Review of Issues Associated with Speededness of GATB Tests

Norman G. Peterson

December 6, 1993

American Institutes for Research
3333 K Street NW, Suite 300
Washington, DC 20007

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.

Acknowledgments

I thank Jeffrey Bell for assistance in conducting the literature search and finding and organizing the reports. Chris Sager, George Wheaton, Scott Oppler, and Teresa Russell read portions of this report and greatly improved it. I am solely responsible, however, for the entirety of this document.

Table of Contents

Review of Issues Associated with Speededness of GATB Tests

Introduction

The purpose of this literature review was the collection and integration of research findings and opinion on several issues related to the degree of speededness of four tests on the General Aptitude Test Battery (GATB) (USES, 1970). These tests--Computation, Three-Dimensional Space, Vocabulary, and Arithmetic Reasoning--purport to measure psychological constructs generally measured with power tests, but the GATB tests appear to be more highly speeded than tests in other, modern multi-aptitude batteries, according to the criterion of number of items completed by 90% of examinees (Hartigan & Wigdor, 1989, p. 108). (Note, definitions of speed and power vary, see discussion below). The impetus for the review primarily arises from concerns raised by the National Research Council's Committee on the GATB (Hartigan & Wigdor, 1989). Their concerns can be summarized as three major points:

- the meaning of constructs measured by speeded power tests may be different from the meaning conventionally attached to those constructs
- the speed component of the tests may cause the tests to be differentially valid for different racial or ethnic groups
- "the severe time limits of the GATB subtests might produce an adverse psychological reaction in examinees as they progress through the examination and might thereby reduce the construct validity of the subtests" (p. 106).

We searched the literature to identify articles and reports that might address several specific issues that are related to these concerns, especially as they pertain to the four constructs measured by the GATB tests:

- the extent to which speeded and power tests containing similar items measure similar constructs and the implications of the findings for maintaining construct linkages with the GATB validity base should the GATB tests be made less speeded
- the relative merits of speeded power tests, including the degree to which optimal speededness is influenced by the population tested, the construct being measured, and the purposes of testing
- the extent to which test speededness may result in differential validity for members of different ethnic or racial subgroups

In completing the review we attempted to focus on research and opinion relevant to the adult population and tests used for employment or educational screening purposes, since those are most appropriate for the past and future contemplated uses of the GATB as an employment screening test. We relied on reviews of earlier literature, with some exceptions, and examined original articles for the more recent research and opinion.

Methods of Assessing Speededness of Tests

Claims and conjectures about the influence of speed on the construct, predictive, or differential validity of tests must rest on some assessment of the speededness of a given test. The traditional definition of a pure power test is one for which all examinees have time to consider and attempt all items and is usually thought of as providing scores uninfluenced by rate of work. Such tests generally have items with a range of difficulties, often arranged in order by difficulty (i.e., more difficult items appear later in the test). A pure speed test is one for which no examinees have time to consider and attempt all items and all items are so easy that the number of items attempted is equal to the number correct. In practice, very few tests fall into the pure power or speed category. They instead are partially speeded, sometimes for theoretical or construct validity reasons, but perhaps more often for administrative efficiency. Given that a test is not a pure power or speed test, some assessment must be made of the degree of speededness of the test. Rindler (1979) notes that investigations using speededness as an important variable have been hampered by "the murkiness of the theoretical literature on test speededness and by the resulting inadequacy of currently applied measures of speed (p. 262)."

Single Test Administration Methods

She provides an excellent review of methods that have been proposed as assessments of the the speededness of tests, dividing these methods into two primary groups, those estimating speededness based on single test administrations and those requiring multiple test administration. The primary single group methods include:

- Gulliksen's (1950) ratio of standard deviations of items unattempted to items unattempted plus items incorrectly answered or omitted after consideration (the standard deviation of items unattempted with equal 0 when the test is completely unspeeded and the ratio will be zero, but when the test is completed speeded, the standard deviation of items incorrectly answered will be 0 and the ratio will be 1.0),

- the Educational Testing Service's (ETS) "rule of thumb" approaches (Donlon, 1973). That is, a test is speeded if less than 100% of examinees reach 75% of the items and less than 80% of the examinees finish the test. (Later ETS publications [Donlon, 1979] mention a third measure, the number of items completed by 80% of the examinees, and Donlon [1975] also combines these indices into a graphical presentation of percentage of examinees still working against percentage of test completed.)

- Donlon's (1973) "time-needed estimates" (also expanded upon in later work [Donlon, 1975, 1979] in an attempt to overcome some of the flaws in the single-administration methods),

- Stafford's (1971) speededness quotient.

Even simpler methods, not mentioned by Rindler, are sometimes used to assess the degree of speededness of a test from a single administration such as the percentage of items attempted that are answered correctly (Swarthout, 1988; Peterson, Russell, Hallam, Hough, Owens-Kurtz, Gialluca, & Kerwin, 1990).

She points out serious flaws in the single test methods--flaws which are due to the estimation of speed as a function of the number of unattempted items. These single-group methods make three assumptions in practice, she maintains, all of which are problematic:

a) all of the unattempted items are found at the end of the test;

b) all consecutive omissions that appear at the end of the test are unattempted (rather than considered and omitted) items;

c) unattempted items are the only elements of test performance that reflect speededness (Rindler, 1979, p. 265).

Rindler points out that the first assumption is valid only if examinees begin with the first item and work consecutively to the end, that the second assumption miscategorizes an unknown number of considered-and-omitted items that appear as part of the string of omissions at the end of the test (this is ameliorated but not eliminated with correction-for-guessing methods), and that the third assumption is theoretically incorrect. That is, test instructions that ask examinees to "work quickly and accurately" or the like urge particular and differential paces of responding, speeding each attempted response to some degree. In summary, she cites three sources of systematic error in single administration methods:

1) The failure to account for the number of items subjects skip over and have insufficient time to return to consider.

2) The failure to account for the number of items subjects do not read but, rather, fill in at random as time approaches its limit.

3) The failure to account for effects of cognizance of time limits on overall rates of response and consequent accuracy (p. 266).

Multiple Test Administration Methods

The methods requiring multiple test administrations stem from the concept that a test, practically speaking, is unspeeded if order of examinees remains unchanged when the test is given under speeded and unspeeded conditions. Cronbach and Warrington (1951) proposed *tau* which is consistent with this concept. *Tau* is the square of the correlation between the scores on the speeded test and an untimed administration of a parallel form of the speeded test, corrected for attenuation due to unreliability of the tests. "Essentially, *tau* allows determination of the proportion of the reliable time limit score variance reflecting common factors in the no-time-limit parallel test. It clearly depends neither upon proper identification of unattempted items nor upon unattempted items as the sole representatives of speededness in test performance (Rindler, [1979], p. 266)." Just as clearly, these methods are impractical for routine use, but are certainly attainable in research settings. Donlon (1980, p. 15) summarizes the differences between the Gulliksen (single-administration) and the Cronbach & Warrington (multiple-administration) methods as follows:

> The Gulliksen approach basically asks is there any evidence of being affected by speed? The Cronbach and Warrington approach basically asks whether, given that some records are affected by speed, there is evidence that things would be different, in the sense of a ranking of candidates, under conditions of power. Of these two aspects, the Cronbach and Warrington concern is the one most frequently overlooked in practice. This is largely due to the difficulty of obtaining an adequate estimate. In a sense, the Cronbach and Warrington question and the Gulliksen question are two different but equally important aspects of the concern for speed.

Other Methods

Rindler notes that a number of other methods of assessing test speededness involve multiple scorings of the same test performance using a procedure in which examinees take the test under normal time limits, then finish the test in unlimited time. Usually three scores are calculated: (1) the number correct in the normal time limit, (2) the number correct on the completed test, and (3) the time it took to complete the test (Baxter, 1941; Davidson & Carroll, 1945). These scores are then subjected to correlational and regression analyses. Other methods include the administration of the same test (rather than parallel forms) under timed and untimed conditions to the same group; splitting a test in half and administering one half in a timed condition and one half in an untimed condition; timing the responses to each item or to the entire test when given in an untimed administration; and limiting time of exposure to items through the use of audio-visual equipment (or, today, through computer administration). Rindler notes that computing correlations between timed and untimed administrations of a test (or test halves) is not very meaningful unless there is a basis of comparison such as the expected correlation of the tests administered under the same time limit, and that techniques using substantially different administration materials (audi-visual or computer administration) have limited generalizability to the typical testing conditions.

Summary

This discussion of measures of speededness sets the background for reviewing the research relevant to other issues since the measures of speededness limit the interpretations that can be made of findings. For the GATB, it is the Cronbach and Warrington type of measure that seems most relevant for an evaluation of the extent to which different constructs are being measured under speeded and power conditions, whereas comparison of the Gulliksen types of measures seems more appropriate for identifying the impact of speededness on score differences between subgroups. As noted by Rindler, it would be very useful to conduct research that compares the single-administration indices to the multiple-administration indices.

Before leaving this topic, we note that Item Response Theory techniques have begun to be applied to the problem of measuring the impact of speed on item and test performance (Bejar, 1985; Davey, 1990; Douglass, 1981). These methods appear to hold some promise, but they require further investigation before reliance can be placed on their routine use.

Relative Merits of Power Tests and Speeded Power Tests

Should tests intended to measure a "power" construct ever be speeded? If so, under what conditions should this occur? "Modern" general ability tests (modern approximately defined as the period after World War II) favor the use of tests with items arranged in order of item difficulty and with time limits that "are sufficiently generous that nearly everyone finishes all the items on which he is likely to succeed (Cronbach, 1990, p. 258)". Cronbach notes that earlier tests of general ability tended to be highly speeded. There is relatively little direct discussion in the literature about the specific reasons for this change in approach to testing, but the previously cited concerns of the National Research Council (imposition of time limits on power tests changes the construct measured, differential validity of speed component for race/ethnic groups, and adverse psychological reactions to speed) seem plausible reasons for the change. Wise (personal communication, October 21, 1993) mentions several reasons for preferring non-speeded power tests:

- examinees are generally anxious about taking tests and speeded tests are particularly anxiety producing; consideration of examinee comfort level argues for non-speeded tests

- speeded power test scores will be more highly correlated with speeded test scores (than with non-speeded power test scores), resulting in a narrower coverage of the predictor domain for batteries with such tests, which is likely to result in lower validities, especially if the criterion domain of interest is broad

- speeded tests are more susceptible to coaching even if a correction for guessing is used because time allocation strategies (e.g., decisions of examinees regarding how much time to spend on any given item, whether to skip troublesome items) are so important on speeded power tests; under the completely power test extreme, time allocation strategies are irrelevant

Ree (personal communication, September 1, 1993) noted that scores on speeded tests are much more sensitive than are scores on power tests to changes in answer sheets, or, more generally, to alterations in administration conditions, as evidenced by the effects on speeded test scores of the difference in the answer sheets (from those used in operational testing) used in the Armed Services Vocational Aptitude Battery (ASVAB) norming study by the National Opinion Research Center (Ree & Wegner, 1990). This point of view is echoed by Wesman (1960), Wise (personal communication, August 30, 1993), and Hartigan & Wigdor (1989).

Construct Validity Considerations

Wesman (1960) warns against considering speed to be a unitary trait rather than a dimension upon which tests can be made to vary. He provides several illustrations that he contends show speed is a quite different phenomenon from one setting to another. The thrust of his argument is that the type of speed called upon by the test should "match" the type of speed called for on the job or in the classroom, if the test is intended to predict job or classroom performance.

Ackerman & Humphreys (1990) discuss the role of speed in the context of construct validity of a test and the appropriate choice of the dependent variable, i.e., the necessity of choosing to emphasize speed or accuracy in the construction, administration, and method of scoring a test. As have others, they point out that scores on many modern tests confound speed and accuracy (as the GATB speeded power tests do). Indeed, they (Ackerman & Humphreys, 1990, p. 246) point out that:

> In some situations, two otherwise identical tests may also measure different constructs *when one of the tests uses simple items that allow for correct responses but long latencies from low-ability subjects, and another uses a power test format or some similar design with wide-ranging difficulty levels* [italics added]. Empirical studies have yielded general support to this claim (e.g., Lohman, 1979; Lord, 1956).

Regarding the GATB tests of concern, this statement would seem to argue that making the tests less speeded would markedly change the construct measured. I would argue that this might only be partially true because the GATB tests do contain items of varying difficulty, ordered by increasing difficulty (Swarthout, 1988). Note that the same concern is relevant in instances in which tests are changed to contain less difficult items administered under the same or more speeded conditions.

Ree (personal communication, September 1, 1993) offered the opinion that one can certainly alter the speededness of a measure, but this change will certainly alter what one is measuring, reinforcing the views expressed by Ackerman and Humphreys. He suggested that research be conducted to compare performance on timed and untimed administrations of the GATB measures in order to gain as precise an understanding as possible of the effect of the changes.

Other discussions of the "proper" role of speed revolve around its contribution to the measurement of general ability or *g* (Jensen, 1983; Cronbach, 1990) and how important it is perceived to be in the measurement of g. While these discussions are stimulating and important to theories of mental ability, their pertinence to the GATB tests centers on the likely effect of altering the contribution of speed to the GATB scores. According to Jensen, laboratory measures of speed (usually speed of reaction) correlate about .5 (uncorrected for restriction of range) with measures of g. Ree (personal communication, September 1, 1993) offers the opinion that, all else equal, making a test less speeded should reduce its loading on g, and, therefore, its expected validity for predicting job performance across a wide variety of jobs. He cites his research in the Air Force that showed little improvement in prediction of training performance by the addition of specific abilities to general ability (Ree & Earles, 1991b). As just mentioned, however, opinions differ over the centrality of speed to the measurement of "g", and reasonable persons can differ over the importance of the increment to criterion-related validity obtained by adding specific ability to general ability.

Criterion-related Validity

There is no need to invoke *g* to make the argument that reducing speed for a given test will result in changing the construct measured, however; it is sufficient only that there be a dominant dimension other than speed that is the construct being measured. The two versions of the test will then have differing loadings on the "speed" factor, even if their loading on the dominant dimension remains the same--thus the tests are no longer "parallel." (See Ackerman & Humphreys, 1990, p. 235.) However, if the loading on the dominant dimension remains unchanged or nearly so, it may be that the validity (for predicting job performance) would remain essentially the same as well.

Kendall (1964) has conducted a somewhat similar analysis, but with an emphasis on the effects of speed on predicting a criterion:

> For tests containing multiple-choice items ordered as to difficulty, the general proposal is made that adjustment of the time limit, for a fixed amount of material, may change the nature of the test. For example, with very long time limits power scores may be obtained; with very short time limits speed scores may be obtained, especially when easy items appear at the first of a test. Intermediate time-limit scores reflect some composite of speed and power variance (e.g., Davidson and Carroll, 1945; Baxter, 1941). Thus, adjustment of time limits may alter the relative weightings of speed and power factors, changing the factorial composition of predictor scores. If the criterion measure [that the test is intended to predict] contains both speed and power variance, then there may be a unique predictor time limit for which the relative weightings of speed and power in the predictor scores most closely approximate the optimal weightings required for maximum validity (p. 790).

Given this analysis, substantially changing the time limit for tests like the GATB would lead one to expect a change in the validity coefficients, given no matching change in the speed/power influences in the criterion predicted. The GATB has been used to predict job performance for so many jobs, however, that it is extremely difficult to decipher the probable overall effect on its criterion-related validity, even given the credibility of this analysis.

Kendall, in the same article, investigated the validity of varying time limits for a heterogeneous, multiple-choice test with items ordered by difficulty for a criterion of total scores on the same test given under full-time conditions, from a separate administration. Six different time limits (ranging from 15 minutes to 30 minutes) were used in a within-subjects design (Canadian military recruits), and correlations with the full-time limit (which was about one hour) score were calculated. His analyses showed that validities did differ across time limits with the maximal validity occurring for the 25-minute limit, though the validities were very similar for all but the shortest time limit, which was the lowest.

Furthermore, Kendall reported, regarding the use of tests of unequal difficulty for the same samples, that "it was clearly established that the maximal [validity] time limits were longer for the easier tests. This result corresponds to using a longer maximal time limit for subjects of higher ability (Kendall, 1962). If such a conclusion is substantiated in further research the practical result would be an increased flexibility in the adaptation of a given test to different conditions. Short time limits might be best when using the test with less able groups, longer time limits might prove best when using the same test with more able groups (p. 798)." It should be kept in mind that Kendall was using a test with heterogeneous item content, but his finding also concords with Ree's opinion that all tests are speeded for persons of low ability. Boese (personal communication, August 31, 1993), citing specific analyses of GATB validities as moderated by level of education (Boese, 1993), offers the opinion that GATB tests might be made less difficult and administered under the same or increased conditions of speededness in order to increase their validity for less-educated examinees. He also agreed with the author that this might act to decrease the validity for the more-educated examinees. Both of these conclusions seem in accord with Kendall's reasoning and conclusions.

Bell and Lumsden (1980) carried out related research on test length and validity. Their sample consisted of Australian 12th-graders; the tests were (a) the Australian Scholastic Aptitude Test (a 100-item four-choice objective test, which purports to measure abilities relevant to the study of humanities, social sciences, mathematics, and science) and (b) a 40- item, multiple-choice test in economics and reading comprehension. They successively removed items according to the strategy by which the test was constructed, i.e., they deleted items with the lowest point-biserial correlations with the total test score. Validity coefficients for predicting an appropriate external criterion (student grades or scores on achievement tests) were recomputed for each successively shorter version of the test. Across four samples, they found that much shorter length tests (down to about 40% of full test length) could provide the same or very nearly the same levels of validity. There are some problems with their methodology, notably a lack of replication of the findings (but their N's are respectably large enough to expect stability of the results) and the assumption that the same constructs would be measured if the tests were actually to be administered in their shortened versions. These results do seem to indicate that power tests (they do not report that the tests were administered under power conditions, but I am assuming this to be the case since they were using point-biserial correlations to shorten their tests and the test content appears to be of the sort that would be measured in a power setting) might be shortened

considerably without sacrificing too much validity. The extent to which these results generalize to administering GATB tests under power conditions is open to question because of the different samples and purposes of testing and, importantly, because the GATB power tests are speeded such that few persons attempt items near the end of the tests--which implies that shortening the tests would have relatively little effect on validity. The results do seem to indicate that giving shorter versions of the GATB tests under power conditions (because of constraints on available administration time) might be possible without substantial sacrifices in predictive validity.

Testing Disabled Persons

Nester (1993) discusses issues associated with testing handicapped persons. She notes that many people with disabilities will require extra testing time, beyond that provided in standard test administrations. This creates a problem of determining the appropriate time for testing those with a variety of disabilities, but this problem is largely manageable if the test of concern is not speeded. When power tests are given under partially speeded conditions, however, there can be severe problems. When unlimited time on such tests is provided to persons with disabilities, the resulting scores may not be interpreted in the same way as scores achieved under standard times. ETS researchers discovered this effect when they investigated the validity of the SAT for predicting freshman grades (Willingham, Ragosta, Bennet, Braun, Rock, and Power, 1988). Ideally, Nester notes, speeded power tests would be eliminated. She states (p. 80), "Anyone who is in a position to make policy with respect to test time limits should argue for the use of very liberal time limits, with a completion rate of 90-95% of all test-takers. For such tests, disabled test-takers could be given unlimited time without having an undue advantage."

Summary

Almost all experts agree that scores on speeded power tests are influenced by both a speed component and at least one other component, presumably the dominant dimension or construct the test was intended to measure. The relative influence of these two components on test scores appears to vary with the degree of speededness, the type and difficulty of the items, the ability of the examinees, and other factors in the administration situation (e.g., instructions to examinees). There do not appear to be any strictly analytic methods of determining the influences of these components on a particular test; research must be conducted to determine the relative influences of speed on scores on a particular test at particular time limits. If the dominant dimension of the test is considered "g," then some authors would expect the validity of the test for predicting job performance to decrease if the influence of speed on the scores is decreased. Others argue that increases and decreases in validity are determined by the match between the speed and power components of the test and the criterion to be predicted, which seems to imply a change in validity if the speed/power balance of the test is changed but the criterion measures remain unchanged. Whether this change will be positive or negative is difficult to discern in general, but especially for the GATB. Such a determination would seem to involve detailed analyses of the speed/power contributions to scores on the job performance criteria and the tests involved

in each situation. For a battery like the GATB, used for hundreds of jobs, such analyses of job performance criteria seem infeasible. Most GATB validity studies have used rating scales as criteria, and the sensitivity of this methodology to speed/power differentials in job performance is unknown.

On more practical grounds, scores on speeded tests are noted as being much more susceptible to deviations from standard testing conditions, or even planned changes such as modifications in the shape of answer bubbles. This concern generalizes to all administration conditions, such as adherence to time limits. As Nester noted, speeded power tests create more problems for adapting test administration to the disabled populations than do full power tests.

Given that the GATB contains several intentionally speeded tests, some argue that reducing the speededness of the GATB power tests results in broader coverage of the ability domain, therefore increasing the potential for criterion-related validity. This contention seems correct on its face, but the amount of increase in coverage is unknown and does not seem to be a strong argument by itself for changing the speededness of the GATB power tests.

Theoretical and practical arguments for changing the speededness of tests of power constructs seem to favor a reduction in speededness, given that the purpose for using the tests are still fulfilled. I am primarily concerned here with the purpose of predicting job performance, rather than counseling, and, based on the information reviewed in this section, no certain answer can be given about the effect on criterion-related validity of reducing the speededness of the GATB power tests. Such a determination is complicated by the fact that the GATB subtests have extremely short time limits, for important practical reasons, and administration of the tests under power conditions within these time limits might severely reduce the reliability of the tests. If the reliability is substantially reduced, then validity must be reduced. In such a case, time limits would have to be increased if true power administration were to be obtained without sacrificing reliability, and, therefore, validity. Empirical data are required to make this evaluation and judgment.

Relationships Between Speeded and Power Tests of Similar Items

General Evidence

Much of the early work on the relationship between scores on tests of intelligence obtained under speeded and power conditions led to the erroneous conclusion that speed and power scores were very highly correlated, often in the .90's. Baxter (1941) and Davidson and Carroll (1945) were among the first to point out that these conclusions, for the most part, were based on part-whole correlations (usually the "unlimited time" score was the time-limit score plus the additional score achieved by the examinees in the allowed extra time). They used different and more sophisticated methodologies to investigate the relationships between these two types of scores and their relationships to other scores. Later researchers attempted to use these methods or improvements of them, so their work deserves fairly close examination.

Baxter, using a single-group administration method for 100 college students, obtained scores for number of items correct in the normal time limit on the Otis Self-Administering Test (an omnibus test of intelligence), number of items correct in unlimited time, and number of seconds required to complete the entire test. (Examinees were not allowed to go back and work on items attempted during the normal time limit.) Somewhat confusingly for modern readers, he labeled the first of these scores the "power" score, the second score the "level" score, and the third score the "speed" score. These scores were correlated with each other and with other test scores and school grades. His "power" score correlated .62 with the "level" score (this correlation is a part-whole correlation) and .75 with "speed" (this correlation is also somewhat inflated because of the testing procedure). "Speed" and "level" correlated -.06. The multiple correlation of "speed" and "level" with "power" was .998, and the contributions of speed and level to the power score were about 60% and 40%, respectively. Baxter reported separate correlations of the three scores with grades (honor point ratio), a vocabulary test given prior to college entry, and the Army Alpha test. The pattern of correlations was similar for grades and the vocabulary test, with speed correlating .28 and .18, respectively, while the level and power scores correlated .43 to .49 with these criteria. For the Alpha test, which was given under a shorter than normal time limit, the pattern was a bit different--.52 and .05 for the speed and level, but .68 for power. Multiple correlations and beta weights were computed to identify the contributions of speed and level to the prediction of the various criteria. Some of the conclusions that Baxter reached on the basis of these data were: speed and level vary independently, speed and level account for the entire variance of power (the score attained in the normal time limit) with speed contributing slightly more than level, and prediction of external criteria (grades and other test scores) by using separate speed and level scores in multiple correlation is greater than that obtained by the normal, standard time score.

Baxter's results are limited in generalizability by the type of tests and criteria he used and the relatively homogeneous and small sample he employed, but his methods show that a careful and informative analysis is possible within a single-group administration <u>if</u> one can obtain unlimited time scores. Results from these kinds of methods, however, are subject to the problems of interpretation that Rindler identified and were discussed above.

Davidson and Carroll (1945) used procedures similar to those used by Baxter to obtain speed, level, and power scores on a battery of tests (item types included addition, arithmetic reasoning, same-opposites, common sense, disarranged sentences, number series, verbal analogies, directions, disarranged morphemes, and letter grouping) administered to 91 college students. However, unlike Baxter, they did allow examinees to go back and work on previously attempted items after they had attempted each item once. Thus, their level score included items that may have been reattempted. Their speed score was identical to Baxter's. They factor analyzed scores, "Essentially, the procedure involved locating the time-limit variables [the "power" scores] in the factor space defined by the speed and level variables (p. 418)," to avoid spuriously inflating the factor loadings of the speed and level scores with the power scores. After extraction and rotation to simple structure, they presented a six-factor solution that contained four interpretable factors: numerical speed, level of reasoning, speed of reasoning, and general speed. They also regressed level and speed scores on the power scores. The multiple correlations ranged from .711 (one of the arithmetic reasoning subtests) to .843 (common sense items). The beta weights for speed and level varied considerably across item types and appeared to be partially a function of the particular time limits set for the "power" score. They concluded:

> Factor analysis revealed that in all cases speed scores were linearly independent of level scores and that time-limit scores [speeded power scores] could be represented as factorially complex measures having loadings on both speed and level dimensions of ability. Of the factors which were identified several were similar to verbal, numerical, and reasoning factors isolated in previous studies. In the domain of reasoning ability both level and speed factors were identified. A general speed factor involving nearly all of the speed scores were found. It is concluded that because of their factorial complexity, time-limit scores should be used with considerable caution both in factorial studies and in studies involving the prediction of criteria (p. 426).

Although Davidson and Carroll's sample was relatively small and homogeneous, the empirical conclusions they reached are consistent with the theoretical statements and opinions earlier discussed in the section on the relative merits of power tests and speeded power tests, that is, that the relative influences of speed and level (power) on speeded power tests are determined by a multiplicity of factors, making it difficult to determine the effects of changes in time for administration on the construct validity and criterion-related validity of speeded power tests.

More recent studies confirm and expand the findings of Baxter and Davidson and Carroll. Powers, Swinton, and Carlson (1986) performed item-level factor analyses of the Graduate Record Examination (GRE) Aptitude Test on large samples of examinees (several thousand). The form of the GRE used in their study contained verbal analogies, opposites, sentence completion, reading comprehension, algebra, arithmetic, geometry, and data interpretation item types. It was administered as three, separately-timed tests of verbal ability, reading comprehension, and quantitative ability. The tests were administered under normal timed conditions; no speed manipulations were conducted. They used the results of the item factor analyses together with other item statistics (such as percent attempting) to interpret the influence of speed on the factor structure. They reported in summary, "Although distinct components of speed were found to be associated with each of the three sections of the operational forms, the component found in Section I (discrete verbal items) is problematic because of its relatively large contribution to the test's total common variance. Since the GRE Aptitude Test is purported to be primarily a power test, it is suggested that Section I be reexamined in light of its relative speededness--especially since speed and ability emerge here as uncorrelated traits (p. 50)." These remarks are similar to those made by Baxter and Davidson and Carroll, again reflecting the consistent finding of the relative independence of speed and level scores, and their differential impact, apparently dependent on item type and degree of speededness, on scores obtained from a speeded power test.

Kingston (1984), also using GRE data, examined the effects of several changes to GRE forms that were made in 1981. Of most interest here are the changes that were made to the verbal section. Fewer items were administered and more time was given to complete the items. In addition, instructions (and scoring) were changed from formula scoring (corrections for guessing) to number right-scoring. Kingston analyzed residuals obtained by predicting new GRE scores from a prediction equation that had been derived on a sample of examinees who had completed two of the older GRE forms. Variables in the prediction equation included old GRE scores and additional data (demographics, major, etc.). Thus, to the extent that the residuals were large, the new scores were being determined by variables not in the equation or because the variables were not weighted appropriately. The cross-validated multiple correlation for the prediction equation in the old-old group was .90. The multiple correlation for the equation for

predicting the old-new group was .89. Thus, the prediction did not "shrink" much across the change in time limit and instructions/scoring change, therefore, the prediction equation was adequate for use in analyzing the mean residuals by predicted GRE scores. Examination of the mean residuals plotted by predicted GRE verbal score showed that the residuals were positive for the lower half of the predicted score range and negative for the upper half of the score range. The mean residual was strongly linearly related to predicted score for most of the score range and the standard deviations of the residuals also were consistent throughout the score range. Kingston says, "All in all, the residual pattern for the predicted verbal scores indicates a small but consistent change in the sources of variability underlying the verbal measure (p. 4)."

Since GRE test items were ordered by difficulty, only the more able examinees were able to answer the last few difficult items. Thus, Kingston notes, the pattern of residual results could occur because the new test was less speeded, resulting in the more able examinees being more likely to have had their scores underpredicted. Alternatively, or in addition, more able examinees might be more "test-wise" or more readily grasp the test instructions and have a greater propensity than less able examinees to guess under number right-scoring conditions. This would also lead to the observed pattern of residuals.

Lin (1986) provides a concise summary of the conflicting results concerning the role of speed in the construct and predictive validity of familiar constructs, and he presents an investigation of the effects of speed on the construct and predictive validity of two measures (verbal ability as measured by sentence completion items and quantitative ability as measured by non-graphic items), the effects of time limits on examinees' time allocation and test performance, and a test of the classical assumption (made by Gulliksen) that examinees proceed in an orderly, sequential fashion through a test. He used a computer administered method and his sample consisted of 93 graduate students at Cornell University. Although his methods and sample are dissimilar to the current GATB and likely applicant population, his investigation was elegant, precise and focused on important issues associated with the GATB.

Lin administered two different test forms of each item type, each form administered under timed and untimed conditions--with each examinee receiving both item types in timed and untimed formats, but for different forms (counterbalanced across examinees). To investigate the effects of speed and level, he regressed the time administration score (number correct) on one form against the level (number correct) and speed (time to complete) scores in the untimed condition on the other form. He found that the multiple correlations were .49 and .69 for the two sentence completion forms, with near zero beta weights for speed (.035 and -.113) and high weights for level (.549 and .629). For quantitative items, he found similar levels of multiple correlations for the two forms, .49 and .73, but significant beta weights were found for both speed (-.341 and -.446) and level (.465 and .571).

Lin also investigated the effects of age, sex, and "feelings of being rushed" to see if they moderated these regressions, but his sample sizes did not afford him much power and those results are not discussed here.

He compared the correlations of the timed and untimed tests with scores on the GRE verbal and quantitative tests to test the predictive (actually, postdictive) validity of the two versions of the tests. These were not statistically significantly different, but the N's were only forty for the comparisons. The largest absolute difference in the coefficients was .13 for predicting GRE verbal from the timed and untimed sentence completion tests for one of the experimental groups. The other three differences were all .04 or less.

Lin examined the computer records of the examinees and concluded that 74 and 75 of the 80 examinees answered the items consecutively for the sentence completion and quantitative items, respectively. However, 65% of the examinees reported that they would be more likely to answer the items non-consecutively if the tests had not been administered on a computer.

Lin asked examinees about "feelings of being rushed." About this variable, he summarizes, "The relationship is such that students feeling more time constraints are found to have lower aptitude scores [on the separately administered GRE], which also significantly correlates with both timed and untimed scores. In addition, a scrutiny of the standings on the times and untimed score distributions for those 15 examinees who did not finish reveals that 12 of these examinees remain at the lowest 12 ranks on both tests."

As noted above, Lin's investigation used methods and subjects that are not similar to GATB administration conditions and examinee population; however, most of his findings are in line with other results earlier reviewed, including the differential impact of level and speed on timed tests across different item types and the similar levels of validity for timed and untimed versions of the test (Kendall, 1964; Bell and Lumsden, 1980). What is perhaps most interesting are the two findings about item completion and the relationship between being "rushed" and test performance. The examinees completed items consecutively under computer administration, consistent with the assumption Gulliksen made which underlies his psychometric model of power and speed, but they also reported that they are likely <u>not</u> to follow the assumption under more traditional (paper-and-pencil) testing conditions. Also, the feeling of being rushed was related with scores on the separately administered GRE (lower scores for those feeling more rushed) and with lower scores on both the timed and untimed tests. This finding accords with Ree's (personal communication, September 1, 1993) comment that "all tests are speeded for persons with low ability."

<u>Evidence Specific to the GATB</u>

The findings reviewed so far in this section are illuminating, but are, for the most part, of somewhat limited generalizability to the GATB. The Baxter and Davidson and Carroll studies used omnibus intelligence tests on samples from fifty years ago, and the other studies used primarily college student samples and the GRE. More proximate to the GATB are the conclusions reached in Hartigan & Wigdor about the convergent validity of the GATB aptitudes with constructs measured on other test batteries and some empirical results from Wise and McDaniel (1991) about the relationships of ASVAB and GATB subtests.

Jaeger, Linn, and Test (1989) present a detailed summary of the convergent validity evidence for the GATB aptitudes. Since the GATB tests are speeded power tests, their convergent validities with tests on other batteries provide evidence about the degree to which the GATB scores measure the same constructs that are, presumably, measured on less speeded tests, or, alternately, the extent to which other batteries may have incorporated similar "mixtures" of speed and power in their test scores.

The aptitude evidence can be considered isomorphic with GATB subtests in the case of the vocabulary (V aptitude) and three-dimensional space (S) subtests since these tests are the single measures of those aptitudes; this is not so for the N aptitude, since it is made up of scores from both arithmetic reasoning and computation. Still, the evidence is useful.

About the evidence for V, they say, "Considering the variety of measures with which GATB-V was correlated, and the less-than-perfect reliabilities of the GATB subtests that contribute to V and the tests with which it was correlated, a median validity coefficient of .72 provides adequate evidence of convergent validity for the GATB verbal ability measure (p. 307)." Concerning S, they say, "A median concurrent validity coefficient of .62 with a range from .30 to .73 and a fourth of the coefficients below .58 suggests that somewhat different spatial perception constructs are measured in various batteries, or that the reliabilities of spatial ability measures are somewhat lower than those of corresponding verbal ability measures. Although these data do not cast serious doubt on the construct validity of the spatial ability aptitude, they are not as supportive as the evidence amassed for the verbal ability measure (p. 308)." Concerning N, they state, "A median convergent validity coefficient of .68 is somewhat smaller than would be desired for a measure of numerical ability. However, a claim to convergent validation for GATB-N is reasonably well supported by the data at hand, since three-fourths of the coefficients exceed .61 and a fourth are larger than .75 (p. 307-308)." Thus, the convergent validity of the power subtests of the GATB is characterized as "adequate" for verbal, "not in serious doubt" for spatial, and "reasonably well supported" for numerical aptitude by these authors. Apparently, scores on the GATB speeded power tests measure constructs fairly similar to constructs measured on the other batteries included in the evidence summarized.

ASVAB-GATB Analyses. Wise and McDaniel (1991) collected data on military recruits that allowed confirmatory factor analyses of the correlations of ASVAB and GATB subtests. They administered the first seven subtests of the GATB along with a selected subset of ASVAB subtests, in counterbalanced order, to over 2100 newly enlisted military personnel in the U.S. Army, Air Force, Navy, and Marines. They also had available the full set of ASVAB subtest scores, taken as part of the military entrance process and on record for each recruit. The bulk of the analyses were completed with the ASVAB scores of record. (Analyses were also run with the ASVAB subtests administered concurrently with the GATB to see if the results would differ substantially. They concluded there were higher correlations among the subset of readministered ASVAB subtests, compared to the same subset when administered during the application process. This resulted in some differences in the confirmatory analyses, but for purposes of this discussion these can be ignored.) These analyses are of particular interest because the ASVAB contains some tests that are thought to measure constructs very similar to those purportedly measured by GATB tests. In particular, these include the Arithmetic Reasoning tests from both batteries, GATB Vocabulary and ASVAB Word Knowledge (and perhaps the ASVAB Paragraph Comprehension) tests, the GATB Computation and the ASVAB quantitative tests, especially Number Operations, and the GATB 3-D Space and ASVAB

Mechanical Comprehension, although the ASVAB Mechanical Comprehension is acknowledged as a much more heterogeneous mixture of item types. Except for the ASVAB Numerical Operations test, an intentionally speeded measure, the ASVAB tests are all administered in an acknowledged power condition. Thus, an examination of the relationships of these tests on a recent sample that is fairly similar to the GATB population is very informative.

Wise and McDaniel posited and tested several factor models of the ASVAB and GATB correlations, based on prior factor analyses of the two data sets. They paid particular attention to a six-factor model, but their results also provide support for a five-factor model. Both models had approximately equal goodness of fix indexes (.949 and .951 for the five- and six-factor models, respectively) and root mean square residuals (.060 and .058, respectively). Table 1 shows the pattern matrix (i.e., the loadings of the observed variables on the underlying, proposed factors) and the factor correlation matrix for the five-factor solution. In this model they posited a separate speed factor composed of six of the seven GATB tests (3-D Space was not posited as part of this factor) and two speeded ASVAB tests know to consistently load together (Coding Speed and Number Operations). They also posited quantitative (QUANT), verbal (VERBAL), technical (TECH), and perceptual (PERCEPT) factors. Note that the factor correlation matrix shows very low correlations between the SPEED factor and the VERBAL and TECH factors, but fairly substantial correlations between the SPEED and the QUANT and PERCEPT factors.

Table 1. Pattern Matrix and Factor Correlation Matrix for ASVAB and GATB Subtests; data from 2189 Military Recruits (from Wise and McDaniel, 1991)					
Pattern Matrix	**Factors**				
Subtests	**QUANT**	**VERBAL**	**TECH**	**PERCEPT**	**SPEED**
GATB-Arithmetic Reasoning	.530				.375
ASVAB-Arithmetic Reasoning	.818				
GATB-Computation	.186				.674
ASVAB-Math. Knowledge	.644				
ASVAB-Word Knowledge		.796			
GATB-Vocabulary		.663			.338
ASVAB-Para. Comp.		.604			
ASVAB-General Science		.449	.451		
ASVAB-Auto & Shop			.739		
ASVAB-Elect. Info.			.751		
ASVAB-Mech. Comp.			.805		
GATB-Name Comp.				.394	.409
GATB-Form Matching			.227	.509	.182
GATB-Tool Matching			.072	.727	.045
GATB 3-D Space			.569	.344	
ASVAB-Number Operations					.632
ASVAB-Coding Speed				.231	.470
Factor Correlation Matrix	QUANT	VERBAL	TECH	PERCEPT	SPEED
QUANT	1.000				
VERBAL	.590	1.000			
TECH	.657	.576	1.000		
PERCEPT	.109	.038	.017	1.000	
SPEED	.508	.100	-.054	.556	1.000

Note: the blank cells in the pattern matrix indicate parameters that were set to zero in the confirmatory factor analysis.

Of most interest are the loadings of the four GATB "power" tests in this factor structure. Note that Arithmetic Reasoning has a higher loading on the QUANT factor (.530) than on the SPEED factor (.375), but that this difference is not large. Computation, on the other hand, loads much more highly on SPEED (.674) than on QUANT (.186). GATB Vocabulary loads .663 on the VERBAL factor and .338 on the SPEED factor. Finally, the GATB 3-D Space loads .569 on the TECH factor and .344 on the PERCEPT factor. This pattern of loadings seems to indicate that Computation scores are primarily and significantly influenced by the speed factor, and Arithmetic Reasoning and Vocabulary scores are secondarily influenced by the speed factor, with their primary loadings being on their intended construct factor. 3-D Space did not define the SPEED factor in this model, but it did show a higher loading on the TECH factor, primarily defined by the ASVAB power technical tests, than it did on the PERCEPT factor, primarily defined by the GATB Form Matching and Tool Matching tests, which are speeded. Interestingly, the GATB Name Comparison test, an intentionally highly speeded test, loaded nearly equally on the PERCEPT and SPEED factors which provides an indication that both of these factors are "speed" factors to a similar degree.

A Reanalysis. While these results are interesting, other analyses of the correlations are also useful. In order to further examine these data, I corrected the correlation matrix for attenuation due to unreliability of the tests (I used alternate forms reliability coefficients for the ASVAB [Waters, Barnes, Foley, Steinhaus, and Brown, 1988] and averages of six alternative forms reliabilities for GATB subtests [Manual for the USES General Aptitude Test Battery, 1985]) and then performed three analyses. First, the corrected correlations between GATB and ASVAB test scores purportedly measuring the same or similar constructs were examined. Second, a principal components factor analysis of the corrected correlation matrix was done and the unrotated solution was examined to see if the loadings on the first factor, which can be considered a measure of psychometric "g" (Ree and Earles, 1991a), were similar for the GATB and ASVAB tests of the same or similar constructs. Finally, a principal factors analysis (multiple correlation estimates of communality were used) with varimax rotation of five factors was done to examine the pattern of loadings for the GATB and ASVAB test scores, allowing an examination of the comparative influence of the factors on the scores.

Table 2 shows the correlations between GATB and ASVAB test scores that are purported to measure the same or similar constructs. Bolded coefficients indicate that the correlation represents the relationship between two measures of purportedly highly similar constructs as measured on a power, ASVAB test and on a GATB, speeded power test. Italicized coefficients indicate a correlation between measures of purportedly similar, but somewhat distinct constructs from the ASVAB and the GATB. Finally, the underlined coefficient represents a coefficient between two measures of purportedly highly similar constructs, measured on a speeded ASVAB test (Number Operations) and on a GATB test (Computation).

Table 2. Correlations Between Selected ASVAB and GATB Tests, Corrected for Attenuation Due to Unreliability						
Tests	ASVAB					
GATB	Arith. Reas.	Math. Know.	Number Oper.	Mech. Comp.	Word Know.	Para. Comp.
Arith. Reas.	**.725**	*.584*	.513	.370	.363	.364
Computation	**.533**	**.512**	.665	.102	.124	.180
3-D Space	.478	.338	.081	*.669*	.291	.237
Vocabulary	.510	.493	.291	.423	**.681**	*.561*

The true score correlation for Arithmetic Reasoning is .725, which compares favorably with correlations between GATB parallel forms (these range from .75 to .79, [USES Manual, 1985]). The true score correlation between GATB and ASVAB Vocabulary is .681, which is about .20 less than the correlations between GATB parallel forms (USES Manual, 1985). The true score correlation of 3-D Space with the ASVAB Mechanical Comprehension, a similar, but by no means identical construct--it contains many item types not found in 3-D Space--is .669, about .14-.15 less than GATB parallel forms correlations. Finally, Computation has true score correlations that are higher with the speeded ASVAB Number Operations test than with the non-speeded ASVAB measures of quantitative ability. This is not surprising since the ASVAB Number Operations items are much more similar in content to the GATB Computation test, although GATB items have a much wider range of difficulty, than they are to ASVAB Arithmetic Reasoning and Math Knowledge items. The .665 true score correlation between GATB Computation and ASVAB Number Operations is about .18-.20 less than GATB parallel forms correlations.

Another way of evaluating this set of recent ASVAB/GATB correlations is to compare the uncorrected coefficients to the median convergent validity coefficients of GATB test scores with scores on other measures of purportedly similar constructs, reported by Jaeger, Linn, and Tesh (1989, see above). For Arithmetic Reasoning the comparison is .594 versus .68 (the convergent validity for the N aptitude), for Computation, it is .514 versus the same .68, for Vocabulary it is .601 versus .72, and for 3-D Space it is .534 with .62. Thus, except for Computation, the observed correlations here are about 9 to 12 points below the median convergent validities.

All of this information points to the conclusion that the GATB Arithmetic Reasoning test seems to be measuring a construct similar to that measured by the ASVAB Arithmetic Reasoning, while GATB Computation seems to be measuring a construct somewhat more similar to that measured by the speeded ASVAB Number Operations tests than to the ASVAB power tests of quantitative ability. This latter conclusion is not too surprising since the ASVAB power tests of quantitative ability have obviously different item types than does GATB Computation. GATB 3-D Space seems to be more closely related to ASVAB Mechanical Comprehension than would be expected based on the differences in their item types. GATB Vocabulary does seem to be measuring a construct fairly similar to ASVAB Word Knowledge, but there is a fairly large amount of variance that is not in common.

Shown below are the loadings of the 17 GATB and ASVAB tests on the first principal component of the matrix of correlations. Examining the comparable ASVAB and GATB tests shows the extent to which they similarly measure psychometric "g" as measured by this particular battery. Tests with the same superscripts are the most relevant comparisons. Note that ASVAB and GATB Arithmetic Reasoning load .79 and .80, ASVAB Mechanical Comprehension and GATB 3-D Space load .68 and .67, ASVAB Word Knowledge and GATB Vocabulary load .60 and .78, and, finally, ASVAB Number Operations and GATB Computation load .49 and .62. These findings reinforce the position that GATB Arithmetic Reasoning and ASVAB Arithmetic Reasoning measure similar constructs as do GATB 3-D Space and ASVAB Mechanical Comprehension. However, these findings show that the GATB Vocabulary test loads higher on "g" as measured by this set of tests than does the ASVAB Word Knowledge test. A similar statement holds for GATB Computation compared to ASVAB Number Operations.

Test	Loading
GATB-Arithmetic Reasoning	.79[1]
ASVAB-Arithmetic Reasoning	.80[1]
GATB-Computation	.62[2]
ASVAB-Math. Knowledge	.70
ASVAB-Word Knowledge	.60[3]
GATB-Vocabulary	.78[3]
ASVAB-Para. Comp.	.58[3]
ASVAB-General Science	.70
ASVAB-Auto & Shop	.48
ASVAB-Elect. Info.	.59
ASVAB-Mech. Comp.	.68[4]
GATB-3-D Space	.67[4]
GATB-Name Comp.	.54
GATB-Form Matching	.64
GATB-Tool Matching	.50
ASVAB-Number Operations	.49[2]
ASVAB-Coding Speed	.51

(Tests with the same superscripts are the most relevant comparisons.)

These similar loadings (e.g., ASVAB and GATB Arithmetic Reasoning load .79 and .80) across comparable ASVAB and GATB tests on the first principal component are somewhat inconsistent with results from the Wise and McDaniel confirmatory factor analysis. In Table 1, three of the GATB power tests have a smaller loading on their corresponding constructs than the corresponding ASVAB tests (e.g., GATB Arithmetic Reasoning loads .530 on QUANT and ASVAB Arithmetic Reasoning loads .818 on QUANT). Table 1 also indicates that these GATB power tests have substantial loadings on the SPEED factor (e.g., GATB Arithmetic Reasoning loads .375 on SPEED). How is it possible for the GATB Arithmetic Reasoning test to have a substantially smaller loading on QUANT than the ASVAB Arithmetic Reasoning test and a substantial loading on SPEED in the confirmatory factor analysis, and at the same time the GATB and the ASVAB versions of this test have almost identical loadings on the first principal component in a principal components analysis? The answer is that the Wise and McDaniel confirmatory factor analysis is partially confounding a general factor with their SPEED factor by having a fairly large number of the tests in the battery defining the SPEED factor, some of which

should be expected to load fairly highly on a general factor--as they were shown to do above. This point should be kept in mind during the following discussion of the exploratory factor analysis results.

Table 3 shows the five-factor structure when the corrected correlations are subjected to a principal factors analysis (multiple correlation estimates of the communalities) with a varimax rotation. In general, the same factors emerged from this exploratory factor analysis as for the five-factor model in Wise and McDaniel's confirmatory analysis, although the SPEED factor in the exploratory analysis is primarily defined by just the ASVAB Coding Speed and Number Operations tests and does not show the high loading from GATB Computation found in Wise and McDaniel.

GATB Arithmetic Reasoning and Computation loaded most highly (.75 and .71, respectively) on the QUANT factor with fairly high loadings on the PERCEPT factor (.32 and .38). Arithmetic Reasoning showed a low loading on the SPEED factor (.12) while Computation showed a moderate loading (.30). In the Wide and McDaniel confirmatory solution, Computation showed a much higher loading on its SPEED factor relative to the QUANT factor. Computation loads much more similarly to Arithmetic Reasoning in this exploratory analysis than it does in the Wise and McDaniel five-factor solution.

A comparison of the GATB Arithmetic Reasoning and ASVAB Arithmetic Reasoning factor loadings in Table 3 provides some indication of the extent to which their scores are influenced by the same factors. Their communalities are very similar (.783 and .774) and their highest loadings are similar (.75 and .70 on the QUANT factor). They also have very similar, low loadings on the VERBAL (.25 and .29) and SPEED (.12 and .12) factors. They diverge on the loadings on the TECH and PERCEPT factors. GATB Arithmetic Reasoning loads .32 on PERCEPT and .18 on TECH, while ASVAB Arithmetic Reasoning loads .42 on TECH and .08 on PERCEPT. Since PERCEPT is primarily defined by speeded tests (GATB Tool Matching, Form Matching, and Name Comparison) and TECH is primarily defined by ASVAB technical power tests, this provides some evidence that speed influences GATB Arithmetic Reasoning scores more than ASVAB Arithmetic Reasoning.

Table 3. Principal Factor Solution for GATB and ASVAB Subtests						
	Factors					
Subtests	QUANT	VERBAL	TECH	PERCEPT	SPEED	H^2
GATB-Arithmetic Reasoning	.75*	.25	.18	.32	.12	.78
ASVAB-Arithmetic Reasoning	.70*	.29	.42	.08	.12	.77
GATB-Computation	.71*	.05	-.08	.38	.30	.75
ASVAB-Math. Knowledge	.63*	.32	.21	.08	.21	.59
ASVAB-Word Knowledge	.12	.83*	.24	.00	.03	.76
GATB-Vocabulary	.33	.67*	.18	.36	.04	.73
ASVAB-Para. Comp.	.18	.71*	.18	.00	.12	.58
ASVAB-General Science	.17	.63*	.60*	.02	-.03	.79
ASVAB-Auto & Shop	.04	.10	.88*	.00	-.07	.78
ASVAB-Elect. Info.	.06	.28	.85*	.02	-.01	.81
ASVAB-Mech. Comp.	.21	.26	.83*	.12	-.08	.82
GATB-Name Comp.	.28	.13	-.13	.68*	.32	.68
GATB-Form Matching	.21	.05	.22	.79*	.13	.74
GATB-Tool Matching	.09	-.02	.07	.84*	.16	.75
GATB-3-D Space	.17	.17	.57*	.49*	-.08	.63
ASVAB-Number Operations	.46	.06	-.14	.28	.68*	.78
ASVAB-Coding Speed	.27	.13	-.12	.44	.70*	.79
Eigenvalues of Rotated Factors	2.63	2.52	3.35	2.73	1.31	12.53

Note: Asterisks (*) mark the highest loading in a row plus any loading which has a squared value greater than ½ of the square of the highest loading (to indicate significant secondary loadings).

A comparison of GATB Computation to ASVAB Number Operations loadings show highly similar communalities (.75 and .78, respectively), highly similar loadings on VERBAL (.05 and .06) and TECH (-.08 and -.14), and fairly similar PERCEPT loadings (.38 vs. .28), but substantially different loadings on the QUANT and SPEED factors. GATB Computation loads highly on QUANT (.71) and moderately on SPEED (.30) with ASVAB Number Operations showing the reverse pattern (.46 on QUANT and .68 on SPEED). These patterns seem to indicate that the GATB Computation test scores are heavily influenced by the QUANT factor with moderate influence from the two speed factors (SPEED and PERCEPT), while the reverse is the case for ASVAB Number Operations test scores.

Comparing GATB 3-D Space's pattern of loadings to ASVAB Mechanical Comprehension shows fairly large differences. The ASVAB Mechanical Comprehension test has a higher communality (.82 vs. .63) and loads much higher on TECH than does GATB 3-D Space (.83 vs. .57). ASVAB Mechanical Comprehension shows a very low loading on the PERCEPT factor, while 3-D Space loads substantively there (.49). These patterns confirm earlier results showing that these test scores are influenced by similar, but still substantially different factors.

Again, examining, the pattern of GATB Vocabulary and ASVAB Word Knowledge loadings across the five factors is instructive. Their communalities are similar (.73 and .76) and their highest loadings are on the VERBAL factor, but ASVAB Word Knowledge loads substantially higher than does GATB Vocabulary (.83 vs. .67, respectively). They load similarly on TECH (.24 and .18) and SPEED (.03 and .04), but GATB Vocabulary loads higher on both PERCEPT (.36 vs. .00) and QUANT (.33 vs. .12). Since PERCEPT, as noted above, is primarily defined by speeded tests, this seems to indicate that GATB Vocabulary scores are somewhat more influenced by speed than are ASVAB Word Knowledge scores. Comparing ASVAB Paragraph Comparison to GATB Vocabulary leads to much the same conclusion, except that ASVAB Paragraph Comprehension has a lower communality (.58) than either ASVAB Word Knowledge or GATB Vocabulary.

In summary, the findings reviewed in this section which include the convergent validity evidence summarized by Jaeger, Linn, and Tesh (1989) and the analyses of a recent, highly relevant data set seem to point to the following general conclusions:

- GATB Arithmetic Reasoning measures a construct very similar to the ASVAB measure by that name (their true score correlation was .73 and they load almost identically on "g" in a sample of military recruits), and its scores appear to be somewhat influenced by speed but not inordinately so. The convergent validity evidence is difficult to interpret here since it is summarized at the aptitude level, rather then the subtest level, but it is characterized by Jaeger, et. al. as showing that convergent validity is "reasonably well supported" for the N aptitude.

- GATB Vocabulary is characterized by Jaeger et. al. as having "adequate" convergent validity evidence and the GATB/ASVAB analyses show reasonably high true score correlations between it and ASVAB Word Knowledge (.68). It loads higher on "g" than did ASVAB Word Knowledge and the confirmatory factor analysis shows a moderate loading on the speed factor in that analysis, but, for reasons noted above, this loading might be an overestimate of the influence of speed on Vocabulary scores. The exploratory factor analysis supported these findings.

- GATB Computation, based on the evidence from the GATB/ASVAB analyses, appears to measure a construct fairly similar to the ASVAB Number Operations, certainly more similar to that ASVAB test than to the other quantitative ability tests on the ASVAB (the true score correlation between GATB Computation and ASVAB Number Operations is .67, compared to .53 and .51 for the other ASVAB quantitative tests). However, it does load higher on "g" than does ASVAB Number Operations, perhaps due to its wider range of item difficulties, and in the exploratory factor analysis showed a factor pattern with similarities to both Arithmetic Reasoning (both loaded highest on the quantitative factor) and to Number Operations (similar loadings on the verbal, technical, and perceptual factors).

- GATB 3-D Space's convergent validity evidence is characterized by Jaeger et. al. as "not in serious doubt", but they note that somewhat different spatial ability constructs appear to be measured across batteries. This is an apt description of the ASVAB Mechanical Comprehension and GATB 3-D Space relationship. Their true score correlation is .67 and this is almost identical to their loadings on "g". As would be expected from their differences in item content, however, ASVAB Mechanical Comprehension shows higher loadings on the Technical factor identified in the factor analyses above, and 3-D Space shows higher loadings on the Perceptual factor.

These results show that examinees appear to respond to the GATB tests in a manner that does not reflect a high degree of speededness (except perhaps for Computation), though these tests are admittedly speeded by most definitions of speed. A possible explanation for much of this finding is the fact that GATB test items are ordered by difficulty. This ordering operates to lessen the effect of speed on examinees' scores since many examinees, when they have run out of time, have reached the point in the order of items at which they are unlikely to get any more items correct. Some of the findings reviewed in the next section tend to substantiate this hypothesis.

Differential Effects of Speededness

In this section we review literature that speaks to the effect of increasing or decreasing the speededness of a test on test scores achieved by different race and sex subgroups. We are particularly interested in interaction effects, that is, differential changes across subgroups associated with increases or decreases in testing time on achieved test scores and the relationships of test scores with one another or with external criteria. Generally speaking, there are well-documented subgroup differences in achieved test scores (Ackerman and Humphreys, 1990; Cronbach, 1990; Linn, 1982) with some racial and ethnic groups (black and Hispanic) achieving generally lower scores than the majority group (white) and others (Asian) achieving generally higher scores, though these results vary across type of test and particular groups being

compared. I will not repeat those findings here, instead, I will focus on fairly specific attempts to evaluate the impact of changing time limits on the achieved scores of subgroups or attempts to ascertain the degree of speededness of tests for different subgroups, primarily race and sex subgroups. I also will not attempt to review the literature on differential validity and test fairness which is far too voluminous to summarize here. I note that these phenomena depend as much on subgroup differences on criterion scores as they do on test scores differences (Ackerman and Humphreys, 1990) and are, therefore, much more difficult to interpret in terms of the likely effect of changes in the speededness of a particular test.

Studies of the effects of changes in speededness of tests on subgroups' scores have been conducted on the GRE, the LSAT, the SAT, the Basic Occupational Literacy Test (BOLT), and the Employee Aptitude Survey (EAS). The samples vary from law school aspirants to adults in a variety of training settings such as basic education classes and government-sponsored training courses for unemployed adults.

EAS

Dubin, Osburn and Winick (1969) compared black (N = 235) and white (N = 232) high school students' performance on speeded (normal time limits) and unspeeded (tripled time limits) versions of two tests from the Employee Aptitude Survey: Numerical Reasoning (solving number series items) and Verbal Reasoning (drawing valid conclusions from a list of facts). Parallel forms were used for the speeded and unspeeded versions. They also compared race differences on speeded tests of Numerical Ability (solving computational problems) and Space Visualization (counting three-dimensional blocks) with and without first completing a parallel, practice test. Information is not provided to determine if the "unspeeded" versions are actually unspeeded, but in any case the "power" time limits are three times longer than the normal time limits.

Dubin et al. performed analyses of variance to determine if the increase in time limits resulted in differential improvements in test performance for blacks or whites (both groups significantly improved their performance when given more time, indicating that the tests were significantly speeded under normal time administrations). To do this, they matched black and white examinees on sex, grade level, and socioeconomic status (SES) index. They found no differences in improvement between the black and white groups for either the Verbal or Numerical Reasoning test. They also found no effects for SES or SES x Race. They obtained similar findings for the Space Visualization and Numerical Ability tests with regard to practice, that is, a significant main effect for practice, but no interaction with race, SES, or race and SES. These findings indicate that blacks did not benefit more than whites from significantly increased time limits on verbal and numerical reasoning tests nor from an opportunity to practice for speeded tests of computation and three-dimensional spatial ability.

BOLT

A study of the Basic Occupational Literacy Test (BOLT) by the Utah State Dept. of Employment Security (1981) showed significant, differential changes in several BOLT scores across sex, age, race, and economic status subgroups when normal (considered to be a power condition) and one-half normal time limit scores were compared. The BOLT subtests analyzed included fundamental and advanced versions of reading vocabulary, reading comprehension, arithmetic computation, and arithmetic reasoning. Examinees were assigned to either the fundamental or advanced batteries based on a modified BOLT Level Grid that used stated educational level and Wide Range Scale scores.

Results included t-tests for the difference between one-half normal time scores (called Brief Time) based on one form and normal time scores (called Standard Time) based on another form. Overall t-tests showed that there were significant differences between Brief Time and Standard Time scores for all the advanced tests (N=493-664), but no significant differences for the fundamental tests (N=118-209). The mean raw score differences were about one-half point or less for the fundamental tests and two of the four advanced tests. These findings indicate that cutting the time in half may not have substantially speeded the tests, on average, for this sample of examinees. The authors cautioned that the results may have been influenced by the use of the modified BOLT grid for assigning examinees to levels of test, and the sample sizes were also much larger for the advanced level tests, resulting in more powerful tests of mean differences for those tests.

Analyses of variance were performed for the Brief Time, Standard Time, and the differences between Brief and Standard Time scores. Age, sex, ethnic group or race, and economic status were used for analyses of the advanced tests and age, sex, and ethnic group were used for the fundamental tests. Results indicated significant overall effects for the proposed models for both the fundamental and advanced tests. Examination of the significance of main and interaction effects for the models showed a larger number of significant efforts for the advanced tests than for the fundamental tests. Focusing on the results for the Brief/Standard Time difference scores, they found no significant main or interaction effects for reading comprehension or arithmetic computation at either level. For reading vocabulary they found a significant effect for ethnic group at both levels, and a significant age by ethnic group effect at the fundamental level. For arithmetic reasoning, they found no significant effects at the fundamental level, but, at the advanced level, they found significant effects for ethnic group and three higher-level interactions, all including economic status (sex by economic status, age by sex by economic status, and sex by ethnic group by economic status).

Finally, they performed a posteriori contrasts (Duncan Multiple Range Tests) on the Brief/Standard Time difference scores for age, sex, economic status, and ethnic group. No significant contrasts were found for three of the four fundamental tests (reading vocabulary, reading comprehension, and arithmetic reasoning). For fundamental arithmetic computation, older examinees gained significantly more than younger examinees when given the full time and non-minority examinees gained more than black examinees (sample sizes were too small for other ethnic groups to provide conclusive results). Significant contrasts were found for three of the four advanced tests; none was found for arithmetic computation. For advanced arithmetic reasoning, significant contrasts were found for age (older examinees gained more than younger examinees), sex (females gained more), economic status (non-disadvantaged gained more), and

ethnic group (blacks gained more than non-minorities). For reading comprehension, the economically disadvantaged group gained more than non-disadvantaged and blacks gained more than non-minorities. For advanced reading vocabulary, blacks gained more than non-minorities.

Based on these results, the authors concluded that the BOLT time limits could not be reduced without adverse effects on minority groups. While the pattern of results is neither particularly strong nor consistent, it does seem to indicate a greater effect of *reduced* time limits for the arithmetic reasoning (advanced) test with regard to adverse effects on minority groups. Also, blacks benefited more than non-minorities from increased time limits for three of the four advanced tests.

The BOLT study results are not consistent with the EAS results, which showed no differential improvement of black examinees over white examinees. Aside from the obvious differences in the two test batteries, the EAS study used high school students only and directly matched black and white examinee groups on sex, SES, and grade level before making comparisons. These factors might account for the differences in results, but the BOLT study also included and examined similar variables (age, sex, economic status) and did not produce results which might indicate that matching would change results (such as a large number of higher-order interactions of race with SES, age, and sex).

LSAT

Evans and Reilly (1973, 1972) conducted research with applicants to graduate schools on test speededness and adverse effects on minority groups. Using law school candidates (Evans and Reilly, 1972) from fee-free (candidates are from predominantly black colleges) and fee-paying testing centers, data were collected for speeded (35 items in 40 minutes) and unspeeded (27 items in 40 minutes) reading comprehension tests on the Law School Admission Test (LSAT). Only the 27 common items were scored using no correction for guessing. Their analyses confirmed expectations that candidates in the fee-paying centers obtained higher scores than those in the fee-free centers and that all candidates performed better on the unspeeded versions of the tests. Analyses also showed that the normally timed LSAT and experimentally speeded versions of reading comprehension did not meet the usual Educational Testing Service's criteria for a power test (i.e., all candidates reach 75% of the items and 80% reach the last item) for the fee-free candidates, but the experimentally unspeeded version came extremely close to doing so (about 98% of the examinees completed 75% of the items and about 90% completed all items for the unspeeded version, compared to about 80% of the examinees completing 75% of the items and about 63% completing all the items for the speeded version). For the fee-paying candidates, power or near-power conditions appeared to prevail in all conditions.

Most importantly, analysis of variance showed that there was no interaction effect (type of test center by speededness of form) on scores. They also correlated type of test center with test score for speeded and unspeeded forms and these were only slightly different, .46 for speeded samples and .42 for unspeeded samples. These results indicated that examinees in the fee-free centers did not benefit more from the less speeded conditions than did examinees in the fee-paying centers. In addition, Evans and Reilly reported that the KR-20 reliabilities were higher in the unspeeded condition for the fee-free examinees than in the speeded condition, somewhat counter to expectations, and rising the possibility that the speededness of the LSAT reading comprehension section was producing less reliable results for fee-free candidates than for fee-paying candidates.

They concluded, "Results of the analyses show: (1) the test is somewhat more speeded for fee-free candidates than for regular candidates, (2) reducing the amount of speededness produces higher scores for both regular and fee-free center candidates, and (3) reducing speededness is *not* [italics theirs] significantly more beneficial (in terms of increasing the number of items correct) to fee-free than to regular center candidates (p. 130)."

ATGSB

Evans and Reilly (1973) followed up the LSAT study with a study of the quantitative section of the Admission Test for Graduate Study in Business (ATGSB). The design was essentially similar to the LSAT design, except that they used three experimental versions of the quantitative section: normal time limit (80 seconds per item), speeded time limit (69 seconds per item), and unspeeded time limit (96 seconds per item). They also had three experimental groups of candidates: black candidates from fee-free testing centers, black candidates from fee-paying or regular testing centers, and white candidates from regular testing centers. However, in the ATGSB quantitative test the items were ordered by difficulty and a correction for guessing was used. Neither of these held for the LSAT study and could account for differences in the results between the two studies.

Analyses showed that all three versions of the test were moderately speeded for all groups of candidates, that scores were not significantly different across the three versions of the test (unlike the LSAT where scores were significantly different across versions), and that increasing or decreasing the time allowed per item resulted in no differential effects in scores across the three groups of candidates (like the LSAT study). Evans and Reilly concluded, "Aside from the possibility that quantitative abilities are more resistant to speed effects than verbal abilities, the major reason for this discrepancy in results may be the way in which the experimental sections in each of the studies were constructed. In the previous study, the items were not clearly ordered for difficulty and there was no correction for guessing, while the reverse was true for the experimental section in the present investigation. Both ordering items in terms of their difficulties and correcting for guessing would tend to weaken any effects due to speed (p. 181)." Note that this explanation for the discrepancy in results is consistent with the theoretical statements earlier discussed about the complex interplay of time-limit, item-difficulty levels, applicant populations, and administration conditions in determining the influence of speed on test scores. These results also support our earlier contention that ordering items by difficulty will tend to reduce the effects of speed on test scores.

GRE

Wild and Durso (1981) built on the results from the two Evans and Reilly studies, extending the research to the Graduate Record Examination (GRE). They investigated the effects of increasing the time allowed to complete experimental sections on the GRE: verbal items (a mixture of items from reading comprehension passages and stand-alone, discrete items) and quantitative items (a combination of discrete mathematics items and data interpretation items). The experimental sections were completed under two time limits, 20 minutes (the normal time limit of 46 seconds per item for verbal and 86 seconds for quantitative) and 30 minutes (the "unspeeded" time limit of 69 seconds per item for verbal and 129 seconds for quantitative). Items were ordered by difficulty, corrections for guessing were used, and the data were collected as part of the normal test administration process. Scores were obtained from operational sections for the same item

types in the same administration. Verbal operational scores were based on two separately timed sections: 25 minutes for 55 discrete questions and fifty minutes for 40 reading comprehension questions. Quantitative operational scores were based on 55 questions answered in 75 minutes. Information was collected on sex, race (black or white only), and number of years since receiving the baccalaureate degree (in five categories of five-year intervals).

As with the Evans and Reilly studies the primary concern was with differential change in test scores for the sex, race, and "time since degree" groups with the increase in testing time. These differential changes were identified by appropriate interaction effects (e.g., race by test version). Using regression analyses that controlled for initial differences on verbal and quantitative ability (using the operational GRE scores), Wild and Durso found no evidence for such significant interaction effects. Significant t-tests occurred at the chance level (7 of 261 significant for verbal and 15 of 257 for quantitative, when approximately 13 would be expected by chance).

Item and test score analyses revealed that the experimental 20-minute version of the verbal test did not meet either of the two typical ETS criteria for power tests (i.e., all candidates reach 75% of the items and 80% reach the last item) for blacks under the 20-minute version, and only marginally met one of the criteria under the 30-minute versions. For whites, one of two criteria was marginally met under the 20-minute version and one was definitely met under the 30-minute version. The quantitative test showed one criterion marginally met for blacks under the 20-minute version and both criteria met for black males and one met for black females under the 30-minute version. A similar pattern occurred for whites, except that white males definitely (rather than marginally) met one criterion under the 20-minute version. As expected, scores were generally higher under the 30-minute version for all groups on both quantitative and verbal tests, and whites showed higher scores across all test versions. Regarding reliabilities, the experimental 20- and 30-minute verbal tests were about equally reliable with black reliabilities somewhat lower (.73 vs. .79) and were somewhat lower than the reliabilities for the two operational sections (which showed reliabilities ranging from .80 to .88 depending on subgroup and type of item, i.e., reading passage or discrete). For the quantitative test, reliabilities were generally lower for the 20-minute version compared to the 30-minute version, and were lower for blacks than for whites. Black male and female reliabilities for the 20-minute version were .64 and .50, respectively, compared to .72 and .52 for the 30-minute version; comparable values for white males and females were .75 and .70 (20-minute) versus .77 and .73 (30-minute). Operational test reliabilities for the quantitative test, in contrast, range from .83 to .91, depending on sample subgroup, with females showing lower reliabilities. Thus, the operational tests show higher reliabilities for both tests, but the quantitative tests showed comparatively more degradation in reliability when they were given in the experimental version than in the operational versions. The quantitative tests also showed slightly more degradation in reliability in the shorter versions of the experimental tests.

ATGSB/GRE Summary

Taken together, the Evans and Reilly and Wild and Durso studies indicate that black candidates for graduate school admission do not benefit more from increased time to complete verbal and quantitative tests than do white candidates in terms of obtained scores, though there may be other benefits such as increased reliability of scores. The studies do indicate that tests as normally given to these populations appear to be much more speeded for black examinees than for white examinees, but generally not nearly as speeded as the GATB power tests are for its population of applicants. Obviously, the GATB population is different from the population sampled in these studies so generalizing these findings to the GATB is risky.

SAT

Donlon (1977) investigated male-female differences with respect to speededness of the Scholastic Aptitude Test (SAT). Donlon compared 1000 male and 1000 female SAT examinees (generally SAT examinees are persons in their senior year of high school aspiring to attend four-year colleges) using data from a standard administration of the SAT. He used graphic techniques employing normal probability paper to compare the two groups at an overall level and within ability quintiles (quintiles were separately computed for the sexes). The total group comparisons on percentage of group completing all items showed very little difference between males and females for any of the four SAT sections, though the item completion data did vary as expected with the rate-of-work requirements for the four sections. The sections varied from 45 (70% completion) to 54 seconds (86% completion) per item for the two verbal sections and 72 (81% completion) to 77 (93% completion) seconds per item for the two mathematical sections. (The two verbal sections each contained four item types: analogies, sentence completion, antonyms, and passage-related reading comprehension items, and the two mathematical sections contained "general mathematical material," except that the second mathematics section contained some data sufficiency items, which ask only that the examinee indicate whether or not there is enough information to answer each item. These mathematical items take less time, so the second section was expected to be less speeded than the first section for this reason as well as the longer time allowed per item.) Thus, the first of the two verbal sections and the first of the two mathematics sections were more speeded than the second sections, but all four sections pass or come extremely close to passing the two ETA criteria for power tests -- except for the first verbal section, where only 70% complete all items (criterion equals 80%). None of the sections is very speeded.

Examination of plots for males and females in the high and low ability quintiles showed that, for verbal, the high quintile groups for males and females both attempt more items than the low quintile groups for the sexes, as would be expected since male and female performance for the total test is about equal on verbal. For mathematics, where males outperform females on the total test, however, the low-scoring females finished more of the test than did the low-scoring males -- the opposite of expectations based on the superior performance of males on the total test. Indeed, for the first mathematics section, the low-quintile females actually show a slightly higher completion percentage than the high-quintile females. Thus, the low-scoring females are working at a much faster rate than anticipated based on their level of ability. The authors offer no firm rationale for this finding. They concluded, in general, that "There is no evidence of appreciable [sex] differences in rate-of-work on any section of the Scholastic Aptitude test. There is some evidence that low-scoring females on the mathematical sections work at faster

rates than would be anticipated by the data for males or higher-scoring females. There is some evidence that while the time limits are so generous as to reduce variance in failure to complete, there may be sizable differences in time to review, go back over the test, etc., the score implications of this variance are seldom considered (p. 11)."

Schmitt and Bleistein (1987) used differential item functioning analyses to identify factors that might influence black examinees' performance of verbal analogy items on the SAT. These types of items previously had been identified as being unexpectedly more difficult for black examinees than for white examinees. They noted, "The most significant finding is that Black students appear to need more time to complete the SAT verbal sections than White students with comparable total SAT verbal scores. This differential speededness effect makes analogy items appear differentially more difficult for Black examinees. Once differential item functioning statistics were corrected for speededness, a smaller number of analogy items were identified as differentially more difficult." Although they did not submit the antonym items to the same intensive analysis, some of the data in their report indicate that a small percentage of these items consistently shows up across forms as more difficult for black examinees. (Antonym items, of course, are similar to the item type used on GATB vocabulary whereas verbal analogies are not similar.) Of course, "correcting" for speededness does not change the fact that black examinees appear to be more adversely affected by speed on the SAT verbal items.

These two SAT studies sample from populations that are likely to be somewhat different from those that routinely take the GATB, but perhaps more similar than those sampled by the Evans and Reilly and Durso and Wild studies of graduate school applicants. Also, the SAT studies did not directly manipulate speededness or amount of time available to complete a test, rather they used indexes of speededness for operational tests to compare the degree of speededness for race and sex groups. The findings seem to indicate that there are no sex differences with regard to speededness of the SAT, with a minor qualification about low-scoring females, but that black examinees do appear to be adversely affected by speed for verbal analogy items on the SAT, and perhaps, for antonym items.

Overall Summary

Taking all the studies in this section into account, and weighting the EAS and BOLT studies a bit more heavily because of their more similar (to GATB) item types and sampled populations, the evidence is, at best, mixed concerning the beneficial effects for blacks (no evidence for other groups was found and reviewed here) of reducing the speededness of tests of verbal and quantitative abilities similar to those measured by the GATB. No research bearing directly on spatial ability and direct manipulations of speededness was located and reviewed (Dubin, Osburn, & Winick [1969]) studied effects of practice, but not speededness on a spatial ability test). With regard to spatial ability, results from a recent, large-sample study of U.S. Army recruits (Peterson, et. al., 1990) showed that blacks performed relatively similarly, compared to whites' performance, on both highly-speeded and less-speeded tests of spatial ability. Females performed similarly, compared to males, on two highly-speeded tests and two less-speeded tests of spatial ability (about 1/3 standard deviation lower), but females and males showed no difference in performance on two other, less-speeded tests of spatial ability. Speededness was not directly manipulated within tests in these data, so degree of speededness is confounded with the particular item types used on each of the tests. Finally, in this regard, Ackerman and Humphreys note that, "On the whole, blacks do relatively better on tests in which speed makes a

substantial contribution to variance (p. 268)." (They appear to be speaking of highly speeded tests made up of easy items, since they say "speed makes a substantial contribution to variance" and make this statement immediately after describing such tests.)

Adverse Psychological Reactions

Hartigan and Widgor (1989) speculate that "the severe time limits of the GATB subtests might produce an adverse psychological reaction in examinees as they progress through the examination and might thereby reduce the construct validity of the subtests (p. 106)." They opine that those most affected by this reaction would be those "least experienced with standardized tests, a group in which minority examinees will be overrepresented (p. 106)." They note the results by Dublin, Osburn, and Winick (1969) that do not support this speculation (Dublin et. al.'s study is reviewed above), but mention several studies cited by Dublin et. al., as indicating that extra practice and test familiarity reduce test performances between blacks and whites.

These studies included Boger (1952) who studied the effects of perceptual training (pictorial, jigsaw, and wood puzzles) on group IQ test scores for black and white students (N = about 50 in each group divided into experimental and control groups) in rural, one- or two- teacher elementary schools (grades 1-4). Students received several types of training each day for seventy-eight consecutive school days. Both black and white students in the experimental groups showed statistically significant gains. No analyses of interaction were completed; but inspection of the means in their report shows similar levels of gain for blacks and whites. It should be noted that both groups of students scored, on average, below national averages on the tests prior to training, so some amount of the improvement may be due simply to regression to the mean. The generalizability of these results to the issue of the speededness of GATB tests for adult populations seems dubious.

Eagleson (1937) studied the effects of training on a simple visual discrimination task administered in individual settings on a special apparatus. The task was to bisect a line by using a knob to move the bisection indicator, and errors in bisection were recorded (i.e., distance from actual bisection point). Subjects included black and white males (N = 12 and 7, respectively) in Civilian Conservation Corps camps; 11 white females, 14 white males, and 25 black males at Indiana University; and 50 black and 50 white high school students (24 male and 26 female in each group). Results indicated that all groups improved over the course of training at this task and the performances of the groups were converging, but blacks, starting from initially poorer performance, improved at a faster rate. The generalizability of results using this simple, apparatus-based task to the GATB, paper-and-pencil, speeded power tests does not seem strong.

Katzenmeyer (1962) compared the performance on the Lorge-Thorndike Intelligence Test of 193 black students and 1061 white students who were enrolled in the first two years of a school desegregation program in Jackson, Michigan in 1957-1959. Gains in scores from kindergarten to second grade were compared for the two groups. Blacks initially scored significantly lower than whites, but their gain scores were significantly greater than whites over the two years. Katzenmeyer attributes the more significant gain to desegregation particularly the social interaction with white students, though the absence of control groups (groups of white and black

students not participating in desegregation) allows the possibility of competing explanations. Even if this attribution is correct, the relevance of the findings to issues of concern with GATB speededness appears small.

Klineberg (1928) summarizes data relating to speed and accuracy indices computed on several tests for groups of Native Americans, blacks, and whites and attempts to interpret the results in a "nature versus nurture" context. I must confess to being unfamiliar with most of the tests mentioned (Mare and Foal, Healy "A"), but a Form Board test is mentioned, which is generally a speeded test. Most of the indices computed appear to depend on a single administration of a test and are subject to the problems associated with those indices, as detailed by Rindler (1979) and reviewed above. Klineberg concludes that whites generally perform better with respect to the speed indices but not on error indices, and argues for an environmental or cultural explanation for the difference. It is difficult to determine the relevance of this research to the present GATB issues.

Vane and Kessler (1964) analyzed the long-term reliability and validity of the Goodenough Draw-a-Man test using samples of black and white children in kindergarten and early elementary grades. Students were tested once a year from kindergarten through the third grade on the Draw-a-Man test, and they were tested on the Stanford-Binet and the American Achievement Test in the third grade. Sample size was 280 for the first testing, but was reduced to 112 by the third grade and only 98 for the correlations of the Draw-a-Man with the other tests. Black and white sample sizes were not reported. The correlational analysis showed the usual pattern of higher correlations between scores on tests given closer in time for the Draw-a-Man (correlations ranged in value from .65 to .80). Correlations of the Draw-a-Man scores with scores on the Stanford-Binet and the achievement test ranged from .40 to .58. No separate correlations by race are reported. The only results reported by race were (p. 488), "The results indicated that the Negro children showed only a slight decline in the Draw-a-Man IQ during the four years, ranging from 98.4 in kindergarten to 95.8 in the third grade. The white children showed a larger decline ranging from 113.0 in kindergarten to 105.3 in the third grade. This decline is considered to be due to the nature of the Draw-a-Man Tests, which has a narrow range that makes it difficult for a bright child to maintain a high IQ as he increases in age." These results, again, do not appear very relevant to the speededness of GATB power tests of adults.

In summary, the literature reviewed just above and cited in Hartigan & Wigdor (1989) appears to underscore their own characterization as "admittedly speculative" the racial or socioeconomic correlates with the speededness of the GATB. Much of the cited research deals with populations that are much younger than the GATB population, uses tests that are quite dissimilar to the GATB, or is based on data that were collected before World War II. None of the research focused directly on speededness of tests and the differential effects of familiarity or practice for race or socioeconomic status, except for Klineberg (1928). The more recent studies focused more directly on this issue, and reviewed in the section just above, offer, at best, mixed support for this speculation.

Summary and Conclusions

Four of the tests on the GATB are intended to measure constructs normally thought of as measured with power tests: Computation, 3-D Space, Vocabulary, and Arithmetic Reasoning. The National Research Council's review of the GATB (Hartigan and Wigdor, 1989) pointed out that these four tests are highly speeded and expressed concerns that, therefore, the meaning of the constructs may be different from the meaning conventionally attached to those constructs. Allied concerns were the possible differential effects of speed on scores achieved by racial or ethnic groups, perhaps because of adverse psychological reactions by such groups to speeded tests. I have attempted to review literature relevant to these concerns and have provided summaries for each of the prior sections. Here, I briefly summarize the results and present some conclusions.

First, the definitions of speededness vary and different definitions meet different applied and research needs (Rindler, 1979). The methods divide into two primary types: those based on a single administration of a test and those based on multiple administrations of tests. The single-administration methods primarily address the question, "how speeded is the test?" while the second set of methods answer the question, "does it make any difference, in terms of ordering examinees, if the test is given under normal conditions (presumably speeded) and under power conditions?" For purposes of evaluating the construct validity of the GATB, the multiple administration methods are most appropriate and should be used in any research intended to evaluate the similarities of GATB tests administered under differing conditions of speededness. The single administration methods are useful, however, for characterizing the degree of speededness that is obtained for any particular administration of a test; in particular, the graphic method, showing the percentage of examinees completing each item on the test, seems most straightforward and interpretable. These methods can be and are routinely used to describe test performance. Obviously, the multiple administration methods are neither feasible nor even desirable in routine operations; but, as already noted, they are most desirable for evaluating the similarity of constructs measured by speeded and power versions of tests.

Expert opinion and research strongly support the view that scores on speeded power tests, like the GATB power tests, are influenced by both a speed component and a "power"component. This latter component usually being thought of as the dominant dimension the test was designed to measure (verbal ability, etc.). However, the relative influence of these two components varies with a number of factors, including prominently, but not necessarily limited to, the actual degree of speededness inherent in the time limit used, the type and difficulty of the items, the ability level of the examinees, and the instructions/scoring methods used. It appears that there is no accurate, analytic method of forecasting the influence of speed on scores for a given test in a given administration condition. Research must be conducted on the appropriate levels of speededness and examinee populations to make this determination. It does appear that ordering items by difficulty operates to reduce the influence of speed on a test's scores, compared to the same test when items are not given in that order.

Regarding effects on criterion-related validity, no clear-cut conclusions appear reachable. Some authors call for the same mix of power and speed on a test as exists for the criterion that is to be predicted, but this appears to be limited to cases where a single job (or group of highly similar jobs) has a fairly specific, particular criterion that is to be predicted; such a requirement seems

much less appropriate for tests in a battery that is to be used for a large number of jobs with a fairly general, broad-brush criterion, such as supervisory ratings. If "g," or general ability, is viewed as accounting for almost all of a given test's predictive validity, then reducing the speededness of a test, if it operates to reduce its loading on "g," would operate to reduce its predictive validity across all jobs. On the other hand, if the loading on the dominant dimension -- at least largely made up of "g" -- does not decline with a reduction in speededness, then no reduction in predictive validity would be expected. The very little empirical research on this issue seems to indicate that changes in the speededness of tests, within some fairly small range of feasible time limits, do change validity coefficients, but not substantially so. Related research shows that power tests could be substantially shortened without sacrificing validity. Almost all of this research, however, was conducted on student populations using grades as criteria or on non-student populations using other standardized tests as criteria; none of it used job performance criteria. The conclusion is, for the adult job performance arena, that research specific to the tests and criteria of concern is the firmest ground for reaching conclusions about the effects of changes in speededness on criterion-related validities.

On practical grounds, experts have noted that scores on speeded tests are much more sensitive than score on power tests to changes in administration conditions, whether intended (as in changes in answer sheets) or unintended (as in deviations from time limits). Also, it is easier to adapt testing conditions for disabled examinees if the standard testing conditions are power conditions.

Considerable attention was given to analyses showing relationships between the ASVAB and GATB tests in a recent sample of military recruits, since the ASVAB power tests measure constructs similar to the GATB and the population sampled is similar to the GATB, except for age distribution. These analyses indicate that scores on the GATB tests are more influenced by speed than are scores on their complementary ASVAB tests, but not unduly so in the cases of Arithmetic Reasoning, Vocabulary, and 3-D Space. Computation appears to be fairly strongly influenced by speed; it appears to measure a construct fairly similar to ASVAB Number Operations as well as constructs measured by Arithmetic Reasoning. This appears to be due to the fact that GATB Computation measures the four basic mathematical operations, as does ASVAB Number Operations, but with items of varying difficulty, arranged in order of difficulty. In contrast, all of the Number Operations items are very easy, if attempted. I believe the fact that all the GATB speeded power tests have items arranged in order of difficulty operates to lessen the influence of speed on their scores, as evidenced by these analyses.

A number of fairly recent studies that attempted to identify the differential effects of speed on race/ethnic and sex subgroups were reviewed. For the most part, these used college populations and tests that are much less speeded than the GATB tests, even when the experimental conditions were manipulated to increase speededness. Still, the evidence is not strong that reducing the amount of speededness will have a beneficial effect on scores for blacks or females (relative to whites and males). My review of the earlier studies that led to Hartigan & Wigdor's speculation in this regard indicates that this literature is not highly relevant to the populations and types of tests of concern with the GATB.

A number of recommendations seem to fall out of this set of conclusions:

- Definitions of speededness differ and are more or less useful for different operational and research contexts; the ETS "rules-of-thumb" and related graphic depictions (Donlon, 1973, 1975) should be used on a routine basis to characterize the power/speed nature of tests while the *tau* (Cronbach and Warrington, 1951) measures requiring multiple test administrations should be used to make comparisons of speeded and unspeeded versions of tests in research settings.

- Expert opinion and empirical research lead to the conclusion that scores on speeded power tests are influenced by both speed and power components, but the relative influences of these components cannot be analytically determined in a general way; therefore, research using the tests, administration conditions, and populations of most concern must be done to determine the comparability of more- and less-speeded versions of tests in terms of construct validity, adverse impact, and criterion-related validity.

- Research conducted in the GATB examinee population directed at the comparability of constructs measured by the present GATB tests and GATB tests given in much less speeded, if not completely power, conditions would provide crucial evidence about the likelihood that criterion-related validities from the past uses of the GATB can be generalized to less-speeded versions; if correlations between standard and less-speeded forms are nearly equal in value to the parallel-form reliabilities for the GATB, then a strong case can be made for linkage to past validity studies.

- The same types of research could provide information about the levels of score change for ethnic/race, sex, and age groups (i.e., adverse impact), but would not provide evidence capable of evaluating changes in regressions of tests on job performance criteria; additional research including the collection of job performance criteria for persons tested on standard and less-speeded versions of the GATB is required to directly evaluate differential validity and fairness for the two versions.

- Available research does provide 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 blacks 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); this argues for conducting the research outlined in the two points above.

- Resources are limited for research. Therefore, I recommend that research designs for accomplishing the above goals take into account the likely changes that already are planned for the GATB and incorporate them insofar as possible without limiting the capability to make the desired inferences from research results; in particular, since operational administration time is at a premium, I recommend that the less-speeded research versions of the GATB use fewer items and the standard time limits, but

with the caveat that the number of items should not be fewer than that required to adequately represent the full range of item difficulty and content. This will insure that the shorter form is as parallel as possible with the longer, standard form. This has the advantage of closely mirroring the likely operational setting, and anchors the end of the continuum with regard to number of items to use in a power setting. Extrapolations about time required for administration of longer forms and expected levels of internal consistency reliability can be made from data collected using this test. Research using all, or a large proportion of the present number of items in a power setting would require impractically long testing sessions, in either research or operational settings.

R E F E R E N C E S

Ackerman, P. L., & Humphreys, L. G. (1990). Individual differences theory in industrial and organizational psychology. In M.D. Dunnette and L.M. Hough (Eds.) Handbook of industrial and organizational psychology (Second Edition Volume I). Palo Alto, CA: Consulting Psychologists Press, Inc.

Baxter, B. (1941). An experimental analysis of the contributions of speed and level in an intelligence test. Journal of Educational Psychology, 32, 285-296.

Bejar, I. I. (1985). Test speededness under number-right scoring: An analysis of the test of English as a foreign language (ETS-RR-85-11). Princeton, NJ: Educational Testing Service.

Bell, R. & Lumsden, J. (1980). Test length and validity. Applied Psychological Measurement, 4, 165-170.

Boese, R. R. (1993). An argument for speeded tests over power tests in assessing the less-educated. Southern Assessment Research & Development Center: Raleigh, NC.

Boger, J. H. (1952). An experimental study of the effects of perceptual training on group IQ test scores of elementary pupils in rural ungraded schools. Journal of educational Research, 46, 42-52.

Cronbach, L. J. (1990). Essentials of psychological testing. New York, NY: Harper Collins Publishers, Inc.

Cronbach, L. J., & Warrington, W. G. (1951). Time limit tests: Estimating their reliability and degree of speeding. Psychometrika, 14, 167-188.

Davidson, W. M., & Carroll, J. B. (1945). Speed and level components in time-limit scores: A factor analysis. Educational and Psychological Measurement, 411-427.

Davey, T. (1990, April). Modeling timed test performance. A paper presented at the Annual Meeting of the American Educational Research Association, Boston, M.A.

Donlon, T. F. (1973). Establishing time limits for tests. A paper presented at the Annual Meeting of the Northeast Educational Research Association, Ellenville, NY.

Donlon, T. F. (1975, April). Estimating rate-of-work parameters in the assessment of test speededness. A paper presented at the Annual Meeting of the National Council on Measurement in Education, Washington, D.C.

Donlon, T. F. (1977). Sex differences in test speededness on the Scholastic Aptitude Test. A paper presented at the Annual Meeting of the Northeastern Educational Research Association, Ellenville, NY.

Donlon, T. F. (1979, April). Time for review as an aspect of test speededness. A paper presented at the Annual Meeting of the New England Educational Research Organization, Provincetown, MA.

Douglas, J. B. (1981, April). Item bias, test speededness, and Rasch tests of fit. A paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.

Dubin, J. A., Osburn, H., & Winick. (1969). Speed and practice: Effects on Negro and White test performances. Journal of Applied Psychology, 53, 19-23.

Eagleson, O. W. (1937). Comparative studies of white and negro subjects in learning to discriminate visual magnitude. Journal of Psychology, 4, 167-197.

Evans, F. R., & Reilly, R. R. (1972). A study of speededness as a source of test bias. Journal of Educational Measurement, 9, 123-131.

Evans, F. R. & Reilly, R. R. (1973). A study of test speededness as a potential source of bias in the quantitative score of the admission test for graduate study in business. Research in Higher Education, 1, 173-183.

Gulliksen, H. (1950). The reliability of speeded tests. Psychometrika, 15, 259-269.

Hartigan, J. A., & Wigdor, A. K. (Eds.) (1989). Fairness in employment testing. Washington, D.C.: National Academy Press.

Jaeger, R. M., Linn, R. L., & Tesh, A. S. (1989). A synthesis of research on some psychometric properties of the GATB. In J.A. Hartigan and A. K. Wigdor (Eds.) Fairness in employment testing. Washington, D.C.: National Academy Press.

Jensen, A. R. (1983, August). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. A paper presented at the 91st Annual Meeting of the American Psychological Association, Anaheim, CA.

Katzenmeyer, W. G. (1962). Social interaction and differences in intelligence test performance of negro and white elementary school pupils. Doctoral dissertation, Duke University.

Kendall, L. M. (1962). An investigation of hypotheses regarding the way adjustment of time limits affects validity. A paper presented at the 33rd Meeting of Eastern Psychological Association, Atlantic City, NJ.

Kendall, L. M. (1964). The effects of varying time limits on test validity. Educational and Psychological Measurement, 24, 789-800.

Kingston, N. (1984, April). Analysis of shifts in scale and construct through the use of repeater data. A paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Klineberg, O. (1928). An experimental study of speed and other factors in racial differences. Archives of Psychology, 15, 109-123.

Lin, M. H. (1986). The impact of time limits on test behaviors. A paper presented at the 67th Annual Meeting of the American Educational Research Association, San Francisco, CA.

Linn, R. (1982). Ability testing: Individual differences, prediction, and differential prediction. In Wigdor, A. K. and Garner, W. R. (Eds.) Ability testing: Uses, consequences, and controversies. Washington, D.C.: National Academy Press.

Lohman, D. (1979). Spatial ability: A review and reanalysis of the correlational literature (Tech. Rep. No. 8). Aptitude Research Project, Stanford University, School of Education, Stanford, CA.

Lord, F. M. (1956). A study of speed factors in tests and academic grades. Psychometrika, 21, 31-50.

Nester, M. A. (1993). Psychometric testing and reasonable accommodation for persons with disabilities. Rehabilitation Psychology, 38, 75-85.

Peterson, N. G., Russell, T. L., Hallam, G., Hough, L. M., Owens-Kurtz, C., Gialluca, K., & Kerwin, K. (1990). Analysis of the experimental predictor battery: LV sample. In J.P. Campbell and L. M. Zook (Eds.), Building and retaining the career force: New procedures for accessing and assigning Army enlisted personnel, Annual Report, 1990 Fiscal Year (ARI FR-PRD-90-6). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Powers, D. E., Swinton, S. S. & Carlson, A. B. (1977). A factor analytic study of the GRE aptitude test (GREB No. 75-11P), Princeton, NJ: Educational Testing Service.

Ree, M. J., & Earles, J. A. (1991a, April). Estimating psychometric g: An application of the Wilk's theorem. A paper presented at the Annual Meeting of the American Psychological Association, San Francisco, CA.

Ree, M. J., & Earles, J. A. (1991b). Predicting training success: Not much more than *g*. Personnel Psychology, 44, 321-332.

Ree, M. J., & Wegner, T. G. (1990). Correcting differences in answer sheets for the 1980 Armed Services Vocational Aptitude Battery reference population. Military Psychology, 2, 157-170.

Rindler, S. E. (1979). Pitfalls in assessing test speededness. Journal of Educational Measurement, 16, 261-270.

Schmitt, A. P. & Bleistein, C. A. Factors affecting differential item functioning for black examinees on Scholastic Aptitude Test analogy items (Research Report 87-23). Princeton, NJ: Educational Testing Service.

Stafford, R. E. (1971). The speededness quotient: A new descriptive statistic for tests. Journal of Educational Measurement, 8, 275-277.

Swarthout, D. (1988). An investigation of the effect of correction for guessing on General Aptitude Test Battery scores (AWP #3822). Detroit, MI: Northern Test Development Field Center.

U.S. Employment Service (1985). Manual for the USES General Aptitude Test Battery: Reliability and comparability forms C and D. Washington, D.C.: U.S. Department of Labor Employment and Training Administration, U.S. Employment Service.

Utah State Department of Employment Security (1981). The speed-power study of the USES Basic Occupational Literacy Test (BOLT) analysis and report. Salt Lake City: Western Test Development Field Center, Utah State Department of Employment Security.

Vane, J. R., & Kessler, R. T. (1964). The Goodenough Draw-a-Man Test: long-term reliability and validity. Journal of Clinical Psychology, 20, 487-488.

Waters, B. K., Barnes, J. D., Foley, P., Steinhaus, S. D., & Brown, D. C. (1988). Estimating the reading skills of military applicants: Development of an ASVAB to RGI conversion table. Alexandria, VA: Human Resources Research Organization.

Wesman, A. G. (1960). Some effects of speed in test use. Educational and Psychological Measurement, 20, 267-274.

Wigdor, A. K. & Garner, W. R. (Eds.) (1982). Ability testing: Uses, consequences, and controversies (Part II: Documentation Section). Washington, D.C.: National Academy Press.

Wild, C. & Durso, R. (1979). Effects of increased test-taking time on test scores by ethnic groups, age, and sex (GREB-76-6R). Princeton, NJ: Educational Testing Service.

Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). Testing handicapped people. Boston, MA: Allyn and Bacon, Inc.

Wise, L. L. & McDaniel, M. A. (1991, August). Cognitive factors in the ASVAB and the GATB. A paper presented at the Annual Convention of the American Psychological Association, San Francisco, CA.