF; the remainder of the items were calibrated and ready for selection into the final forms. Each of these steps is described briefly here and discussed in more detail in Mellon et al. (1996).

Item Position Effects. The effect of an item's position on its apparent difficulty was assessed by comparing difficulty estimates from the forward and reversed ordering of the new items for each test. Table 2-4 summarizes the distribution of differences in item difficulty (proportion correct) across the two booklets for each test. Although the results showed a significant correlation between item position and proportion passing, the size of the differences was generally modest. The Computation test showed the largest item position effects, which was consistent with concerns that the test was partially speeded. Computation was subsequently treated as a speeded test. Measured Progress and PARDC decided that combining data from the two item orderings was the best approach to minimize the effects of item position. The pooled data were used in the remaining analyses.

Test Dimensionality. Dr. Neal Kingston of Measured Progress analyzed the dimensionality of each power test using the TESTFACT program (Wilson, Wood, & Gibbons, 1991). Application of ordinary linear factor analysis procedures to dichotomously scored item variables typically yields extraneous difficulty factors. TESTFACT largely corrects this problem by incorporating a logistic model of the relationship between the underlying factors and the item responses.

Table 2-4
Effect of Changing Item Position on Item Difficulty

	Test			
Statistic	Arithmetic Reasoning	Computation	3-D Space	Vocabulary
Minimum Difference	-.14	-.18	-.21	-.15
1st Quartile Difference	-.04	-.04	-.04	-.03
Median Difference	-.01	.01	.00	-.01
3rd Quartile Difference	.05	.07	.03	.03
Maximum Difference	.14	.23	.17	.37
r_{pd}	-.62	-.76	-.31	-.56

Notes. The .37 maximum difference in difficulties was found for Vocabulary item 1 (in the forward order), which was an extreme outlier. The next largest difference for Vocabulary was .11.
r_{pd} is the correlation between initial item position and change in difficulty.

Form A items for each test were analyzed first. Results showed a strong first factor for each test. The Computation test had the strongest and most interpretable second factor, which contrasted division and other items. Because Computation was subsequently treated as a speeded test, IRT analyses were not used with this test. For the remaining three tests, the first factor accounted for roughly half of the total variance and the remaining factors were small and largely uninterpretable. One of the 3-D Space factors appeared to be related to difficulty, but this was most likely an artificial result not fully eliminated by the use of TESTFACT. Consequently, the Arithmetic Reasoning, 3-D Space, and Vocabulary tests were judged sufficiently unidimensional to support IRT analyses.

Item Calibration. Table 2-5 shows the number of examinees used in the analysis of each of the power tests after editing and pooling across booklet pairs. Measured Progress used the BILOG program (Scientific Software, 1990) to estimate both classical item parameters (proportion passing and item-total biserial correlation) and IRT parameters. Twelve items were found to have biserial correlations less then .15 and were eliminated from the IRT analyses. Parameter estimates (difficulty, slope, and guessing) for the three-parameter logistic IRT model were obtained for the remaining items. Item fit (including item plots with fit probabilities less than 0.3) was examined, but no other items were eliminated on the basis of item statistics. In general, the strategy was not to screen out items altogether solely on the basis of their statistics, but rather to avoid selecting items for the final forms that were too easy or difficult or had marginal discrimination unless no better items could be found.

Table 2-5
A Description of the Samples by Ethnic and Gender Subgroups

Test	Whites	African Americans	Hispanics	Males	Females
Arithmetic Reasoning	857	700	364	1,102	946
Vocabulary	583	917	362	1,125	900
Three-Dimensional Space	699	741	389	1,105	879

Differential Item Functioning (DIF). DIF refers to a situation where the probability of passing an item differs for individuals at a given level of true ability who differ only on the basis of group membership (typically defined by ethnic or gender distinctions). HRStrategies used two approaches to analyzing the item tryout data for possible DIF. Separate analyses were conducted comparing African Americans to Whites, Hispanics to Whites, and females to males using IRT-based approaches (Raju, Drasgow, & Slinde, 1993) and a Mantel-Haenszel statistic (Holland & Thayer, 1988) that does not require IRT parameter estimation. Each of these approaches is described briefly here, followed by a summary of the screening decision that resulted from the DIF analyses. Mellon et al. (1996) provides more detail on both the application of the different procedures and the results.

IRT approaches involve estimating separate item characteristic curves (ICCs) for the two groups being compared. ICCs give the probability of a correct answer as a function of true ability. ICCs could not be estimated for some items in some ethnic groups. The largest difficulty was with Vocabulary items in the Hispanic samples. These items were eliminated from further DIF analyses (they were considered, however, for inclusion on Forms E and F). For the remaining items, the EQUATE program (Baker, Al-Karni, & Al-Dosary, 1991) was used to adjust the underlying ability scales for true differences in the two samples being compared using the equating method developed by Stocking and Lord (1983). Differences in the ICCs for the two groups being compared were summarized in terms of (a) the area between these two curves after the lower asymptote (guessing parameter) was constrained to be the same, and (b) a chi-square statistic given by Lord (1980). The two statistics agreed closely, and the chi-square statistic was used in subsequent item screening.

The Mantel-Haenszel estimates DIF as a ratio of the probability of passing the item for members of the two groups that is assumed to be constant across different ability levels. Items have DIF to the extent that this odds ratio differs from one (or the log of the ratio differs from zero). Observed number correct scores are typically used to equate ability levels for individuals within each ethnic or gender group, and proportion passing statistics are analyzed within each score level. Software written by Hambleton and Rogers (1993) was used to generate Mantel-Haenszel statistics for each of the power test items.

The primary goal of the DIF analyses was to identify those items having clearly aberrant DIF. Items were flagged if DIF statistics were significant at p<.01 using either of the two procedures. After inspection of the flagged items, 15 items were dropped for extreme DIF values. One of these items, a Form A vocabulary item, had been flagged for DIF in the Hispanic sample by the IRT procedure but not the Mantel-Haenszel procedure. All other items had been flagged by both procedures. Table 2-6 shows the number of items screened out by the DIF analyses as well as by other factors and the number of items that remained for possible use in the final forms.

That only items with extreme DIF were dropped reflected a strategy of controlling DIF at the test level by balancing items with positive and negative DIF, rather than controlling tightly at the item level. Eliminating items with marginal DIF could inadvertently narrow the content of the test, perhaps reducing its validity.

Table 2-6
Item Screening Results for the Power Tests

	Test		
Screen	Arithmetic Reasoning	Vocabulary	3-D Space
Total E & F items tried out	66	160	130
# dropped for biserial < .15	1	3	8
# excluded from IRT DIF analyses	1	10	7
# dropped for DIF (African American-White)	0	4	0
# dropped for DIF (Hispanic-White)	0	0	0
# dropped for DIF (Female-Male)	1	3	0
Total new items remaining	63	140	122

Calibration and Screening of Speeded Test Items

Item analysis procedures for speeded tests are considerably less standardized relative to power test item analysis procedures. In some models, where all of the emphasis is on assessing speed rather than accuracy, traditional concepts of difficulty and discrimination do not apply. Without computer administration, however, it is not possible to record response latencies on an item-by-item basis and so analyses of speed at the item level are not feasible.

For the speeded tests on the GATB, there is some evidence that items do vary in terms of difficulty, with response errors being more common for some items than for others. It was therefore essential that measures of item difficulty and discrimination be analyzed so the new forms could be balanced with respect to these statistics. Given the importance placed on test fairness, analyses of DIF across gender and ethnic groups were also deemed essential.

A special item analysis program was developed to compute the proportion passing (difficulty) and point biserial correlation (discrimination) for each item using only those examinees who reached the item. If the proportion passing an item were based on all examinees, items toward the end of a speeded test would automatically have low values, because very few examinees reach that point in the test and attempt to answer the item. Although only items in the first block of each speeded test were analyzed, the relationship of item position to the number passing the item was substantial. Computing the proportion passing only for those examinees who reached the item greatly reduced the artificial confounding of an item's difficulty estimate and its position in the tryout booklet.

The relationship between item position and point biserials computed on all examinees was equally substantial. For example, for the Object Matching test, the correlation between item position and point biserials averaged .91 across the eight item blocks when all examinees were included in the computation of the point biserials. When examinees who did not reach the item in question were excluded, however, the average correlation decreased to .65.

The specific item and test statistics computed for the speeded tests included:

- *Mean and standard deviation of the total score on the analyzed items.*
- *Mean and standard deviation of the item-corrected total test scores.* The item-corrected total test score is the total test score **excluding** the item in question.
- *Item difficulty*. The difficulty of an item is assessed in classical test theory by the proportion of examinees who answered the item correctly. Because this computation was based only on those examinees who reached the item in question, the results are referred to as "corrected difficulty values."
- *Point biserial correlations*. Point biserial correlations are an index of item discrimination in classical test theory. Point biserials were computed based on all examinees and also based only on the examinees reaching the item in question (corrected point biserials). Only the corrected point biserials were used in the item screening and form construction analyses.
- *Response-alternative information*. For each item response alternative, the proportion of examinees who endorsed that alternative, the mean corrected total test score for those examinees, and the point biserial correlation between the response alternative and the corrected total test score were computed.

The statistics were also computed separately for separate examinee groups, including females, males, African Americans, Hispanics, and whites. Because total scores are based more on speed than on accuracy, the models for DIF that group examinees on the basis of total score and look for differences in item passing rates within total score levels do not appear applicable. In the present analyses, DIF was analyzed in terms of effect sizes computed as the standardized difference in corrected passing rates (i.e., proportion correct calculated only on those individuals who reached the item) between each focal group (females, African Americans, or Hispanics) and the corresponding reference group (males or whites). Although this strategy can cause problems when applied to power tests (e.g., lower reliability and validity of the resulting test), the approach is reasonable when selecting items for highly speeded tests, because few if any substantial group differences are typically encountered for such tests. Item selection for the Computation subtest was supplemented by examining Mantel-Haenszel DIF statistics computed by HRStrategies.

Item response alternative data were also examined, primarily to identify very poor items (e.g., those with positive point biserials for the distractor; a distractor endorsed by examinees whose mean test score excluding that item was higher than the mean score for those who correctly responded to the item). No items were excluded for this reason.

Experimental items for the Computation test were tried out in two item orders: forward and reversed. Correlations between item difficulty estimates from the odd numbered booklets (forward direction) were correlated with the difficulty estimates obtained from the corresponding even numbered booklets (reverse direction). The correlations were .79 (for booklets 1 and 2) and .81 (for booklets 9 and 10), suggesting that item difficulties are highly similar, regardless of order of presentation. Consequently, item data were pooled across the corresponding booklets in the computation of item statistics.

To minimize the negative impact on subgroups, items evidencing a subgroup effect size greater than 0.25 (that is, one-quarter standard deviation) were removed from further consideration. Item response alternative data were also examined for the presence of distractors with positive point biserials, or mean item-corrected total test scores that were (a) based upon a sample size of at least ten or more, and (b) higher than the mean item-corrected total test scores for those who responded correctly to the item. No items were removed from the Object Matching and Name Comparison tests for these reasons. Eighteen experimental Computation items were flagged by the DIF analysis. Of these, eight were included in the proposed Forms E and F (six on Form E, two on Form F). In no instance did the selected items possess effect sizes of greater than one-half standard deviation. Table 2-7 details the characteristics of the items flagged by the DIF analysis.

Selection of Items for Final Versions of Forms E and F

Items for Forms E and F power tests were selected by first creating information graphs for the base form (A) and draft new Forms E and F. Information graphs were calculated for the thetas between -2 and 2. Information for Form A was calculated by summing the information for all items, and then scaling down to the length of the new forms. For example, Form A Vocabulary had 40 items. Forms E and F will have 19 items. Thus, at every theta, information for Vocabulary Form A was multiplied by 19/40.

In designing the new Forms E and F, an attempt was made to improve measurement significantly over Form A between thetas of ±1, where approximately 70 percent of examinees score, and where most decisions are likely to be made.

Table 2-7
Summary Item Statistics for Computation Items Flagged by the DIF Analysis

Original GATB Form	Original Item #	Item Type	Sub-group showing DIF	p-value	Corrected Point Biserial	Effect Size	New Form	New Item #
E	12	Add	W-H	.804	.328	-.104	E	21
E	16	Mult	M-F	.797	.333	-.243	NA	NA
E	21	Div	M-F	.872	.362	.040	E	4
E	23	Mult	W-B	.844	.317	-.002	F	11
E	26*	Add	M-F	--	--	--	--	--
E	30	Add	W-H	.906	.230	-.161	E	9
E	36	Sub	W-B	.669	.449	.465	E	34
E	40	Add	W-H	.908	.246	-.039	E	5
E	60	Div	M-F	.873	.261	.105	NA	NA
E	65	Div	M-F	.763	.311	-.311	NA	NA
E	67	Div	W-B	.336	.220	.044	NA	NA
F	7	Div	W-B	.357	.394	.556	NA	NA
F	8	Div	M-F	.837	.307	.031	F	8
F	19	Sub	W-H	.746	.424	-.048	NA	NA
F	29	Add	W-H	.724	.443	-.100	NA	NA
F	31	Sub	W-B	.670	.483	.527	NA	NA
F	42	Div	W-H	.722	.423	.298	NA	NA
F	47	Sub	M-F	.638	.337	-.066	E	34

* Form E item 26 was misprinted in one test booklet. It was omitted.

Selection of Power Items

The analyses and report of power test item selection were conducted by HRStrategies. Excerpts from their report of procedures and results (HRStrategies, 1994) are included here. Previous work conducted by Measured Progress had produced draft forms for all three power tests. These forms were preliminary, however, in that DIF analyses had not been completed. These draft forms were used as a starting point in drafting the new Forms E and F of the power tests using the full pool of items that passed the screening for DIF.

Items were selected in accordance with professionally accepted practices (see Hambleton & Swaminathan, 1985; Lord, 1977) to meet the following objectives:

- The new Forms E and F power tests must be fair[3] for candidates of all backgrounds.
- The new Forms E and F power tests must be as parallel as possible.
- The new Forms E and F power tests should improve upon the measurement properties of Form A.
- To the extent possible, the new Forms E and F power tests should provide measurement that is uniformly precise over a broad range of abilities.

Item selection proceeded as follows:

- The draft forms created by Measured Progress served as a starting point. These initial forms were based on a desire to have informative (i.e., discriminating) items that spanned the range of item difficulties. In actuality, there were many informative, difficult items and few informative, easy items. Therefore, item selection focused on the need to provide adequate precision below theta = -1.
- Any item discarded during the DIF screening was deleted from the form.
- Using software written in Pascal, item information functions were computed at 13 theta points for all items in the pool, separately by gender and ethnic group. The number of theta points used appeared more than adequate for comparing the smooth information curves observed in the present research.
- Informative items were sought to replace those deleted as a result of the DIF analyses.
- The test information functions were then computed for Forms A, E, and F as the sum of the item information functions. The use of a spreadsheet facilitated computation and graphing of the test information functions for various choices of items.

The above steps were iterated when comparisons between forms suggested a lack of correspondence between E and F, or when the measurement properties of A were superior to those of E and/or F for some range of abilities.

Some of the items included in the forms may have been associated with small to moderately large DIF statistics during screening. Therefore as a final check, measurement equivalence for ethnic and gender groups was checked at the test level.

Assessment of Test-Level Equivalence. The final step in forms development was to assess the equivalence of the forms on a total test-score level to ensure that expected scores of equally able persons were essentially the same. It was crucial to make this last, test-level comparison because decisions will be made on the basis of the total test score.

The procedure for this comparison was to

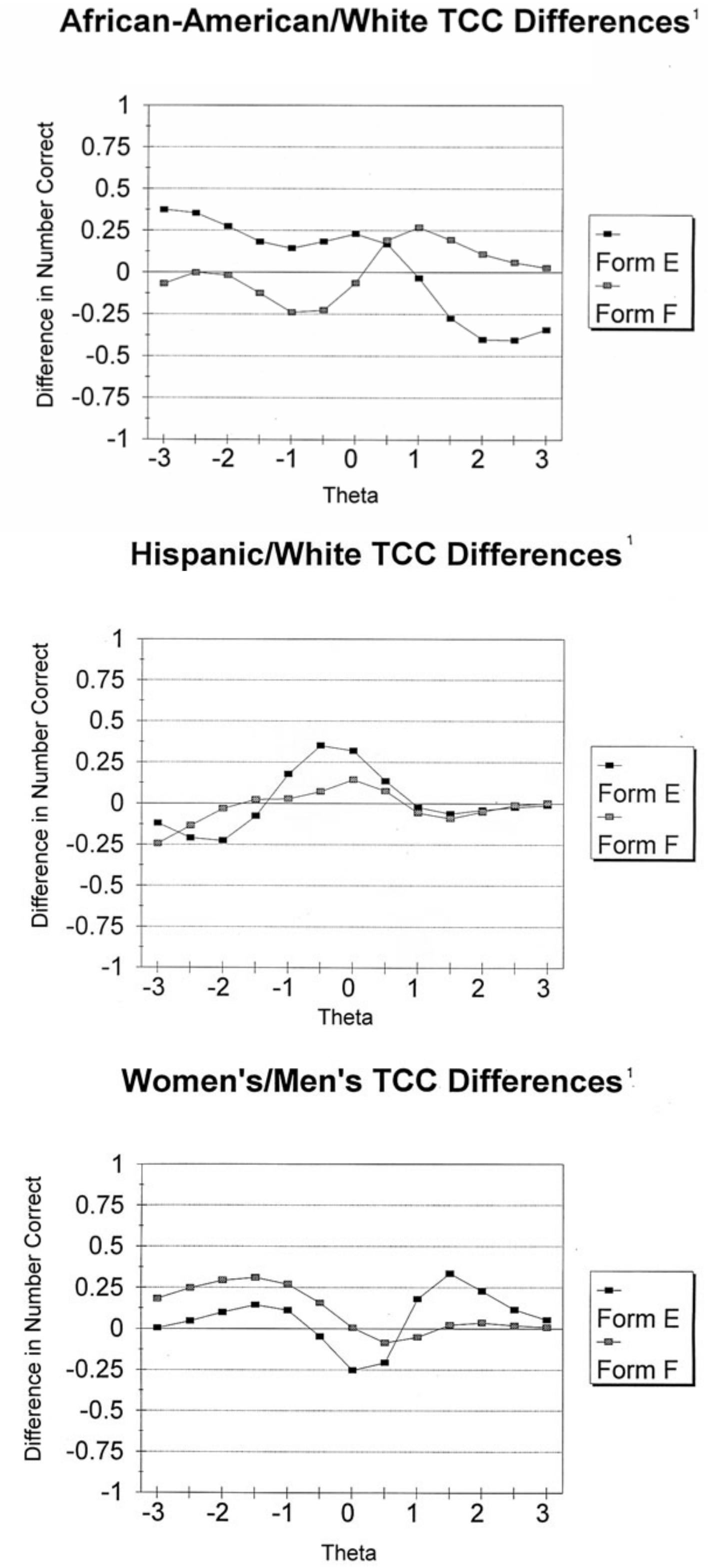
[3] Fair, as it is used here, means that individuals with the same level of ability have the same expected score.

- Use the final linear linking parameters computed previously to match each focal group ability metric to the reference group metric (Stocking & Lord, 1983) using EQUATE (Baker, Al-Karni, & Al-Dosary, 1991); and
- Plot the TCCs for each group on a common metric.

The plots of the expected number correct differences for the final forms are presented in Figures 2-1 through 2-3 for the final Arithmetic Reasoning, Vocabulary, and Three-Dimensional Space tests, respectively. These difference plots subtract the reference group expected score from the focal group expected score so that regions where the lines lie above zero indicate that the focal group is advantaged; the distance above or below zero indicates the degree of advantage or disadvantage for the focal group, respectively, in raw score points. In general, the graphs show that deviations from perfect fairness are small. In no case was the difference in expected scores greater than half a raw score point. Further, differences did not consistently favor either the reference or focal groups; the lines lie above and below zero for all comparisons. As a result, it was concluded from these analyses that the two forms of each of the three power tests are fair for African Americans, Hispanics, and whites, and for males and females.

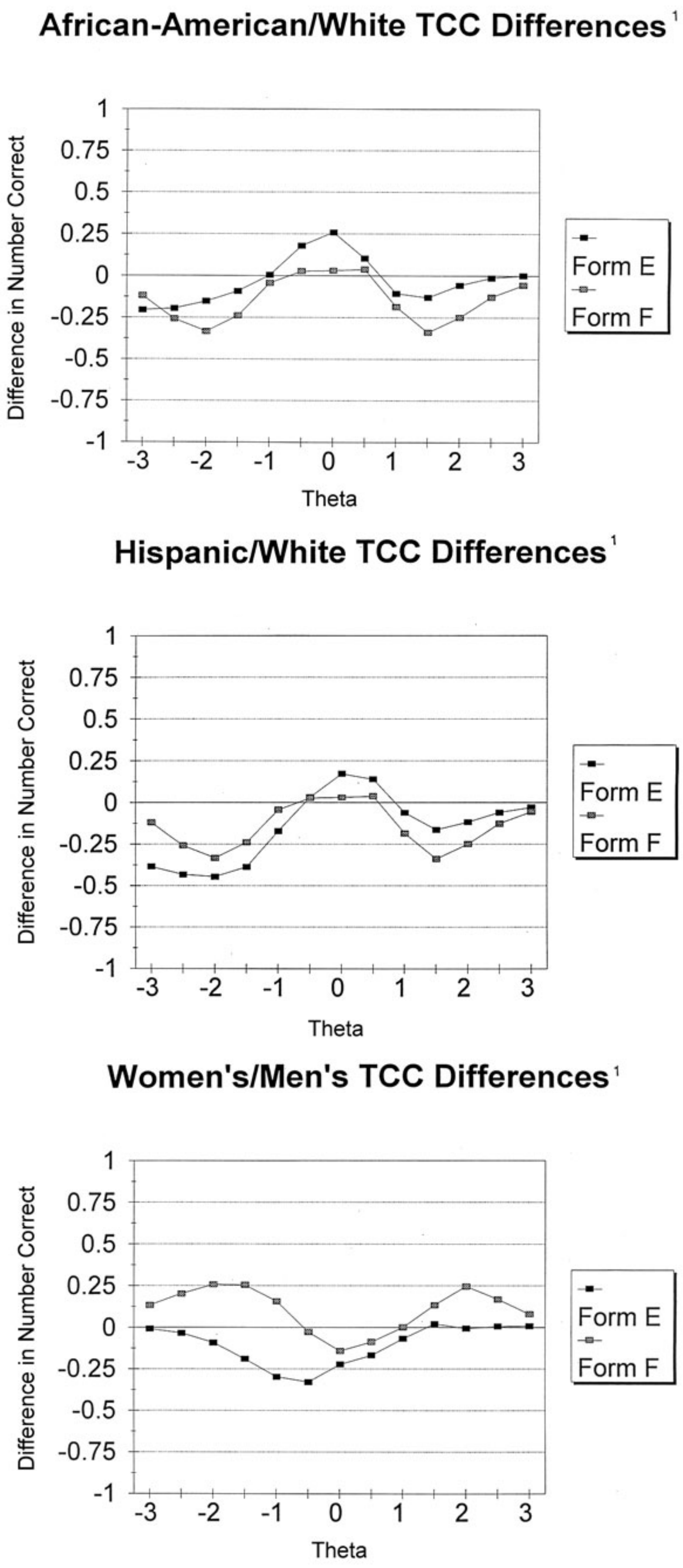
The draft forms were modified three times to create parallel, precise, and fair instruments. Each modification was evaluated by developing test information function graphs which were examined and discussed.

Discrepancies were observed between the expected scores of some reference and focal groups for some forms. Therefore, the objective of the first set of modifications was to ensure that the forms were fair. This was accomplished by deleting items exhibiting DIF in undesirable directions and adding items showing either no DIF or DIF in the desired direction. For example, if one form showed advantage for males near theta=0, then one or more items with DIF favoring males near zero would be deleted and replaced with items showing no DIF or showing DIF favoring females in that range.



[1] Differences between the expected number correct scores of focal and reference group members. Points above zero indicate advantage for the focal group.

Figure 2-1. Test Characteristic Curve (TCC) Differences for Arithmetic Reasoning



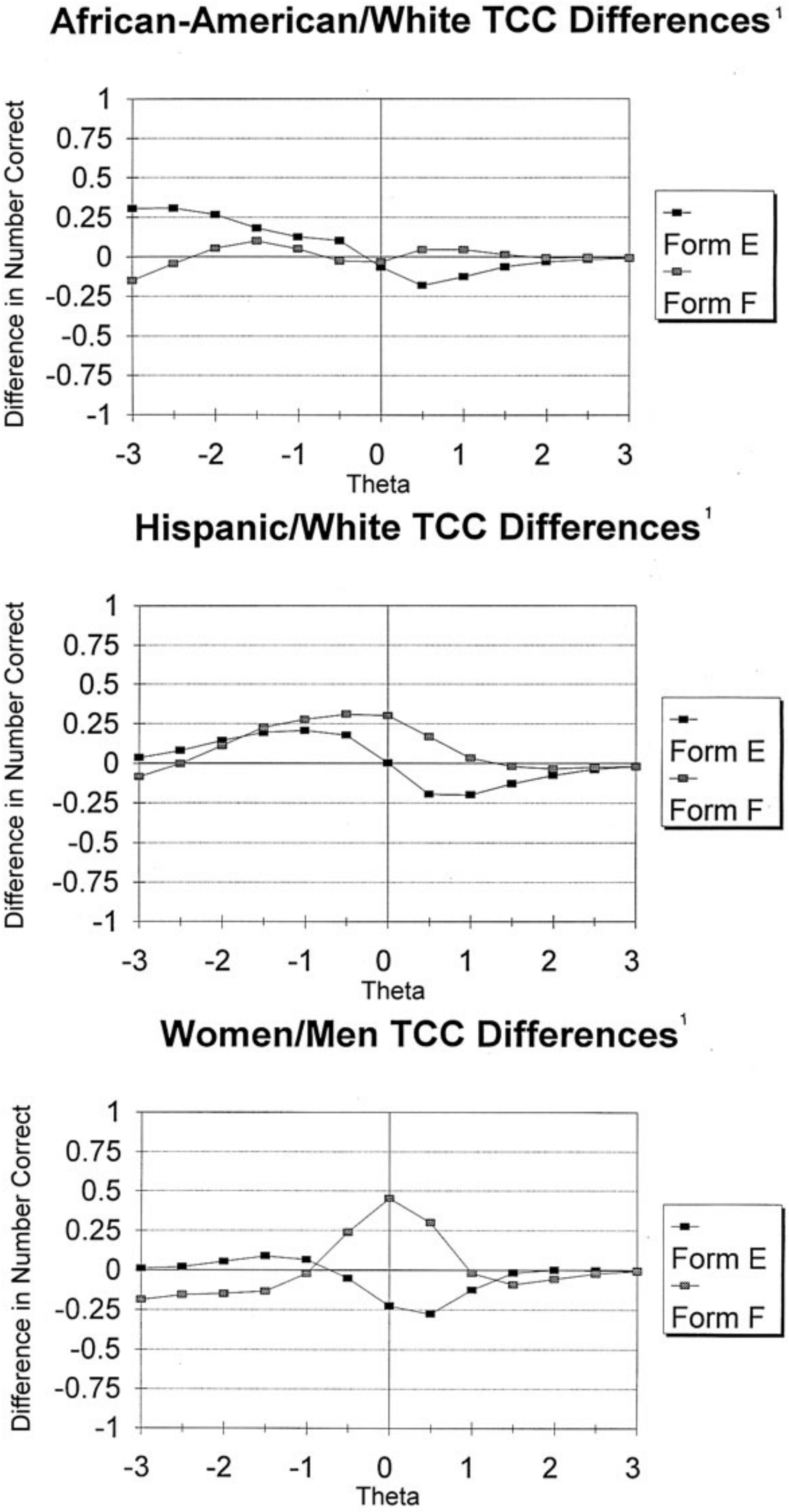
African-American/White TCC Differences[1]

Hispanic/White TCC Differences[1]

Women's/Men's TCC Differences[1]

[1] Differences between the expected number correct scores of focal and reference group members. Points above zero indicate advantage for the focal group.

Figure 2-2. Test Characteristic Curve (TCC) Differences for Vocabulary


African-American/White TCC Differences[1]

Hispanic/White TCC Differences[1]

Women/Men TCC Differences[1]

[1] Differences between the expected number correct scores of focal and reference group members. Points above zero indicate advantage for the focal group.

Figure 2-3. Test Characteristic Curve (TCC) Differences for Three-Dimensional Space

The resulting forms provided less test information than (i.e., were not as precise as) the original forms and in some regions were somewhat less precise than Form A. Thus, a second round of modifications was made to make the forms more precise without sacrificing fairness. This required adding a small number of Form A items to the final versions of Forms E and F. For Arithmetic Reasoning, the new Forms E and F contain two and one items from anchor Form A, respectively. For Three-Dimensional Space, Forms E and F contain two items each from Form A. No Form A items were added to the Vocabulary tests. The third set of modifications was performed to erase a moderate disparity between the expected scores of African Americans and whites on one of the forms of Arithmetic Reasoning.

The resulting forms were fair within the limits imposed by the pool of items. Some compromise had been made in terms of precision to have equitable forms. Compared to the gains in equity, however, the loss of precision was small, and the new forms are generally comparable to Form A. The degree of difference between reference and focal groups is roughly comparable to that found by Drasgow (1987) for the ACT English Usage and Mathematics Usage tests.

Measurement Properties of the Final Forms. The measurement properties of each iteration of Forms E and F, as well as the anchor Form A, were examined from several perspectives, including inspection of three types of graphs based on parameters estimated in the total sample. First, curves displaying test information at each of 13 theta levels were evaluated. Second, the test information curves of Forms E and F were graphed along with graphs showing the ratio of Form E or F information relative to Form A information at each of these points. In addition, because test information may not be easily interpretable to all readers, a third graphic based on a heuristic procedure was used to provide estimates on a reliability scale.

For the absolute and relative information graphs, the information for a hypothetical, shorter Form A was estimated by multiplying by the number of items on Forms E and F divided by the number of items on Form A. For example, Form A of the Three-Dimensional Space test had 40 items, whereas Forms E and F had 20. To make the forms comparable, the test information plotted for Form A was multiplied by 20/40 = 0.50 at each theta point prior to plotting.

The absolute information graphs for the final forms are provided here as Figures 2-4 through 2-6. The graphs generally show that the new forms are about as informative as Form A, except for the Vocabulary test, where Forms E and F have much more information than Form A at low to moderate ability ranges. These differences are magnified in the Relative Information graphs presented in Mellon et al. (1996). On the Relative Information graphs, it is important not to over-interpret large relative differences between curves of small absolute information. For example, on page 10, the graph shows that the new forms are more than 60 percent less informative than Form A at theta = +3. The absolute amount of information for any form at theta = +3, however, is very low.

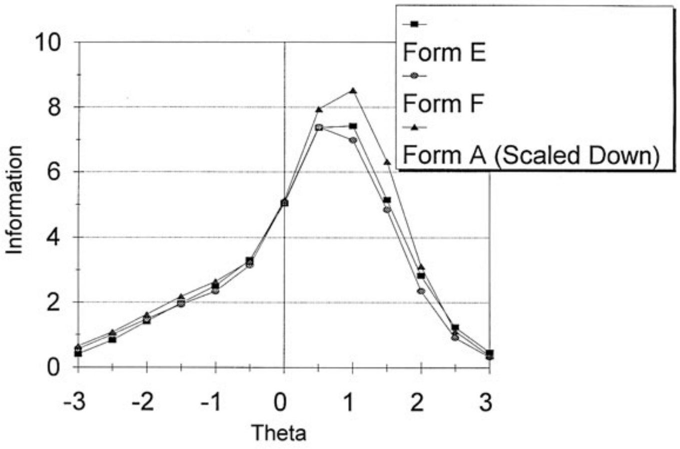


Figure 2-4. Arithmetic Reasoning Absolute Information Graph

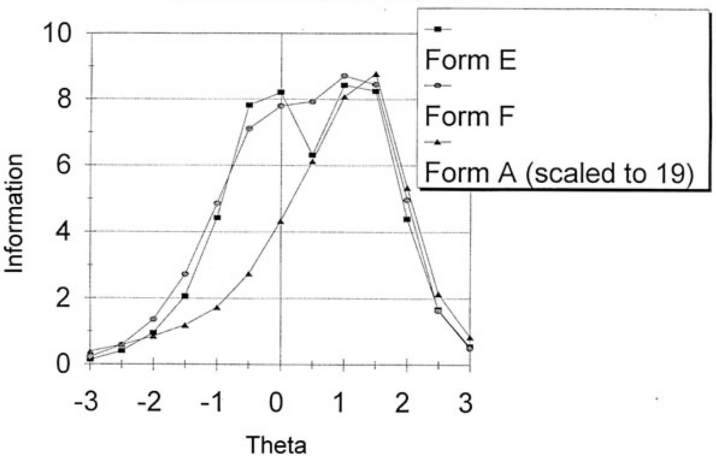


Figure 2-5. Vocabulary Absolute Information Graph

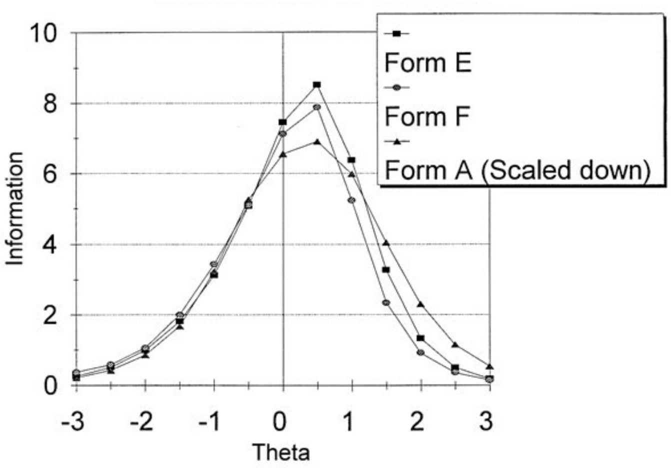


Figure 2-6. Three-Dimensional Space Absolute Information Graph

In sum, the information graphs demonstrate that the new Forms E and F are about as informative as Form A. For Arithmetic Reasoning, the new forms are not quite as precise as Form A, particularly over the range 0 < theta < 2.5.

Selection of Items for the Speeded Tests[4] [5]

Item Selection Strategy. The statistics for items constituting the extant GATB forms suggest that the speeded test items assess something other than simply a speed component. A test that assesses speed alone should contain only items with very high *p* values. The current GATB forms comprise items that demonstrate a range of *p* values. Therefore, pursuant to recommendations from the ARDP, items were selected that were similar statistically to current GATB forms.

Specifically, items were selected that approximated the item difficulty data calculated from operational administrations of GATB Forms B and D. These data were not used as the primary criteria for item selection; rather, the data from Forms B and D were used to provide rough guidelines for the item difficulty distributions of the selected items. More stringent attempts to match the item statistics of the selected experimental items to those of the items from Forms B and D were deemed unwise, given that the data from Forms B and D (a) were obtained several years ago, (b) reflect a different sample of examinees, (c) were collected under different circumstances (operational conditions, as opposed to the current experimental conditions), and (d) appear to treat omitted items as missing observations rather than incorrect responses.

Because of the differences between the data for the experimental items and those for Forms B and D, experimental item difficulty distributions for the Form Matching, Object Matching, and Name Comparison tests were examined and compared to the same distributions for the first n_i items (n_i = the number of experimental items that were analyzed for test *i*) from the operational forms. In all instances, the ranges of difficulties were approximately equivalent.

For the Computation test, comparison of the item statistics of the experimental items to item statistics from a previous form was accomplished by embedding anchor items from GATB Form A within the experimental Computation test. The section on item selection for Computation Forms E and F details a comparison of the item statistics for the experimental items with the statistics for the Form A items. These analyses provide a more useful comparison of the characteristics of the selected items to previous GATB forms, because the data describe item responses from the same individuals, under the same administration conditions, at the same testing time.

In general, the approach used to select items for each of the speeded tests was to maximize the ability of the tests to discriminate between those high on the attribute and those low on the attribute, while minimizing differential impact on various subgroups. For the Computation test, DIF information (a list of items found significant through Mantel-Haenszel tests) was computed and reviewed. The section describing the selection of Computation items details the use of these data.

The following sections include a description of the item selection strategies adopted for each of the four tests. Mellon et al. (1996) presents tables that contain item analysis information for the items that were selected, as well as all other experimental items.

Item Selection for the Object Matching Test. Using the statistics from the item analyses as a guide, 42 experimental Object Matching items were selected for each new form. Items were chosen which maximize discriminability and minimize differential impact on subgroups. Hence, items with large point biserial correlations and small effect sizes were targeted. To bolster the effect size data, graphs of corrected item difficulty were produced for the gender and ethnic groups. In addition to these psychometric concerns, experimental items were examined to ensure they represent the various content specifications (e.g., distribution of keyed responses,

[4] Form Matching is not discussed in this section, because the decision to drop it was made before the analyses were conducted.

[5] Much of this section was excerpted from McCloy et al. (1994).

number of shaded areas in the stimulus) as specified in *Test Specifications for GATB Forms E and F* (PARDC, 1995).

Items were sorted by item discrimination values. Because the correlation between item order and point biserial correlation did not vanish, items were selected only if the original target of 95 percent of the examinees reached the item. Hence, those items at the end of the test booklets often were not considered, even though they demonstrated sizable point biserials. The possible influence of careless responders on the point biserial values for these items could have contributed to these large values. Therefore, the additional selection criterion was used.

Characteristics of the Proposed Object Matching Test. The proposed Forms E and F for the Object Matching test and their suggested administration order are given in Appendix J of Mellon et al. (1996). The forms are virtually identical in their mean difficulty (.894 and .893, respectively) and discrimination (.277 and .263, respectively). In addition, the effect size statistics indicated that the selected items provide little aggregate differential impact across subgroups. These values were all low and similar across test forms: For male/female, white/African American, and white/Hispanic comparisons, the values were -0.062, 0.109, 0.008, for Form E; and -0.028, 0.101, and 0.015 for Form F; respectively.

Item Selection for the Name Comparison Test. Using the item statistics as a guide, 90 items were selected for each new form. Items were chosen to maximize discriminability and minimize differential impact on subgroups. Hence, items with large point biserial correlations and small effect sizes were targeted. In addition to these psychometric concerns, the experimental items were examined to ensure they represent the various content considerations (e.g., male and female names, Hispanic names, number of characters in the stimuli) as specified in *Test Specifications for GATB Forms E and F* (PARDC, 1995).

Once selected, the items were sorted by keyed response and corrected difficulty values, and the content specifications of the items were tallied. Although the psychometric qualities of the initial proposed Forms E and F were satisfactory, the content specification tally indicated 23 male items appeared on Form E, whereas Form F comprised only 14. Thus, items were shifted between forms, replacing the male items from Form F with indeterminant items from Form E. Once sorted into the final proposed forms, the items were again sorted for each form by keyed response and corrected difficulty. Items were selected to obtain 45 “Same” items and 45 “Different” items.

To determine the order of administration within each proposed form, a table of random units was used to provide a keyed response pattern for each form. Merely ordering items from easiest to most difficult was infeasible for the Name Comparison test, because the Same items were far easier than the Different items. Ordering by difficulty alone would have resulted in all the Same items appearing at the front of the test. Consequently, the sequencing of Same and Different items was randomized. After generating the response pattern, the items were ordered so that the easiest item having the required keyed response was selected. Thus, the item position represents an ordering by item difficulty (from easiest to most difficult), conditional upon the randomly generated keyed response pattern. After examination of the test mockups, however, it became apparent that some adjustments to item ordering were required (e.g., four consecutive Form F items contained the abbreviation “Co.”). Two items on Form E and three items on Form F were exchanged to eliminate this type of content repetition.

Regarding the content characteristics of the selected Name Comparison items, consideration was given to the number of characters in the stimulus name (the name on the left side of the pair) ranging from 6 to 25, the ethnic group the name represents (A = Asian, B = African American, H =

Hispanic, O = Other), the gender of the name (M = male, F = female), whether it was a business (B) or personal (P) name, and the type of difference that occurred for dissimilar items (1 = inverted letters, 2 = same sound/different meaning, 3 = different sound and letter, 4 = addition or deletion of a single letter, and 5 = different word forms but similar meaning). McCloy et al. (1994) and Mellon et al. (1996) include more detailed information on the distribution of items across these categories in the final forms.

An attempt was made to select items that represent a fair distribution of the relevant content characteristics across forms. Although there is less balance for certain characteristics, greater emphasis was placed upon assuring balance of the gender/ethnic name representation and the Business/Personal characteristics, in particular. The Type of Difference characteristic is less balanced across proposed test forms, but the items function equivalently from a psychometric perspective. No items with "Type of Difference" equal to "5" (i.e., different word forms but similar meaning) were chosen for the new test forms. These items demonstrated large differential impact to Asians. Specifically, Asians were disproportionately likely to miss these items that, although structurally very different, mean the same thing. For example, item 40 from books 7 and 8 has a white/Asian effect size of 1.128. The item is

Nat'l Pane Company --- National Pane Company

Clearly, the elements of the item pair differ structurally; yet, they have the same meaning. Confusion of structural similarity with content similarity created numerous problems that are apparent in the data. Because of the substantial differential impact these items cause, they do not appear on GATB Forms E and F.

The proposed Forms E and F are very similar, with mean corrected difficulty values of .921 and .934, respectively. Mean discrimination values are virtually identical (.337 and .338, respectively). In addition, the effect size statistics indicate that the selected items provide little aggregate differential impact across subgroups. These values are all low and similar across test forms: For the male/female, white/African American, white/Hispanic, and white/Asian comparisons, the values are -.061, .075, .009, and .086 for Form E; and -.054, .090, .012, and .078 for Form F, respectively.

Item Selection for the Computation Test. For the Computation test, the primary goal was to select 80 items with desirable psychometric and content properties for inclusion in GATB Forms E and F (i.e., 40 items for each form). The experimental versions of GATB Forms E and F included Computation test items in four booklets: Booklets 1 and 2 (Form E), and Booklets 9 and 10 (Form F). Form E and Form F each had 68 experimental items. Within each form, items were presented in one order in one book and in reverse order in the second book of the same form (e.g., Booklet 1 item 1 was the last item in Booklet 2). In addition, 50 items from operational Form A were embedded in all four booklets as anchor items.

It is also important to note that the Computation test includes four major types of items--addition, subtraction, multiplication, and division. Within each type, content characteristics of the items vary, making some more difficult than others. For example, some addition items involve simply adding two single-digit numbers, whereas others require adding five or six multi-digit numbers. The final version of both Form E and Form F was to include 40 items, 10 of each type of operation, and within each type of operation, items should represent the range of content characteristics as specified in *Test Specifications for GATB Forms E and F* (PARDC, 1995).

Items were chosen to maximize discriminability and minimize differential impact on subgroups. Hence, items with large point biserial correlations and small effect sizes were targeted.

Analysis of Form A Anchor Items. The data for Form A items (which were anchor items on all forms) were compared with the experimental item data to see how comparable they were in terms of item difficulty and discriminability. In general, the entire pool of experimental items was slightly more difficult than the pool of Form A items. Mean difficulties for Form A and experimental items respectively were .80 and .78 for addition, .85 and .79 for subtraction, .79 and .75 for multiplication, and .71 and .66 for division (Mellon et al., 1996, Appendix L). The experimental items were somewhat less discriminating than the Form A items. Mean corrected point biserials for Form A and experimental items respectively were: .31 and .23 for addition, .37 and .33 for subtraction, .42 and .35 for multiplication, and .43 and .38 for division.

Items were selected based on their psychometric properties, and race and sex differences. Items were pooled across forms and organized according to the type of arithmetic operation. Items with point biserials one or more standard deviations below the mean point biserial for that type of operation were identified. As mentioned earlier, Mantel-Haenzel chi-square analyses conducted by HRStrategies (1994) were also used to examine DIF. Items that (a) were flagged by the DIF analysis and (b) showed more than half a standard deviation difference between subgroups, were not selected. Eighteen experimental items were flagged by the DIF analysis. Of these, eight were retained (six on Form E, two on Form F). In no instance did the selected items possess effect sizes of greater than one half standard deviation.

The pool of remaining good items was sorted according to content characteristics, and the items were organized across forms. After making an initial cut across forms, the mean item statistics were computed for each new form. To make the forms as similar as possible (statistically as well as in content), some items were shifted from one form to the other, yielding the final set of proposed items for Forms E and F. Table 2-8 summarizes the statistics for the selected items of each of the proposed forms.

Items were ordered as follows: (1) an addition item, (2) a subtraction item, (3) a multiplication item, and (4) a division item. This sequence of item types was repeated throughout the test. Items were ordered according to difficulty (within type of operation). For example: Item #1 contained the easiest addition item, Item #5 contained the next easiest addition item, . . ., and Item #37 contained the most difficult addition item. This process was repeated for each type of item until all items were selected.

Table 2-8
Computation Test Item Selection: Summary Statistics for Selected Form E and Form F Items

	Item	Corrected Difficulty		Point Biserials		Effect Size		
Form	Type	Mean	SD	All Items	W/O Item	M-F	W-B	W-H
E	Addition	0.79	0.38	.31	.29	-0.08	0.21	0.02
	Subtraction	0.79	0.39	.35	.33	-0.14	0.30	0.12
	Multiplication	0.76	0.40	.40	.37	-0.09	0.24	0.15
	Division	0.69	0.44	.41	.38	-0.07	0.28	0.20
	Grand Mean	0.76	0.41	.37	.34	-0.09	0.26	0.12
F	Addition	0.80	0.38	.32	.29	-0.07	0.23	0.09
	Subtraction	0.77	0.39	.38	.36	-0.11	0.27	0.12
	Multiplication	0.75	0.41	.40	.37	-0.09	0.24	0.15
	Division	0.69	0.43	.46	.43	-0.09	0.35	0.19
	Grand Mean	0.75	0.40	.39	.36	-0.09	0.27	0.14

Forms Equating Study[6]

Forms E and F Test Tryout Data Collection Procedures

The Test Tryout data collection involved a nationwide sample of 8,975 individuals that was representative of the applicant populations of local Employment Service offices. Participants were reimbursed for their travel expenses. The amount of reimbursement varied by condition and geographical region, but it was not dependent on test performance. These sample data were used for equating new Forms E and F to base Form A. The technical requirements and rationale for the procedures discussed below are presented in detail in Segall and Monzon (1995).

Data Collection Design

The data collection design is presented in three sections, each section corresponding to one of the three primary samples included in the GATB equating study. The first section describes the data collection design for the independent-groups (IG) sample. This sample was used to equate the new and old GATB forms. The second section describes the data collection design for the repeated-measures (RM) sample. This sample was used primarily for comparing the reliability and construct validity of the new and old forms. However, a portion of this sample was used as supplemental data for the equating analysis. The third section describes the data collection design for the psychomotor (PM) sample. This sample was used to examine the need for composite equatings, and to examine construct validity issues involving the psychomotor tests.

Independent-Groups Sample. Examinees in the independent-groups sample were randomly assigned to one of three forms (i.e., A, E, and F). As indicated in Table 2-9, a total of 5,892 examinees were tested. Approximately equal numbers of examinees were tested on each of the three forms ($N \approx 1{,}964$). Table 2-9 also displays the numbers of examinees tested on each form at each of the five ARDCs.

Across each of the five ARDCs, there were approximately 40 testing sites. At each site, examinees were randomly assigned to test form (A, E, or F). The old (A) and new (E and F) forms of the GATB possess different test ordering, time limits, and instructions, thus complicating the assignment of test forms to examinees. Consequently, these versions cannot be administered to a single group simultaneously. They must be administered in different testing sessions, where the sessions are separated physically either by location (testing room) or by time. Hence, at a given testing site, one of two methods of assignment was used, depending on whether one or two testing rooms were available:

- At two-room sites, examinees were randomly assigned to Forms A, E, and F upon arrival. Examinees assigned to Form A were tested in one room; examinees assigned to either of Forms E or F were tested in a second room.
- At one-room sites, some sessions were dedicated to Form A, and other sessions were dedicated to the new Forms E and F. All examinees at one-room sites were scheduled for testing prior to their arrival at the test site. At the time of scheduling, each examinee was randomly assigned to one of the three forms (A, E, or F). Once assigned to a specific form, the examinee was given a choice of several test dates that had been dedicated to the assigned form.

[6] Much of this section excerpted from Segall and Monzon (1995).

Table 2-9
Independent-Groups Sample Sizes

ARDC	Form A	Form E	Form F	Total
EARDC	447	370	389	1,206
NARDC	436	370	401	1,207
SARDC	301	330	334	965
PARDC	402	372	392	1,166
WARDC	455	456	437	1,348
Total	2,041	1,898	1,953	5,892

Repeated-Measures Sample. Examinees in the repeated-measures sample were administered two forms of the GATB. These data were used primarily for examining the reliability and construct validity of the GATB. However, a portion of these data were also used to supplement the equating data. These data were used to perform a detailed comparison of measurement properties between the old and new forms.

Each examinee participating in the repeated-measures portion of the study was randomly assigned to one of eight conditions. These conditions and the numbers of examinees in each condition are presented in Table 2-10. Note that conditions 1, 2, 7, and 8 consist of samples of approximately equal size (N ≈ 430). The remaining conditions listed in Table 2-10 also consist of approximately equal sample sizes (N ≈ 218). The rationale for the sample size requirement is outlined in the data analysis section. The numbers of examinees tested in each condition at each site are provided in Table 2-11.

Table 2-10
Repeated Measures Design and Sample Sizes

	Second Test			
First Test	A	B	E	F
A		1 (411)		3 (236)
B	2 (432)		5 (209)	
E		6 (215)		7 (446)
F	4(216)		8 (446)	

Table 2-11
Repeated-Measures Sample Sizes by Test Site

	Condition								
ARDC	1	2	3	4	5	6	7	8	Total
EARDC	43	46	27	22	24	25	53	48	288
NARDC	81	82	41	45	40	40	88	93	510
SARDC	94	100	52	51	47	45	76	77	542
PARDC	107	98	63	49	53	49	103	108	630
WARDC	86	106	53	49	45	56	126	120	641
Total	411	432	236	216	209	215	446	446	2,611

At two-room sites, examinees were randomly assigned to the eight conditions upon arrival. At one-room sites, all examinees were scheduled for testing prior to arrival. This procedure ensured random assignment of examinees to condition.

Psychomotor Sample. This sample of 538 examinees received the five psychomotor tests along with the non-psychomotor portions of Forms A and F. The design is presented in Table 2-12. Examinees were randomly assigned to one of two groups. Each group received three sections: (1) Form A (non-psychomotor), (2) Form A (psychomotor), and (3) Form F (non-psychomotor), with the order of presentation counterbalanced across the two groups. As indicated in Table 2-12, Group 1 received Form A (non-psychomotor) and Form A (psychomotor) portions in the morning session, and Form F (non-psychomotor) in the afternoon. Group 2 received the same battery of tests with the order of the non-psychomotor sections of Forms A and F reversed.

At two-room sites, examinees were randomly assigned to the two conditions upon arrival. At one-room sites, all examinees were scheduled for testing prior to arrival. This procedure ensured random assignment of examinee to condition.

Sample Characteristics

This section provides an evaluation of the demographic characteristics of each of the three samples included in this study: (1) the independent-groups (IG) sample, (2) the repeated-measures (RM) sample, and (3) the psychomotor (PM) sample. An evaluation of the random equivalence of selected groups within each of the three samples is also provided, because the random equivalence of these groups is a key assumption made in the equating, reliability, and validity analyses.

Table 2-12
Psychomotor Data Collection Design

	Group 1 (*N* = 265)	Group 2 (*N* = 273)
Morning		
	1. Form A (non-pmotor)	1. Form F (non-pmotor)
	2. Form A (pmotor)	2. Form A (pmotor)
Afternoon		
	3. Form F (non-pmotor)	3. Form A (non-pmotor)

This section also summarizes the data editing procedure used to remove unmotivated examinees and other highly influential cases from each of the three samples. The numbers of cases removed from each sample are reported.

Demographics and Group Equivalence. For each of the three primary samples, statistical tests of the differences among randomly equivalent groups were conducted by gender, race, age, and education. Non-significant results are consistent with the expectation based on random assignment of examinees to condition and support the assumption of equivalent groups made in the equating, reliability, and validity analyses. All significance tests of differences across groups by form (for the IG sample) and by condition (RM and PM samples) yielded non-significant results with $\alpha = .05$ (cf. Segall & Monzon, 1995, for detailed results).

The results of these analyses indicate diverse samples with respect to gender, race, age, and education. Furthermore, the significance tests performed on the three samples (independent-groups, repeated-measures, and psychomotor) provide reassurance that the assignment procedures worked as intended, producing groups that are randomly equivalent with respect to demographic characteristics. Although the equivalence of the groups on cognitive and psychomotor abilities cannot be tested with existing data, the results based on the demographic variables provide additional confidence in this assumption, since in some instances demographic and cognitive/psychomotor variables tend to be correlated.

Outlier Analysis. Prior to data analysis, a small number of cases with unlikely scores were deleted from the database. Cases for deletion were identified using a procedure suggested by

Hotelling (1931), which identifies cases that are unlikely given that the observations are sampled from multivariate elliptical-shaped distribution.

Separate outlier analyses were performed for the three samples (independent-groups, repeated-measures, and psychomotor). Furthermore, separate analyses were performed for each group, within each sample. Note that it is possible for two types of patterns to be flagged and deleted by the chosen procedure. One type occurs when an examinee receives extreme scores on many tests (e.g., all low scores). Another type of unlikely pattern occurs when an examinee scores high on one test and low on a second that is highly correlated with the first (i.e., alternate forms of the same test). A small number of cases with zero number-right scores on one or more tests were also deleted, because DOL policy dictates that scores should not be provided to such examinees.

Independent-Groups Sample. Table 2-13 provides the editing results for the independent-groups sample. Here, selected cases from the RM sample were combined with the IG sample. Specifically, data from the first test administered (Forms A, E, or F) of Conditions 1, 3, 4, 6, 7, and 8 were combined with the IG data to increase the sample sizes for the equating study. Data editing was performed in each of the three IG/RM groups: "A," "E," and "F." A few cases were removed from each group, ranging from 14 to 34 examinees. The final group sizes used in the equating analysis are listed in the last row of the table.

Table 2-13
Group Sizes for the Edited Independent Groups/Equating Sample

		Form	
Sample	**A**	**E**	**F**
IG	2,041	1,898	1,953
RM-1 (A/B)	411		
RM-3 (A/F)	236		
RM-6 (E/B)		215	
RM-7 (E/F)		446	
RM-4 (F/A)			216
RM-8 (F/E)			446
Total *N*	2,688	2,559	2,615
Number Deletes	34	14	18
Final *N*	2,654	2,545	2,597

Repeated-Measures Sample. Table 2-14 displays the editing results for the repeated-measures sample. Between 3 and 13 cases were deleted from each of the eight groups. The final group sizes after editing are displayed on the bottom row of the table.

Table 2-14
Group Sizes for the Edited Repeated-Measures Sample

				Sample				
Frequency	1	2	3	4	5	6	7	8
Total *N*	411	432	236	216	209	215	446	446
Number Deletes	13	10	4	3	3	5	12	10
Final *N*	398	422	232	213	206	210	434	436

Psychomotor Sample. For the PM sample, four cases were deleted from Group 1, and zero cases were deleted from Group 2, yielding final group sizes after editing of 261 examinees in Group 1 and 273 examinees in Group 2.

Summary Demographic Data. Summary demographic data for the three samples are presented in Tables 2-15 through 2-17, and for the aggregate sample in Table 2-18.

Scoring the GATB

A number of GATB scores are routinely produced and used. Scoring for the new Forms E and F is complicated by the absence of the Form Matching test, and by the use of formula scores for the speeded tests. The method of score computation for the old and new GATB forms is detailed below. This description includes the computation of both test and composite scores.

Table 2-15
Demographic Composition of the Independent-Groups/Equating Sample

Group	Sample Size	Percent of Total Sample
Total Group	7,796	100
Gender		
Female	3,582	46
Male	4,209	54
Not Reported	5	<1
Race/Ethnic Group		
African American	2,931	38
Asian	205	3
Hispanic	971	12
White	3,445	44
Native American	115	2
Other	112	1
Not Reported	17	<1
Age Group		
15	1	<1
16 - 40 Years	5,220	67
Over 40 Years	2,550	33
Not Reported	25	<1
Education		
Less than 6	2	<1
6 - 11 Years	1,401	18
12 Years	3,361	43
13 - 15 Years	1,984	25
16 Years and Over	1,030	13
Not Reported	18	<1

Table 2-16
Demographic Composition of the Repeated-Measures Sample

Group	Sample Size	Percent of Total Sample
Total Group	2,589	100
Gender		
Female	1,084	42
Male	1,502	58
Race/Ethnic Group		
African American	1,184	46
Asian	79	3
Hispanic	355	14
White	876	34
Native American	48	2
Other	43	2
Not Reported	4	<1
Age Group		
16 - 40 Years	1,772	68
Over 40 Years	808	31
Not Reported	9	<1
Education		
Less than 6	1	<1
6 - 11 Years	456	18
12 Years	1,184	46
13 - 15 Years	625	24
16 Years and Over	315	12
Not Reported	8	<1

Table 2-17
Demographic Composition of the Psychomotor Sample

Group	Sample Size	Percent of Total Sample
Total Group	538	100
Gender		
Female	220	41
Male	318	59
Race/Ethnic Group		
African American	225	42
Asian	4	1
Hispanic	37	7
White	251	47
Native American	12	2
Other	9	2
Age Group		
16 - 40 Years	347	64
Over 40 Years	191	36
Education		
6 - 11 Years	78	14
12 Years	242	45
13 - 15 Years	154	29
16 Years and Over	64	12

Table 2-18
Demographic Composition of the Aggregate Sample

Group	Sample Size	Percent of Total Sample
Total Group	8,975	100
Gender		
Female	4,064	45
Male	4,905	55
Not Reported	6	<1
Race/Ethnic Group		
African American	3,447	38
Asian	230	3
Hispanic	1,089	12
White	3,924	44
Native American	140	2
Other	127	1
Not Reported	18	<1
Age Group		
15	1	<1
16 - 40 Years	5,998	67
Over 40 Years	2,949	33
Not Reported	27	<1
Education		
Less than 6	2	<1
6 - 11 Years	1,602	18
12 Years	3,891	43
13 - 15 Years	2,292	26
16 Years and Over	1,169	13
Not Reported	19	<1

Test Scoring. For the purpose of this study, all scores for Forms A and B were computed according to conventions specified in the "Manual for the USES General Aptitude Test Battery" (Section I).

Forms A and B. The set of raw scores for Forms A and B, denoted by

$$\{X_{\mathrm{ar}}(A), X_{vo}(A), \ldots, X_{di}(A); X_{ar}(B), X_{vo}(B), \ldots, X_{di}(B)\} \quad ,$$

are taken as the simple sum of the number of correct responses for the power and speeded tests. The raw scores for the five psychomotor tests are obtained according to the operational procedures. Standard-scores S are obtained from a look-up table. These standard scores are summed in various combinations to form the nine aptitude scores $\{A_g\ A_v, A_n, A_s, A_p, A_q, A_k, A_f, A_m\}$ displayed in Table 2-19. The conversion of raw score to standard score depends on the aptitude for which the test scores will be used. For tests that enter into two different aptitude

Table 2-19
Aptitude Score Composition

GATB Test	Aptitude Score								
	A_g	A_v	A_n	A_s	A_p	A_q	A_k	A_f	A_m
Arithmetic Reasoning	$S_{ar}^{(g)}$		$S_{ar}^{(n)}$						
Vocabulary	$S_{vo}^{(g)}$	$S_{vo}^{(v)}$							
3D Space	$S_{3d}^{(g)}$			$S_{3d}^{(s)}$					
Computation			$S_{co}^{(n)}$						
Name Comparison						$S_{nc}^{(q)}$			
Object Matching					$S_{om}^{(p)}$				
Form Matching					$S_{fm}^{(p)}$				
Mark Making							$S_{mm}^{(k)}$		
Place									$S_{pl}^{(m)}$
Turn									$S_{tu}^{(m)}$
Assemble								$S_{as}^{(f)}$	
Disassemble								$S_{di}^{(f)}$	

scores, there are two conversion tables. For tests that enter into only one aptitude score, there is a single conversion table. Table 2-19 provides the notation for the test standard score and the aptitude score composition. Aptitude scores are formed from the simple sum (down each column of Table 2-19) of test standard scores.

Forms E and F. The set of raw scores for forms E and F, denoted by

$$\{X_{ar}(E), X_{vo}(E), \ldots, X_{om}(E); X_{ar}(F), X_{vo}(F), \ldots, X_{om}(F)\} \quad ,$$

are taken as either (a) the simple sum of the number of correct responses (for the three power tests AR, VO, and 3D), or (b) the chance-corrected formula score (for the three speeded tests CO, NC, and OM). The formula scores for each of the speeded tests are given by the general formula

$$X \;=\; NC - \frac{W}{(\;A\;-\;1\;)} \quad , \tag{1}$$

where NC is the number of correct responses, W is the number of wrong answers (items answered incorrectly—does not include omits or not-reached), and A is the number of response options associated with the test items. For the three speeded tests, the preceding equation simplifies to

$$\begin{aligned} X_{co}(f) &= NC - \frac{W}{4} \\ X_{nc}(f) &= NC - W \\ X_{om}(f) &= NC - \frac{W}{3} \quad , \end{aligned} \tag{2}$$

where f equals “E” or “F.”

Standard scores S for Forms E and F are obtained from look-up tables produced from the equating described in the following section. Thus, the process of computing aptitude scores is identical to that described above for Forms A and B, with one exception. Because Form Matching was dropped from the new forms, the aptitude score A_p is set equivalent to $S_{om}^{(p)}$, rather than computed as $A_p = S_{om}^{(p)} + S_{fm}^{(p)}$ (as in Forms A and B). Note that the distributions of A_p across the new and old forms are ensured to be equal through the appropriate specification of the equating transformation. This point is described in more detail in a later section.

Composite Scoring. In addition to the aptitude score composites described above, two other sets of composites were studied: three *component* composites and five *job-family* composites. Each of the three component composites was computed from the sum of selected aptitude scores:

$$\begin{aligned} \textit{Cognitive}: \quad C_{gvn} &= A_g + A_v + A_n \\ \textit{Perceptual}: \quad C_{spq} &= A_s + A_p + A_q \\ \textit{Psychomotor}: \quad C_{kfm} &= A_k + A_f + A_m \quad . \end{aligned} \tag{3}$$

Scores on the cognitive and perceptual composites were calculated for all examinees in the equating, reliability, and validity analyses. These composites were included in all key analyses. However, scores on the psychomotor composites were calculated only for examinees in the psychomotor sample, which was used to address selected composite equating and validity issues.

In addition, to address these same composite equating and validity issues, five job-family composites were computed from the weighted linear combinations:

$$\begin{aligned} J_1 &= .59 \, x \, C_{gvn} + .30 \, x \, C_{spq} + .11 \, x \, C_{kfm} \\ J_2 &= .13 \, x \, C_{gvn} + .87 \, x \, C_{kfm} \\ J_3 &= C_{gvn} \\ J_4 &= .73 \, x \, C_{gvn} + .27 \, x \, C_{kfm} \\ J_5 &= .44 \, x \, C_{gvn} + .56 \, x \, C_{kfm} \quad . \end{aligned} \tag{4}$$

Because all composites but one (which is redundant with C_{gvn}) are a function of one or more psychomotor tests, these composites were computed only for the psychomotor sample.

Smoothing and Equating

The objective of equipercentile equating is to provide a transformation that will match score distributions of the new forms with the distribution of scores from the reference Form A. This transformation, which will be applied to the new Forms E and F, will allow scores on the new versions to be interpreted relative to the old scale represented by Form A.

One primary objective of the method of equating proposed here was to use smoothing procedures that provide an acceptable tradeoff between random and systematic error. In this study, smoothing (specifically, polynomial log-linear smoothing) was performed on each distribution (of Forms A, E, and F) separately. These smoothed distributions were then used to specify the equipercentile transformation (see Segall & Monzon, 1995, for a more detailed discussion of issues pertinent to the equating transformation and the smoothing procedures).

The data used in this analysis were provided from two samples: the independent-groups sample, and the repeated-measures sample. Data collected on the first-administered test from selected groups of the repeated-measures sample were combined with same-form data of the independent groups sample. The sample sizes used to estimate the Forms A, E, and F distributions appear in Table 2-13.

Zero Cells. The log-linear smoothing procedure is undefined for score levels that have a frequency of zero. For these analyses, a two-step procedure was used for specifying zero-cell entries. First, the maximum score was identified for each test. Then the distribution smoothing was performed for the raw-score range "l - max," where "max" is the maximum score for the test. Any score levels having no observations falling within the "l - max" range were changed to 0.5 for log-linear estimation. For the speeded-test distributions of Forms E and F, the low range was set equal to the observed minimum in the sample, because the minimum can fall below 1 for formula scores.

P-Aptitude (A_p). The new Forms E and F do not contain the Form Matching test. In the old Forms A-D this test was used in the A_p Score (see Table 2-19). The A_p score (Forms A-D) was computed from the sum of Form Matching and Tool (Object) Matching. Rather than equating Object Matching across the new and old forms, Object Matching (Forms E and F) was equated directly to the A_p distribution (Form A). This direct equating will allow scores on Object Matching (Forms E and F) to be transformed to an A_p score having the same distribution as Form A, even though the Form Matching test has been omitted from the new forms.

This matching involved several steps. First, the distribution of A_p was computed from Form A data by summing the standard scores of Form Matching and Tool Matching. Next, the distribution of A_p scores was smoothed by applying log-linear smoothing. The final equating was obtained by matching this smoothed distribution with the smoothed Object Matching distributions of Forms E and F, using the equipercentile procedure.

Polynomial Extrapolation. Several of the new and old tests differ substantially in their length, which leads to a significant difference in their maximum attainable scores. For example, Vocabulary had 60 items in Form A and 19 items in Forms E and F. A maximum score on the Form A version represented several standard deviations above the mean, while a maximum score on Form E/F represents less than two standard deviations above the mean. Consequently, matching the maximum scores on the new and old versions (a consequence of equipercentile equating) does not appear to be appropriate.

Although the three speeded tests were shortened, the test content and time limits remained virtually unchanged across the new and old GATB versions. Because shortening these tests did not cause a noticeable ceiling effect, the maximum score on the new versions was set equal to the equated scores on Form A.

For the three power tests (AR, VO, and 3D), a polynomial extrapolation was used to specify the maximum equated score. The highest five points of the equipercentile transformation were approximated by a second-order polynomial fitted with least-squares. The equated value for the maximum score level of Forms E and F was set equal to the predicted value of the resulting polynomial.

Equating Transformations. A total of 18 smoothings were performed—6 smoothings for each of the three GATB forms. These smoothed distributions were used to compute 12 equating transformations—6 transformations equating Form A and E tests, and another 6 transformations equating Forms A and F.

Tables given in Segall and Monzon (1995) provide a translation between Form E/F raw scores x and Form A standard scores. Linear interpolation was used in conjunction with equipercentile equating to specify appropriate standard scores. The equated standard scores were obtained from

$$S(x) = S_L + \frac{F_{E/F}(x) - C_L}{C_U - C_L} [S_U - S_L] \ , \qquad \textbf{(5)}$$

where $F_{E/F}(x)$ is the cumulative distribution function on the new form evaluated at raw-score level x, C_L and C_U are the lower and upper values of the Form A cumulative distribution interval which contains $F_{E/F}(x)$:

$$C_L \leq F_{E/F} \leq C_U \ , \qquad \textbf{(6)}$$

and S_L and S_U are the Form A standard-scores corresponding to the interval defined by (C_L,Cy). The formula for S(x) was used to specify raw to standard score conversions for each raw-score level of Forms E and F.

Composite Equating

Equating the new and old paper-and-pencil (P&P) GATB forms involves matching test distributions using an equipercentile method. This distribution matching provides a transformation of Forms E and F to standard-score equivalents on the reference form (Form A) scale. Once this transformation is specified for each test, standard-score equivalents can be computed. These standard-score equivalents provided the basis for the computation of GATB composites. The same formulas used to compute composites from standard scores on Form A can be used to compute composite scores from standard-score equivalents on the new forms.

One concern is that the distribution of composite scores from the new forms will differ systematically from the corresponding distributions of the old forms. This difference could result from differences in test intercorrelations between the old and new forms. Different test intercorrelations may result from one or more revisions made to the new P&P-GATB (changes in test lengths, time limits, instructions, etc.). Because the variance of a composite is a function of the correlations among its tests, differences in composite variances could result as a consequence. Higher order moments of the composite distributions could be affected in a similar manner.

Non-Psychomotor Composites. The data used in this analysis were the same data used to estimate the equating transformations. These data were provided from the independent-groups sample and the repeated-measures sample. Data collected on the first-administered test from selected groups of the repeated-measures sample were combined with same-form data of the independent-groups sample. The sample sizes used to examine composite distributions across Forms A, E, and F are provided in Table 2-13.

The distributions of several composites were examined. Comparison of composite score distributions involved a number of steps. First, scores on Forms E and F were transformed to standard-score equivalents using the transformation estimated from the equating. Next, for each composite, scores were computed for three groups:

1. Composite scores were obtained for those examinees taking Form A by applying the composite formulas to the standard scores. Four composites were examined: A_g and A_n (aptitude scores), and C_{gvn} and C_{spq} (Cognitive and Perceptual composites). All other aptitude scores are a function of a single test; thus, the agreement of their distributions across the new and old forms will be guaranteed by the equipercentile

transformation. Because the A_p aptitude score distribution of Form A was equated directly to the A_p score distribution of Forms E and F (which is a nonmonotonic function of Object Matching), these distributions are also matched through the equating transformation, and thus no confirmation is necessary. Composites containing tests were analyzed separately (see section below).

2. Composite scores were obtained for those examinees taking Form E. Composite scores were obtained for A_g, A_n, C_{gvn}, and C_{spq} by applying the composite formulas to the standard score equivalent scores.
3. Composite scores were obtained for those examinees taking Form F. Composite scores were obtained for A_g, A_n, C_{gvn}, and C_{spq} by applying the composite formulas to the standard score equivalent scores.

The distributions of scores for each of the new forms (Forms E and F) were compared to the corresponding composite distribution of the reference form (Form A). Four cut points were used to divide the distribution into five groups. Cut scores were based on the area under a normal density function. The *z* values (computed from Form A means and standard deviations) that divided the distribution into groups having the expected proportions

$$\{.10, .25, .30, .25, .10\}$$

were applied to the composite distributions to produce the observed proportions displayed in Tables 7.1-7.4 of Segall and Monzon (1995). The proportion of examinees falling in each group was compared across the two new versions (E and F) and the single old version (A). The significance of the difference in these proportions was examined using a 3 x 5 contingency table analysis. The Pearson χ^2 statistic was used to test the null hypothesis of no difference among distributions.

Although two composites (A_g and C_{spq}) have marginally significant differences, an examination of the distributions indicated the four composites are very similar across the new and old forms. None of the composites was significant at the .01 level. These results suggest that the standard GATB composite formulas can be applied to the equated standard scores of the new forms, and that these composite scores will have similar distributions across the new and old versions. Therefore, separate composite equating tables for the non-psychomotor composites are unnecessary for the new Forms E and F.

Psychomotor Composites. In this study, the evaluation of composite equating is complicated by the absence of the psychomotor tests from the new Forms E and F. New versions of the psychomotor tests are being developed under a separate data collection and analysis effort. Because some of the composites computed under the job-family system include a combination of psychomotor and non-psychomotor tests, a complete evaluation should examine the similarity of these distributions (across old and new forms). Data for this analysis were collected from the psychomotor sample. This sample was administered one old Form (A), one new Form (F), and one form of the psychomotor portion of the battery (from Form A).

Distributions of composite scores were compared across the two groups using data collected in the morning session only (Table 2-12). For Group 1 (which received Form A psychomotor and non-psychomotor tests), four composites (J_1, J_2, J_4, J_5) were computed according to the procedures described earlier. Group 2 scores for the corresponding composites were computed in a similar manner from standard-score equivalents using the Form F (non-psychomotor) and Form A (psychomotor) portions of the battery.

Four cut points were used to divide the distributions into five groups as in the analysis above. Cut scores were based on the area under a normal density function. The *z* values (computed from Group 1 means and standard deviations) that divided the distribution into groups having the expected proportions {.10, .25, .30, .25, .10} were applied to the composite distributions. The proportion of examinees falling in each group was compared across the new (Form F) and old (Form A) versions. The significance of the difference in these proportions was examined using a 2 x 5 contingency table analysis. The Pearson χ^2 statistic was used to test the null hypothesis of no difference among distributions.

The results indicated that these proportions are very similar across new and old GATB forms, and do not differ significantly from what would be expected from sampling error (see Segall & Monzon, 1995, for detailed results). These results imply that the job-family composite formulas can be applied to the equated standard scores of the new forms—and that these composite scores will have similar distributions across the new and old versions. These results suggest that separate composite equating tables for the psychomotor composites are unnecessary for the new Forms E and F.

One key assumption of this analysis is that the new forms of the psychomotor tests (currently under development for Forms E and F) will be parallel to the form used in this study. If they are not parallel, then the covariances between these tests and the non-psychomotor tests may be poorly represented by those obtained in the current study. That is, if the new and old forms of the psychomotor tests are not parallel, then the results obtained in this study may not generalize to the new psychomotor tests.

Subgroup Comparisons

Although equipercentile equating matches test distributions for the total sample, it does not necessarily guarantee a match for distributions of subgroups contained in the sample. This result follows from the fact that the new and old versions of the P&P-GATB are not strictly parallel. Although we might expect to observe small differences in subgroup performance across the new and old versions as a result of differences in measurement precision, many of the other revisions made to the new forms could also cause group differences. It is therefore instructive to examine the performance of subgroups to determine whether any are placed at a substantial disadvantage by the new forms, relative to their level of performance on the old GATB forms.

In the analyses described below, four subgroups were examined: (1) African Americans, (2) Hispanics, (3) females, and (4) examinees 41 years of age or older. The equating transformation based on the total sample was applied to subgroup members who had taken the new GATB forms (E and F). For each subgroup, mean performance levels were compared across new and old forms. Six test variables were examined: $S_{ar}^{(g)}$, $S_{vo}^{(g)}$, $S_{3d}^{(g)}$, $S_{co}^{(n)}$, $S_{nc}^{(q)}$, and A_p. These variables are monotonic functions of the six test scores. (For Form A, the variable A_p is a function of both Form Matching and Object Matching.)

The significance of the differences among means was examined for Forms A, E, and F using ANOVA. Among the 24 comparisons, only one was significant at the .01 level. Specifically, a significant difference was observed across A, E, and F for African Americans on the Vocabulary test—African Americans administered Forms E and F tended to score slightly higher than those administered Form A. In general, the results indicate similar average performance levels across new and old versions for each of the four subgroups examined (cf. Segall & Monzon, 1995, for detailed results).

Reliability Analysis

A primary issue in the investigation of new GATB forms is that of precision. Several of the new test versions have fewer items than their original counterparts. Although fewer items may be offset by an increase in testing time, it is important to show that the new forms have sufficiently high levels of reliability relative to the old GATB forms. Lower reliability would lead to lower levels of validity.

Method. Four groups of the repeated-measures sample were used in this analysis (see Table 2-10). Data from Groups 1 and 2 were combined to form a sample of 820 examinees whose data were used to compute the alternate form correlations between the old Forms A and B. These correlations are displayed in Table 2-20 for test, aptitude-score, and composite variables. Groups 7 and 8 were combined to form a sample of 870 examinees whose data were used to compute the alternate form correlations between the two new Forms E and F. These alternate form correlations are also displayed in the table.

Fisher's z transformation was used to test the significance of the difference between the alternate form correlations of the new and old GATB forms. As described in Cohen and Cohen (1983, p. 54), the significance of the difference between two correlation coefficients obtained from two different random samples can be evaluated from the normal curve deviate

$$z = \frac{z'_n - z'_o}{\sqrt{\frac{1}{n_n - 3} + \frac{1}{n_o - 3}}} \quad , \qquad \textbf{(7)}$$

where

$$\begin{aligned} z'_n &= \frac{1}{2}[\ln(1 + r_{nn'}) - \ln(1 - r_{nn'})] \quad , \\ z'_o &= \frac{1}{2}[\ln(1 + r_{oo}) - \ln(1 - r_{oo'})] \quad , \end{aligned} \qquad \textbf{(8)}$$

and $r_{nn'}$ is the alternate form correlation for the new test variable (based on Forms E and F), $r_{oo'}$ is the alternate form correlation for the old test variable (based on Forms A and B), and n_n and n_o are the sample sizes for the groups used to compute the alternate form correlations (n_n = 870, n_o = 820). Normal deviates z were computed for each test, aptitude score, and composite variable. The results are displayed in Table 2-20. Also displayed are the probability values associated with these normal deviates, $1 - \mathbf{\Phi}(|z|)$, where $\mathbf{\Phi}$ is the normal cumulative distribution function.

Results. The alternate form reliabilities of the new GATB forms are generally as high as, or higher than, those of the old GATB Forms A and B—an encouraging finding, because the length of the three power tests was decreased. The increase in testing time, however, may have added to the reliability of these power tests, offsetting the detrimental effects of shortening test lengths. Only one comparison displayed a significantly lower alternate-form correlation for the new form: $S_{co}^{(n)}$ (Computation). The magnitude of the difference is small, however, and none of the composites that use Computation displays a significantly lower new-form reliability estimate.

Validity Analysis

This section addresses the third primary issue in the evaluation of the new forms—construct validity. It is highly desirable for the new and old GATB forms to measure identical or highly correlated constructs. The measurement of similar constructs would enable the validity of the new forms to be inferred from the large body of existing validity research conducted on the old forms of the GATB.

Table 2-20
Alternate-Form Reliability Estimates, Normal Deviates, and p-Values

	Correlations		**Significance Test**	
Variable	$r_{nn'}$	$r_{oo'}$	z	p
$S_{ar}^{(g)}$	.800	.803	-.179	.429
$S_{ar}^{(n)}$	.800	.802	-.127	.450
$S_{vo}^{(g)}$	.846	.859	-.925	.177
$S_{vo}^{(v)}$	.850	.858	-.580	.281
$S_{3d}^{(g)}$	.829	.805	1.529	.063
$S_{3d}^{(s)}$	.832	.805	1.648	.050
$S_{co}^{(n)}$	.818	.846	-1.873	.031
$S_{nc}^{(q)}$	.778	.755	1.145	.126
$S_{om}^{(p)}$	.823	.770	2.996	.001
A_g	.908	.886	2.284	.011
A_v	.850	.858	-.580	.281
A_n	.876	.884	-.704	.241
A_s	.832	.805	1.648	.050
A_p	.823	.824	-.049	.480
A_q	.778	.755	1.145	.126
C_{gvn}	.919	.913	.772	.220
C_{spq}	.893	.849	3.694	.000

Note:$r_{nn'}$ = new form reliability;
$r_{oo'}$ = old form reliability;
$p = 1 - \Phi(|z|)$

The construct validity analysis is presented in two parts. The first part describes an analysis of the non-psychomotor test and composite variables based on the repeated-measures data. The second part addresses the construct validity of variables that enter into the job-family composites (i.e., the cognitive, perceptual, and psychomotor composites). These analyses are based on the psychomotor sample.

Non-Psychomotor Construct Validity. For this analysis, all eight groups of the repeated-measures sample were used (see Table 2-10). Data from Groups 3, 4, 5, and 6 were combined to form a sample of 861 examinees. For the purpose of these analyses, scores on Forms A and B were treated as an "Old Test" variable; scores on Forms E and F were treated as a "New Test" variable. Correlations between the old and new batteries (denoted by $r_{n,o}$) were obtained for test, aptitude, and composite variables. These correlations are given in Table 2-21.

The alternate-form reliability estimates computed from Groups 1 and 2 (old forms) and Groups 3 and 4 (new forms) were used to obtain the disattenuated correlations between the new and old forms. These alternate-form correlations were computed as described in the previous section. The disattenuated correlations were computed from the Classical Test Theory expression

$$\rho(\tau_n, \tau_o) = \frac{r_{n,o}}{\sqrt{r_{nn'} \; x \; r_{oo'}}} \quad . \qquad \textbf{(9)}$$

These values also appear in Table 2-21. (Asymptotic standard errors of the disattenuated correlations were obtained using a structural equation modeling approach; see Segall & Monzon, 1995, Appendix I.)

Table 2-21
Disattenuated Correlations Between New and Old GATB Forms

Variable	$r_{n,o}$	$\rho(\tau_n,\tau_o)$
$S_{ar}^{(g)}$	.762	.951
$S_{ar}^{(n)}$	.765	.955
$S_{vo}^{(g)}$	.802	.940
$S_{vo}^{(v)}$	.801	.938
$S_{3d}^{(g)}$	.743	.910
$S_{3d}^{(s)}$	.743	.908
$S_{co}^{(n)}$	.806	.969
$S_{nc}^{(q)}$	.730	.952
$S_{ob}^{(p)}$	.721	.905
A_g	.857	.955
A_v	.801	.938
A_n	.864	.982
A_s	.743	.908
A_p	.748	.907
A_q	.730	.952
C_{gvn}	.887	.968
C_{spq}	.817	.938

As indicated, all disattenuated correlations between old and new forms were extremely high, ranging from .905 to .982. Although numerous changes were made to the test battery (e.g., test format, time limits, test lengths, deletion of the Form Matching Test, change in scoring formulas), these changes do not appear to have significantly altered the dimensionality of the battery. Because of the high correlations between the dimensions measured by the new and old forms, the large number of validity studies conducted on the old GATB forms can continue to provide useful data for inferring the validity of the new GATB forms.

Psychomotor Construct Validity. Five job family composites were used for job counseling and referral with the VG-GATB program. There is a chance that these composites could continue to be used on a limited basis for job counseling with GATB Forms E and F. Because four of the fives composites are computed, in part, on the basis of psychomotor tests, a complete evaluation of the new GATB forms should include the psychomotor tests. If, for example, the covariance between psychomotor and non-psychomotor tests differed across the old and new forms, then the validity of the five job-family composites might be affected. That is, these five composites computed from the old forms might measure different traits than those computed from the new forms.

This analysis used 534 subjects from the psychomotor-sample. As indicated in Table 2-12, each examinee was administered Form A (non-psychomotor), Form A (psychomotor), and Form F (non-psychomotor) portions of the GATB. For each examinee, five scores were computed: $C_{gvn}(A)$, $C_{spq}(A)$, $C_{gvn}(F)$, $C_{spq}(F)$, *and* $C_{kfm}(A)$, where $C_{gvn}(A)$ denotes the cognitive composite computed from Form A tests; $C_{gvn}(F)$ denotes the same cognitive composite computed from Form F tests, etc. To compute the Form F composites, scores were transformed to Form A “standard-score equivalents” using the equating transformation presented in equation 5. The formulae given in equation 3 were applied to the aptitude scores to compute the five composites.

Table 2-22 displays the correlations among these five variables. As indicated in the last row, the patterns of correlations between the psychomotor composite C_{kfm} and the cognitive and perceptual composites (C_{gvn} and C_{spq}) appear to be similar[7] across Forms A and F (.27 vs .23, .44 vs .42). These similar patterns provide some assurance that the same relations among cognitive, perceptual, and psychomotor composites hold for both the new and old GATB forms. These results taken in conjunction with the high disattenuated correlations between the new and old cognitive and perceptual composites (.97 and .94, respectively; cf. Table 2-21) suggest that the dimensions measured by the new and old job-family composites (which are linear combinations of C_{kfm}, C_{gvn}, and C_{spq}) will also be very highly correlated.

Table 2-22
Correlations Among Cognitive, Perceptual, and Psychomotor Composites

Composite	$C_{gvn}(A)$	$C_{spq}(A)$	$C_{gvn}(F)$	$C_{spq}(F)$	$C_{kfm}(A)$
$C_{gvn}(A)$	1.0				
$C_{spq}(A)$	.73	1.0			
$C_{gvn}(F)$	.89	.61	1.0		
$C_{spq}(F)$	.71	.84	.69	1.0	
$C_{kfm}(A)$	.27	.44	.23	.42	1.0

Subgroup Comparisons

Although equipercentile equating matches subtest distributions for the total sample, it does not guarantee a match for distributions of subgroups within the sample. This is because the new and old versions of the paper-and-pencil GATB are not strictly parallel. Differences between the versions in measurement precision, as well as the various revisions made to the new forms, could cause group differences. To examine the impact of the revisions on subgroup distributions, two sets of analyses were performed: (a) analyses examining the level of performance of each subgroup across the old and new forms, and (b) analyses examining adverse impact for selected subgroups (conducted separately for the old and new forms).

Subgroup Performance Across Forms. For this first set of analyses, four subgroups were examined: (a) African Americans, (b) Hispanics, (c) Females, and (d) examinees 41 years of age or older. The equating transformation based on the total sample was applied to subgroup members who had taken the new GATB forms (E and F). For each subgroup, mean performance levels were compared across new and old forms. Six subtest variables were examined: $S_{ar}^{(g)}$, $S_{vo}^{(g)}$, $S_{3d}^{(g)}$, $S_{co}^{(n)}$, $S_{nc}^{(q)}$, and A_p. These variables are monotonic functions of the six subtest scores. Note that A_p is a function of both Form Matching and Object Matching for GATB Form A.)

The analyses involved conducting an ANOVA across the six subtest variables for each of the four subgroups. For each ANOVA, the significance of the difference among means was examined for Forms A, E, and F. Among the 24 comparisons, only one was significant at the .01 level: means across forms for African Americans on the Vocabulary subtest (Svo(g)), with higher scores evidenced on Forms E and F than on A. In general, the results indicate similar average performance levels across new and old versions for each of the four subgroups examined.

Adverse Impact. Adverse impact analyses were conducted for three subgroups: (a) African Americans, (b) Hispanics, and (c) females. The results of these analyses are given in Table 2-23. Differences in mean levels between majority and minority groups are reported for old and new GATB forms on the six subtest variables.

[7] The hypothesis that these pairs of correlations were significantly different was tested using a confirmatory factor model where Φ was set equal to the correlation matrix of the observed variables, and by constraining
$\varphi_{51} = \varphi_{53}$; $\varphi_{52} = \varphi_{54}$. Based on a chi-square difference test, these pairs of correlations did not differ significantly ($\chi^2 = 4.07$, df = 2, $p = .13$).

Each numerical entry in Table 2-23 is an effect size, calculated as

$$\Delta = \frac{(m - M)}{s_T} \quad , \qquad \textbf{(10)}$$

where m is the minority group mean, M is the majority group mean, and s_T is the total group standard deviation. The individual sample statistics used to compute the adverse impact values are given in Appendix A of Segall and Monzon (1995). The first value in the table (-0.99) indicates that African Americans scored 0.99 standard deviation unit lower than whites on the Arithmetic Reasoning subtest from Form A.

Table 2-23
Adverse Impact Statistics Across GATB Forms for Selected Subgroups

		Effect Size (Δ)		
Subgroup	**Score**	**Form A**	**Form E**	**Form F**
African American	$S_{ar}^{(g)}$	-0.99	-0.92	-0.89
	$S_{vo}^{(g)}$	-1.02	-0.87	-0.89
	$S_{3d}^{(g)}$	-0.84	-0.77	-0.79
	$S_{co}^{(n)}$	-0.70	-0.65	-0.61
	$S_{nc}^{(q)}$	-0.70	-0.52	-0.52
	A_p	-0.59	-0.59	-0.57
Hispanics	$S_{ar}^{(g)}$	-0.65	-0.66	-0.51
	$S_{vo}^{(g)}$	-0.82	-0.76	-0.66
	$S_{3d}^{(g)}$	-0.29	-0.33	-0.23
	$S_{co}^{(n)}$	-0.49	-0.50	-0.38
	$S_{nc}^{(q)}$	-0.55	-0.35	-0.31
	A_p	-0.21	-0.22	-0.13
Females	$S_{ar}^{(g)}$	-0.06	-0.07	-0.03
	$S_{vo}^{(g)}$	0.17	-0.01	0.00
	$S_{3d}^{(g)}$	-0.19	-0.17	-0.15
	$S_{co}^{(n)}$	0.15	0.19	0.19
	$S_{nc}^{(q)}$	0.43	0.44	0.45
	A_p	0.26	0.30	0.33

In general, levels of adverse impact tend to be similar across the old and new forms—although some minor trends are evident. For example, GATB Forms E and F tend to display slightly lower levels of adverse impact for African Americans. The adverse impact statistics for Hispanics tend to possess greater variability than for other minority subgroups, a result likely attributable to the small samples ($N \approx 325$) for this subgroup.

Summary of Form Development Activities

This portion of the chapter described the development of GATB Forms E and F (see Table 2-24). The major issues were fairness, speededness, and test security. To address these issues, five steps or phases were implemented. This chapter described each phase: item development procedures, item reduction and test format modifications, item pretest and analysis, test tryout and analysis, and final forms preparation.

Table 2-24
GATB Forms E and F: Test Lengths and Time Limits

GATB Test	Number of Test Items	Time Limit (minutes)
Arithmetic Reasoning	18	20
Vocabulary	19	8
Three-Dimensional Space	20	8
Computation	40	6
Name Comparison	90	6
Object Matching	42	5

Development and Review of New Items. This phase included the development, review, and revision of test items. The ARDP developed measures of item difficulty for all tests except Form Matching, which was comprised of two blocks of items, to arrange items within each test. A structured item review procedure was developed and implemented in three ARDCs with seven panel members to evaluate item bias. Items were then revised based on the comments of the panel members and further analyses.

Changes to Specifications for Test Length and Format and to Supporting Materials. Item reduction was addressed in two ways: (1) ARDP staff examined technical and operational issues, and (2) AIR conducted an empirical study that addressed issues related to item reduction. Recommendations for item reduction based primarily on technical issues were made by ARDP staff. The results of the AIR study led to recommendations for item reduction, test administration time limits, whether a test is speeded or nonspeeded, and test order. The impact of speededness was further reduced by introduction of formula scoring for the speeded tests.

Instructions to the examinees were changed to reflect differences in scoring methods and made much more explicit. Test format modifications were addressed by the Test Aesthetics Project, which included interviews, literature searches, focus groups, and a survey. Specific recommendations were made for the format and content of the GATB administration manual, test booklets and instructions, test items, and answer sheet.

Item Tryout and Statistical Screening. The item pretest and analysis had two goals: (1) conducting an item analysis to obtain preliminary difficulty and discrimination indices, and (2) obtaining a quantitative estimate of ethnic and gender performance differences for each item. The sample comprised 9,327 applicants from USES local offices in the five geographic regions represented within the ARDP. Data were obtained by administering to the sample members 16 test booklets comprising three speeded tests and one power test. Each sample member completed one test booklet. Classical test theory item analyses were performed for the speeded test items. Item selection criteria included difficulty, discrimination, and content considerations. IRT procedures were used for the power test items. The analyses included dimensionality, position effects, item and test fairness, and test information graphs. Item DIF analyses were also performed with Mantel-Haenszel procedures. IRT procedures were used for test-level DIF analyses.

Construction of the Final Version of Forms E and F. After items were screened and calibrated, a final set of items was selected for each Form E and Form F test. Items were selected to yield forms as parallel to each other as possible with respect to content coverage, difficulty, and test information. The forms were also balanced on subgroup difference statistics so that no one form provided any relative disadvantage to females, African Americans, or Hispanics. Insofar as possible, the power tests were also designed to be similar with respect to difficulty and information to Form A, after adjusting for differences in test lengths.

Form Linking Study. Equating of the new GATB forms to the old forms proved successful. Despite the changes made in the new test forms, the evidence suggests that there is sufficient similarity to obviate the need for separate composite equating tables for the non-psychomotor composites. Average subgroup performance levels are similar across the old and new forms, and reliabilities of the new GATB forms are generally as high as or higher than those of the old forms. Construct validity analyses of the old and new forms suggest that the GATB validity data can continue to be used for the new forms.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Baker, F.B., Al-Karni, A., and Al-Dosary, I.M. (1991). EQUATE: A computer program for the test characteristic curve method of IRT equating. *Applied Psychological Measurement, 15,* 78.

Berk, R.A. (1982). *Handbook of methods for detecting test bias*. Baltimore, MD: The Johns Hopkins University Press.

Boldt, R.F. (1983). *Review for perceived bias on ASVAB Forms 11, 12, and 13* (Report No. ETS-RM-83-4). Princeton, NJ: Educational Testing Service.

California Test Development Field Center. (1991). *Proposal to reduce the number of alternatives in GATB Form Matching items*. Unpublished manuscript.

California Test Development Field Center. (1992). *Operational considerations for reducing the number of items in the General Aptitude Test Battery*. Unpublished manuscript.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences (2nd Ed.)*. Hilllsdale, New Jersey: Erlbaum.

Corel Corporation (1993). CorelDRAW! user’s manual – version 4.0. Ontario, Canada: Author.

Daggett, M.L. (1995). *Test aesthetics improvement project*. Unpublished manuscript.

Dale, E., & O’Rourke, J. (1981). *The living word vocabulary*. Chicago: Worldbook-Childcraft International, Inc.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19-29.

Hambleton, R., & Rogers, J. (1993). *MH: A FORTRAN 77 program to compute the Mantel-Haenszel statistic for detecting differential item functioning*. Author: University of Massachusetts-Amherst.

Hambleton, R.K., & Rogers, H.J. (1988). *Design of an item bias review form: Issues and questions*. Albany, NY: New York State Education Department.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Norwell, MA: Kluwer Academic Publishers Group.

Harms, R.A. (1978, March). *The development, validation, and application of an external criterion measure of achievement test bias*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Toronto, Ontario, Canada.

Hartigan, J.A. & Wigdor, A.K. (Eds.) (1989). *Fairness in employment testing: Validity generalization, minority issues, and the General Aptitude Test Battery*. Washington, DC: National Academy Press.

Holland, P. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test Validity*. Hillsdale, NJ: Erlbaum Associates.

Hotelling, H. (1931). The generalization of Student's ratio. *Annals of Mathematical Statistics, 2,* 360-378.

HRStrategies. (1994). *Revised technical documentation for the General Aptitude Test Battery*. Grosse Pointe, MI: Author.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory: Application to psychological measurement*. Homewood, IL: Dow Jones-Irwin.

Lockheed-Katz, M. (1974). *Sex bias in educational testing: A sociologist's perspective* (Report No. ETS-RM-74-13). Princeton, NJ: Educational Testing Serivce.

Lord, F.M. (1977). Practical applications of item characteristic curve theory. *Journal of Educational Measurement, 14*, 117-138.

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

Madaus, G., Airasian, P., Hambleton, R., Consalvo, R.W., & Orlandi, L.R. (1979). *Development and application of criteria for screening commercial standardized tests for the Massachusetts Basic Skills Improvement Policy*. Boston: Massachusetts State Department of Education.

McCloy, R.A., Russell, T.R., Brown, K.M., DiFazio, A.S., & Green, B.F. (1994). *Item selection for the speeded subtests of the General Aptitude Test Battery (GATB), Forms E and F* (HumRRO Final Report FR-PRD-94-10). Alexandria, VA: Human Resources Research Organization.

Mellon, S.J., Daggett, M., MacManus, V., & Moritsch, B. (1996). *Technical report for the development of GATB Forms E and F*. Sacramento, CA: Author.

Olson, A., & Smoyer, S. (1988, April). *Developing quality science programs*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

PARDC. (February, 1995). *Test specifications for GATB Forms E and F.* Unpublished manuscript. Sacramento, CA: Author.

Peterson, N.G. (1993). *Review of issues associated with speededness of GATB tests*. Washington, DC: American Institutes for Research.

Raju, N.S., Drasgow, F.D., & Slinde, J.A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel-Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement, 53*, 301-314.

Sager, C.E., Peterson, N.G., & Oppler, S.H. (1994). *An examination of the speededness of the General Aptitude Test Battery power tests*. Washington, DC: American Institutes for Research.

Schratz, M.K., & Wellens, B. (1981, August). *Minority panel review in the development of an achievement test*. Paper presented at the Annual Meeting of the American Psychological Association, Los Angeles, CA.

Scientific Software. (1990). *BILOG 3: Item analysis and test scoring with binary logistic models*. Chicago, IL: Author.

Segall, D.O., & Monzon, R.I. (1995). *Draft report: Equating Forms E and F of the P&P-GATB*. San Diego, CA: Navy Personnel Research and Development Center.

Southern Test Development Field Center. (1992). *Reduce the number of items on the General Aptitude Test Battery: Recommended number of items for GATB forms A,B,C, and D*. Unpublished manuscript.

Stocking, M.L. & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, 7, 201-210.

Tittle, C.K. (1982). Use of judgmental methods in item bias studies. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias*. Baltimore, MD: The Johns Hopkins University Press.

Wilson, D.T., Wood, R., & Gibbons, R. (1991). *TESTFACT: Test scoring, item statistics, and item factor analysis*. Chicago, IL: Scientific Software, Inc.