# DATA QUALITY ANALYSIS OF THE ARMED SERVICES VOCATIONAL APTITUDE BATTERY

R. Darrell Bock
University of Chicago

and

Robert J. Mislevy
National Opinion Reseach Center

May, 1981

# SUMMARY AND CONCLUSIONS

## PURPOSE OF THE STUDY

This report describes psychometric analyses of the responses of the National Longitudinal Study (NLS) panel to the Armed Services Vocational Aptitude Battery (ASVAB). These data were collected for the purpose of constructing national norms for the aptitudes of the national youth population, ages 16 to 23. The analyses reported herein concern several aspects of the suitability of the data for this purpose.

The central question we address is this: To what extent do the scores of the subjects accurately reflect their abilities? To answer it, we examine the data for possible sources of invalidity: random responses, carelessness, inappropriate levels of test difficulty, ambiguous wording of questions, and cultural bias, among others.

## RESULTS

Subtest reliabilities (the degree to which items within a subtest cooperate to measure the content area) fall within the range of reliabilities of commercially-published aptitude tests. Measurement is of acceptable precision across the range of abilities of the subjects, with all tests most precise in the region from the 25-th to the 75-th percentile.

The incidence of "unexpected" responses (low-scorers answering hard questions correctly and high-scorers answering easy questions incorrectly) was lower than might be expected on multiple-choice tests like the ASVAB subtests. Random guessing was rare in the more technical subtests, more frequent (but still low in absolute terms) in the more academic tests. Little evidence of carelessness on the part of high-scoring subjects was observed, with the notable exception of the final item in the Paragraph Comprehension test. Because of its position in the test booklet, it was inadvertantly omitted by more than a third of the subjects.

The extent of guessing can be quantified by saying that even the lowest-scoring subjects answered the hardest items correctly with about one-in-ten chances. This level is significantly lower than the one-in-four chances that would result from random guessing among the choices offered, indicating that subjects were on the whole taking the care to interact meaningfully with the items. In this respect the derived norms will be of higher quality at the lower ends of the ability distribution than might be expected with tests of this kind.

Levels of guessing were slightly but inconsequentially lower for males than for females. Levels of guessing were also consistent across racial/ethnic groups, given the ability levels of subjects.

Test question bias (some questions being relatively easier for members of one cultural group than another) was not apparent,

with two minor exceptions. (1) The single score reported in General Science tends to over-estimate the Health Science abilities of males but under-estimate their Physical Science abilities, while the reverse is true for females. (2) The single score reported for Word Knowledge tends to over-estimate Hispanic and Black familiarity with "literary" words but under-estimate familiarity with words in popular usage, while the reverse is true for Whites. Otherwise, it may be concluded that ASVAB scores have essentially the same meaning across sex and racial/ethnic groups.

Violations of test administration standards, as defined at the level of NORC area supervisors, were not apparent.

## CONCLUSIONS

Data from responses to the ASVAB are free from major defects such as high levels of guessing or carelessness, inappropriate levels of difficulty, and cultural test-question bias. Accordingly, the level of reliability of the resulting scores from the NLS panel can be expected to equal or surpass that of comparable commercially-published tests of aptitude and achievement.

CONTENTS

## STATEMENT OF PURPOSE AND METHODS

This paper is the final report of data-quality analyses of the responses of the National Longitudinal Study panel to the Armed Services Vocational Aptitude Battery (ASVAB), Form 8A. The analyses and our comments address the suitability of these data for the construction of national norms for the current youth population, ages 16 through 23.

One important aspect of data quality is the degree to which the distributions of number-right scores derived from these data approximate those which would be obtained by testing every member of the youth population. This question is addressed in reports on the sampling design (Frankel, 1981) and on test admininistration procedures (McWilliams, 1980), and by the sampling-theoretic standard errors for percentile points of the final distributions (Frankel, 1981).

Number-right distributions are of interest, however, only insofar as they reflect the abilities of the subjects. A second and separate aspect of data quality is the degree to which test scores are reliable and accurate measures of ability for subjects in the youth population. This question is addressed in the present report, a psychometric analysis of the ASVAB subtests with regard to the responses of the National Longitudinal Study panel.

To this end, a model-fitting approach is taken. Data from each of the ASVAB subtests are fit to an item response curve (IRC) model which quantifies expected patterns of subject/item interaction. The parameters of the models summarize the operating characteristics of the tests; departures from the models highlight disturbances and anomalies in the data. These features of the data are discussed, with particular attention to their implications for the construction, the interpretation, and the eventual use of number-right norms for the youth population.

# PART I

# DATA-QUALITY ANALYSIS WITH ITEM RESPONSE CURVE MODELS

In order to discover and describe the manner in which each of the ten ASVAB subtests operates to measure subject abilities, each subtest has been fit separately to an item response curve (IRC) psychometric model. This part of our report briefly describes the IRC models that are used: a 3-parameter logistic model for the power (unspeeded) tests and a Poisson model for the speeded subtests.

## ITEM RESPONSE CURVE MODELS AND DATA ANALYSIS

The use of item response curve (IRC) psychometric models has long been supported and developed by the armed services, beginning in the 1960's with the work of Allan Birnbaum and continuing in the present with the work of Fumiko Samejima, James McBride, and others. The principal advantages of these models in selection and classification stem from the fact that subject abilities are estimated conditionally on the items the subject has been presented. Tailored testing and linking of test forms, difficult problems for the methods of classical test theory, become quite tractible.

Our choice of IRC models for the analysis of the ASVAB has been motivated from a different point of view. In parameterizing the expected patterns of subject/item interactions, IRC models are excellent vehicles for data analysis of mental test data. To the extent that an IRC explains the data, it expresses the operations of the test and of the items as measuring devices. The amount of information provided about subjects at various levels of ability, by individual items and by the test as a whole, are easily obtained. Departures from the model highlight disturbances and anomalies in the data, such as ambiguities in test items or random guessing behavior of subjects.

Eight of the ten ASVAB subtests are power tests, in which performance depends mainly on subjects' knowledge or reasoning abilities rather than time limitations. To these tests, a Birnbaum 3-parameter logistic IRC (Birnbaum, 1968) was fit with the BILOG computer program (Bock & Mislevy, 1980), using the fixed-effects algorithm outlined in Appendix A.

The remaining two subtests are speeded tests, in which performance depends mainly on subjects' speed and accuracy at simple tasks in a limited amount of time. To both of these tests, a Poisson IRC for speeded tests (Rasch, 1960) was fit.

The following sections review these models in turn.

## A 3-PARAMETER LOGISTIC MODEL FOR POWER SUBTESTS

An unspeeded (power) mental test demands knowledge or reasoning ability from a subject. Although time limits are set and observed, it is assumed that subjects would answer few additional items correctly under more generous time limits. Eight of the ten ASVAB subtests are unspeeded tests: (1) General Science, (2) Arithmetic Reasoning, (3) Word Knowledge, (4) Paragraph Comprehension, (5) Auto & Shop Information, (6) Mathematics Knowledge, (7) Mechanical Comprehension, and (8) Electronics Information.

The model used to analyze the eight power subtests of the ASVAB is based on Birnbaum's 3-parameter logistic item response curve (IRC) model. The IRC provides a statistical model for the probability of a given subject responding correctly to a given test item. Included in this model is the subject's scale score and parameters that characterize the difficulty and reliability of the item. When the model is fitted to the data, it is capable of accounting for the facts that--

(1) Some subjects perform better than others on the items in the subtest.

(2) Some items in the subtest are easier than others.

(3) Some items measure the underlying ability more precisely than others.

(4) Because the test items are multiple choice, subjects can

occasionally answer any item correctly by guessing.

The scale of ability along which persons are measured is defined explicitly by the locations of the items:

The ability, or scale score, of Subject i ($\theta_i$) is his location along the scale.

The location of Item j on the scale is called its threshold ($b_j$). Items' thresholds indicate their relative probabilities of being answered correctly by a person drawn at random from the target population. A subject located at the threshold of Item j would have a 50-50 chance of answering Item j correctly.

The dispersion parameter of Item j ($s_j$) is inversely related to the reliability with which item j measures ability.

Finally, the lower asymptote of Item j ($c_j$) is the probability of a correct response from even the subjects of lowest ability. Lower asymptotes may be useful during the estimation of item parameters, freeing threshold and dispersion estimates from the effects of random guessing. Because guessing behavior differs from one subject to another, however, final estimates of subject ability are instead based on a a robust procedure that does not use the lower asymptotes (see Appendix B).

The exact value of the probability of a correct response to Item j from Subject i is given by the following function of $\theta_i$, $b_j$, $s_j$, and $c_j$:

$$P_{ij} = c_j + (1-c_j)\ \exp(Z_{ij}) \ /\ [1.0 + \exp(Z_{ij})],$$

where

$$z_{ij} = (\theta_i - b_j)/s_j$$

and exp(x) denotes the raising of the base of the natural logarithms to the xth power.

The origin and scale of the ability variable may be chosen arbitrarily. In our analyses, the scale has been set so that the mean of subject abilities in the youth population is zero and the variance is one, after correction for measurement error variation.

The amount of information that Item j provides about subjects at various levels of ability is given by its information curve, which for the logistic item response model is the derivative of the IRC (i.e., the slope of the curve at each point on the scale). The information provided by Item j for ability level $\theta$ is given by

$$I_j(\theta) = \{[1+\exp(z_{\theta j})]^{-1}/(s_j)^2\}\{1-c_j\}\{[1+\exp(z_{\theta j} - \ln(c_j)]^{-1}\}$$

where

$$z_{\theta j} = (\theta - b_j)/s_j.$$

It may be inferred that an item provides most information about subjects whose abilities lie in the neighborhood of its threshold. It may also be inferred that, at their most informative points, items with large dispersions provide less information than items with small dispersions.

The total amount of information provided by a collection of items is given by the sum of their individual item information curves, i.e., the test information curve. When the method of maximum likelihood is used to estimate subjects' abilities from their responses to items with known parameters, the standard error

of estimation will be the square root of the reciprocal of the total test information at the estimated ability. An examination of a test's information curve, then, may be used to examine the levels of measurement precision that are attained at various points along the ability scale. Such analyses will be performed for the power tests of the ASVAB.

We note in passing that simpler item-response models with constant item dispersions and/or lower asymptotes of zero have also been proposed (see Lord & Novick, 1968, and Andersen, 1980). These models offer considerable conceptual and technical advantages in applied settings, and are worthwhile goals during test construction. Preliminary analyses of the data from the ASVAB subtests indicated that these simpler models fit the data poorly, however. The 3-parameter model has been adopted as better suited to the task at hand, namely, data analysis of responses to existing tests.

## A POISSON MODEL FOR SPEEDED SUBTESTS

The items of the speeded subtests of the ASVAB, Numerical Operations and Coding Speed, require little knowledge or reasoning ability. If time were not restricted so severely, almost every subject would answer almost every item correctly. Performance in these subtests, then, places a premium on subjects' speed and accuracy.

The model used to analyze these subtests is based on Rasch's (1960) model for speeded tests. Two assumptions are necessary. First, item content within a test is considered to be homogeneous --an assumption fairly well satisfied for Numerical Operations and almost perfectly satisfied for Coding Speed. Second, a subtest is treated as if it were infinitely long--an assumption also well-satisfied in both tests, as time restrictions are sufficiently strict to prevent all but a few subjects from reaching the end of either subtest.

Briefly, the justification of the model is as follows: It is supposed that (1) the probability of Subject i responding correctly to an item during any small time interval $\Delta t$ depends only on the length of the interval, and that (2) as $\Delta t$ approaches zero the possible outcomes are essentially either <u>one</u> correct response, with probability $P_i$, or <u>zero</u> correct responses, with probability $(1-P_i)$. Then the probability that Subject i will correctly answer $R_i$ items over the course of $N_i$ time intervals is approximated by the following Poisson model:

$$\text{Prob}(R_i|N_i) = \frac{(N_i\ P_i)^{R_i}}{R_i!} \exp(-N_i\ P_i),$$

where $(N_i\ P_i)$ may be interpreted as the expected number of correct responses.

Rasch's original model expressed $P_i$ as the product of a term for subject ability and a term for test difficulty, which is essential for the comparison or linking of multiple tests.

Because our attention is focused on only one subtest at a time, this separation is not necessary. Furthermore, because every subject is allotted the same amount of time, $N_i$ may be absorbed with $P_i$ into a single parameter $\theta_i$, the ability of Subject i with respect to the given subtest and standard time limitations.

Under the assumptions outlined above, the maximum likelihood estimate of $\theta_i$ is simply $R_i$ and its standard error of estimation is the square root of $R_i$. It is customary to analyze the logs of number-correct scores of speeded tests, and for this reason we have approximated $\theta_i$ by $(R_i + 0.5)$. Subsequent analyses of score distributions revealed that the number-correct scores have approximately normal distributions in the youth population, as do the ability estimates from the power tests, while the logs of the number-right scores do not. The number-right metric has been retained, then, so that score units may be more comparable for the power subtests and the speeded subtests. Like the power tests, the speeded tests have been standardized so that the mean of the population is zero and the variance is one, after correction for measurement error variance.

A facsimile of a test information curve is obtained by plotting the squared reciprocals of standard errors against ability estimates. While the form of the model dictates decreasing information with increasing ability, it is useful to examine the test information curve in relation to the population distibution of ability. Moreover, it will be possible to derive statements about local and overall test reliabilities.

## PART II

## OVERALL FIT OF IRC MODELS TO ASVAB POWER SUBTESTS

The 3-parameter logistic model used to analyze the ASVAB power subtests characterizes each item individually, in terms of its difficulty, its reliability, and the level of random guessing it provokes. These parameters attempt to express the item's relationship to the underlying ability variable. The success of this characterization can be examined by the degree to which subjects at various levels of ability respond in accord with the modelled patterns.

The BILOG computer program provides an overall Chi-square value for subtest fit, the sum of the Chi-squares for item fit. The total fit Chi-square may be considered as a summary statistic of the fit of the IRC model to the data. (More useful for diagnosis and quality control are indices of fit for individual items and for individual subjects. These analyses are discussed in Part IV and Part V respectively.) The total fit values are shown as Table 1.

It may be seen that for each power subtest, the Chi-square is about twice the number of degrees of freedom. This fit of IRC models to data from attainment/achievement tests is typical in our

TABLE 1
FIT CHI-SQUARES OF IRC MODEL TO POWER SUBTESTS

| SUBTEST | CHI-SQUARE | DEGREES OF FREEDOM | RATIO |
|---|---|---|---|
| GENERAL SCIENCE | 315.97 | 171.00 | 1.85 |
| ARITHMETIC REASONING | 476.45 | 208.00 | 2.29 |
| WORD KNOWLEDGE | 499.87 | 220.00 | 2.27 |
| PARAGRAPH COMPREHENSION | 173.92 | 79.00 | 2.20 |
| AUTO & SHOP INFORMATION | 360.56 | 172.00 | 2.10 |
| MATHEMATICS KNOWLEDGE | 418.00 | 173.00 | 2.42 |
| MECHANICAL COMPREHENSION | 360.78 | 178.00 | 2.03 |
| ELECTRONICS INFORMATION | 244.90 | 135.00 | 1.81 |

NOTE: DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR.

experience. These Chi-squares cannot be taken strictly at face value, however, because the NLS panel is not a simple random sample from the youth population; it is a stratified sample, including oversampling in certain groups with low average abilities. If it is appropriate to adjust the subtest fit indices in accordance with calculated design effects (and it is not clear as yet whether this is so), then the fit of the IRC models is extremely good; design effects for the ASVAB scores in the NLS study are about 2.0 (Frankel, 1981). At any rate, the fit is certainly good enough to justify interpretation of parameter estimates.

## CONCLUSIONS

The fit of the items in the ASVAB power tests is comparable to that of other tests of educational achievement or aptitude in our experience. The interpretation of item parameters and statements derived from them is justifiable.

## PART III

## SUBTEST INFORMATION AND RELIABILITY

The concern of Part III is the extent to which the ASVAB subtests provide accurate measures for subjects at various levels of ability. The test information curve for each ASVAB subtest indicates the precision of measurement which the subtest can bring to bear at any given point along the ability scale. The square root of the reciprocal of the information at that ability is the standard error of estimation. A test information curve can reveal whether the focus of measuring power of a test is consistent with its intended use.

### METHODOLOGY

Unlike the classical test theory concept of reliability, test information need not refer to the distribution of ability in a particular population of subjects. However, when the distribution of a target population is known or can be estimated, as is the case with the national youth population with regard to the ASVAB subtests, statements about test reliabilities can be derived from the information curves. In addition to traditional population

reliability coefficients, it is possible to determine the relative precision of measurement at various points of the ability distribution.

Reliability is typically defined as the ratio of "true" variation in a population to the sum of "true" and measurement error variation. An estimate of population reliability may be obtained by first integrating the error variation over the population ability distribution to obtain an average value, then computing the ratio of true-score to true-score plus average error variation.

The relative precision of measurement of a subtest can be determined for any point along the ability scale by dividing the standard error of estimation for the point at which measurement is most precise by the standard error of estimation at the point in question. The reciprocal of this quotient is interpretable as well, as the multiple by which confidence intervals around the point in question are longer than confidence intervals with the same level of probability around the point of greatest precision.

## RESULTS

The test information curves and associated standard error curves for each of the ten ASVAB subtests are presented in Appendix C. These curves were computed under the assumptions of the models described in Part I. In particular, it should be recalled that the ability estimates within each test were set to have a

population mean of zero and a true-score variance of one (Note: these values take into account the subject weights from the NLS sampling design). Subsequent analyses of the ability distributions reveal them to be approximately normal in shape, so that the precision of measurement available at percentile points of interest may be approximated from the information curves, using the transformation of normal deviates to percentage points.

Table 2 presents standard errors of estimation associated with abilities of minus two, minus one, zero, plus one, and plus two standard deviations from the population mean in each of the ASVAB subtests, along with the corresponding averages over the population distributions. Under the assumption that the ability distributions are normal, population test reliabilities have been computed using a ten-point Gauss-Hermite quadrature.

Both test information and reliability increase as the number of items in a test increases. Higher reliabilities are to be expected, therefore, from longer tests like Word Knowledge than from shorter tests like Paragraph Comprehension. Population average reliabilities, however, are not neceassarily the most relevant indices for evaluating a test.

Paragraph Comprehension, as a first example, is a very short and therefore less informative test. Population reliability appears to be low (.68) but measurement precision at one standard deviation below the mean is surprisingly high; Paragraph Comprehension is just as precise in this region as Arithmetic Reasoning, which has twice as many items!

TABLE 2
SUBTEST RELIABILITIES AND STANDARD ERRORS OF MEASUREMENT

| TEST NAME | ITEMS | STANDARD ERRORS OF MEASUREMENT | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | -2 SD | -1 SD | MEAN | +1 SD | +2 SD | AVERAGE | RELIABILITY |
| GENERAL SCIENCE | 25 | .59 | .31 | .33 | .39 | .67 | .40 | .86 |
| ARITHMETIC REASONING | 30 | .83 | .33 | .23 | .31 | .59 | .39 | .87 |
| WORD KNOWLEDGE | 35 | .37 | .20 | .23 | .44 | .94 | .40 | .86 |
| PARAGRAPH COMPREHENSION | 15 | .69 | .33 | .45 | .78 | 1.56 | .69 | .68 |
| NUMERICAL OPERATIONS* | 50 | .39 | .52 | .64 | .73 | .77 | .64 | .71 |
| CODING SPEED* | 84 | .25 | .37 | .47 | .56 | .61 | .47 | .82 |
| AUTO & SHOP KNOWLEDGE | 25 | .96 | .50 | .25 | .31 | .72 | .45 | .83 |
| MATHEMATICS KNOWLEDGE | 25 | .89 | .20 | .25 | .29 | .61 | .44 | .84 |
| MECHANICAL COMPREHENSION | 25 | .82 | .42 | .33 | .42 | .65 | .45 | .83 |
| ELECTRONICS INFORMATION | 20 | .94 | .42 | .35 | .47 | .75 | .50 | .80 |

NOTES: * INDICATES SPEEDED SUBTESTS.
TRUE-SCORE VARIANCE IS 1.0 IN ALL SUBTESTS.

A typical item at this point along the Paragraph Comprehension ability scale (see, for instance, Item 4 or 13) requires a subject to identify the main idea of a short written passage. This means that Paragraph Comprehension scores can provide a precise answer to the specific question of whether a subject can or cannot acquire information from written material at the target level; they reliably distinguish readers from non-readers. They are less useful for comparing the abilities of readers who are above, say, one standard deviation above the mean. The Paragraph Comprehension subtest would seem well-suited for initial selection of subjects but poorly suited for placement of selected subjects into advanced training programs.

Auto & Shop Information, as a second example, focuses its measuring power a quarter of a standard deviation above the population mean. A typical item in this region of the scale, Item 20, concerns the fact that a vacuum line must be operated by either the carburetor or the intake manifold--clearly a higher level of performance than required by the Paragraph Comprehension items described above. Auto & Shop Information would seem better suited for placement of selected subjects into training programs than for initial selection.

As dictated by the form and the scaling of the IRC used for the speeded tests, both Numerical Operations and Coding Speed are most informative for the subjects of lowest abilities. It may be

seen that the level of information does not drop rapidly in either test, so that measurement precision even two standard deviations above the populations means are satisfactory--standard errors of measurement of .77 and .61 respectively, in the same range of magnitude as corresponding values for the power subtests.

The most informative regions of each of the ASVAB power subtests are described below:

General Science provides most information in the region three-quarters of a standard deviation below the mean, at about the 25th percentile. An item in this region is Item 7; the subject must recognize that mushrooms, as opposed to three varieties of green plants, lack chlorophyll. The items in General Science cover a broad range of difficulty. As a result, the peak of the information curve is not as sharp as those of other tests, and quite reliable measures are obtained from about 1.5 standard deviations below the mean up to 2 standard deviations above the mean.

Arithmetic Reasoning provides most information in the region one quarter of a standard deviation below the mean, about the 40th percentile. Item 13, with a threshold in this region, tells the subject that a hunting trip yielded 3 rabbits, 9 fish, and 4 squirrels, then asks what fraction of the total catch the squirrels represent. Adequately reliable measurement is obtained from 1.5 standard deviations below the mean to 2.0 standard deviations above.

Word Knowledge, the longest ASVAB power test, provides most precise information three-quarters of a standard deviation below the mean, about the 25th percentile. An item in this region is Item 14, asking the meaning of "gouged." Information becomes less reliable above 1.5 standard deviations above the mean.

Paragraph Comprehension, the shortest power test, is most informative a full standard deviation below the mean, about the 16th percentile. As noted above, items in this region of the scale require the subject to identify the main idea of a short written passage. Information peaks here quite steeply, and falls off rapidly. Paragraph Comprehension scores may indicate that a given subject can acquire information from a written passage of the type presented, but they may not be able to tell how much better or worse he is than another subject who has also scored well on the test.

Auto & Shop Information provides most information in the region a quarter of a standard deviation above the mean, about the 60th percentile. A typical item from this region, as noted above, concerns the operation of vacuum lines. Information peaks fairly sharply; Auto & Shop Information scores are quite reliable for persons within one standard deviation from the mean.

Mathematics Knowledge provides most information in the region a half standard deviation above the mean, about the 70th percentile. One item in this region, Item 17, requires the subject to solve for a specified variable an algebraic expression involving

algebraic division but not exponentiation. The test is not very informative for subjects with abilities more than one standard deviation below the mean.

Mechanical Comprehension provides most information in the region a quarter of a standard deviation below the mean, about the 40th percentile. An item in this region, Item 7, concerns the direction of operation of a simple cam. The information peak is not sharp, and of all the power tests, Mechanical Comprehension spreads its measuring power most widely across the ability distribution.

Electronics Information also provides most information in the region a quarter of a standard deviation below the mean, about the 40th percentile. Item 6, located in this region, requires the subject to identify "capacitor" as a term directly related to electrical systems, as opposed to "lubricator", "gimbal", or "hypoid." This test is less informative about subjects more than 1.5 standard deviations below the mean.

## COMPARISON WITH A COMMERCIAL APTITUDE BATTERY

For a rough comparison of the reliability of the ASVAB subtests with commercially-published vocational aptitude tests, Table 3 presents the reliability coefficients of the ASVAB subtests and of the Differential Aptitude Tests (DAT). The DAT reliabilities are for a target population of the battery, namely, male high-school seniors. Because test lengths differ for the two bat-

TABLE 3
ASVAB AND DIFFERENTIAL APTITUDE TEST RELIABILITIES

| | SUBTEST NAME | # ITEMS | RELIABILITY OVERALL | RELIABILITY SINGLE ITEM |
|---|---|---|---|---|
| ASVAB | GENERAL SCIENCE | 25 | .86 | .20 |
| | ARITHMETIC REASONING | 30 | .87 | .17 |
| | WORD KNOWLEDGE | 35 | .86 | .15 |
| | PARAGRAPH COMPREHENSION | 15 | .68 | .12 |
| | AUTO & SHOP INFORMATION | 25 | .83 | .16 |
| | MATHEMATICS KNOWLEDGE | 25 | .84 | .17 |
| | MECHANICAL COMPREHENSION | 25 | .83 | .16 |
| | ELECTRONICS INFORMATION | 20 | .80 | .17 |
| DAT | VERBAL ABILITY | 50 | .95 | .28 |
| | NUMERICAL ABILITY | 40 | .93 | .25 |
| | MECHANICAL ABILITY | 70 | .91 | .13 |
| | SPATIAL ABILITY | 60 | .95 | .24 |
| | SPELLING | 100 | .96 | .19 |
| | LANGUAGE USAGE | 60 | .91 | .14 |

* DENOTES SPEEDED TEST

teries, all reliability coefficients are also reduced to single-item level by the Spearman-Brown formula. These calculations are not designed for use with speeded tests. Reliability coefficients for speeded tests are typically computed for parallel forms of a given test. Parallel forms were available for the DAT, but not for the ASVAB. For this reason, no comparisons of reliabilities are reported here for the speeded subtests.

The are a number of reasons that these two sets of reliability coefficients are not directly comparable: (1) The two test batteries are designed for different populations. (2) Even subtests with similar names may be measuring different abilities. (3) Reliability coefficients have been estimated in different ways for the two batteries--the ASVAB with our IRC methods, the DAT with split-half and alternate-form methods. (4) Different samples from different populations provided data for the estimation of reliability coefficients. (5) The DAT coefficients relate to the bounded scale of number-correct scores, while the ASVAB coefficients relate to an unbounded logistic scale.

It is apparent nevertheless that item for item, the ASVAB is about as reliable for the national youth population as the DAT is for its target population.

## CONCLUSIONS

The levels of reliability in the ASVAB tests, with respect

to the youth population ages 16 to 23, are comparable to those of commercial batteries of vocational aptitude. The items in each of the power subtests are well-targeted for abilities between, say, two standard deviations below the population mean to two standard deviations above the mean.

Targetting varies from one subtest to another: The shortest test, Paragraph Comprehension, provides precise information at a level about one standard deviation below the mean, but not as much information for subjects very far above the mean. Mathematics Knowledge, on the other hand, is more informative about subjects above the mean than below.

Subtests with relatively high precision for subjects with low abilities include General Science, Word Knowledge, Paragraph Comprehension, Numerical Operations, and Coding Speed. These subtests would be particularly well-suited for initial selection decisions. Subtests with particularly high precision for subjects with high abilities include General Science, Arithmetic Reasoning, Word Knowledge, Auto & Shop Information, Mathematical Knowledge, Mechanical Comprehension, and Electronics Information. These subtests would be especially well-suited for placement decisions for subjects who have already been selected.

## PART IV

## ITEM OPERATING CHARACTERISTICS IN THE ASVAB POWER SUBTESTS

The 3-parameter logistic models for ASVAB power tests identify each item individually in terms of its relationship to the underlying ability variable. This part of the report discusses highlights from these item-level analyses.

The item parameters are defined as follows: Item threshold parameters indicate the point along the scale that a subject would have a 50-50 chance of answering correctly. Dispersion parameters are inversely related to the reliabilty with which items measure the ability. Lower asymptotes (guessing parameters) indicate the minimum probabilities of correct response, from even the lowest ability subjects to the hardest items.

Appendix D presents all item parameter estimates and fit indices. Included in these tables are the traditional indices of item analysis: item proportions-correct and item reliabilities. An item proportion-correct estimates the proportion of subjects in the national youth sample who would have answered the item correctly; an item reliability estimates the correlation between the ability measured by that item and the ability measured by the subtest as a whole. Appendix E illustrates the variable effectively measured by each subtest in terms of the item thresholds.

Indices of item fit indicate items for which the item parameters may not provide a complete summary of the interactions of subjects with the item. Fitted and observed response curves were compared for items with fit probabilities less than .02. Discrepancies were minor, attributable in most cases to the large size of the calibration sample. One type of systematic departure did emerge, however; an analysis of guessing behavior on very difficult items follows.

## GUESSING BEHAVIOR

The ASVAB subtests consist of multiple-choice items, which any subject can answer correctly on occasion with a lucky guess. The inference of ability from correct responses must therefore take into account the nature and the level of guessing behavior. This section discusses guessing in the NLS responses to the ASVAB, considering test directions, subjects' comments in a follow-up telephone survey, and evidence from the responses themselves.

Test directions emphasized to subjects that their performances must reflect their abilities as accurately as possible, to be useful to the government in learning about the abilities of young people and to be useful to the subjects themselves for vocational guidance. The directions also told them that if they were not sure of an answer, then "make the best guess you can." Haphazard responses and mechanical guessing were in this way discouraged, although subjects were to infer that a good guess to

a question they were unsure of would better reflect their ability than a blank.

In a follow-up telephone survey of 1003 subjects, Question 8 asked respondents to select the one of three statements that best described how they handled items they were not sure of. Of those responding, 71 percent chose the response saying they almost always made a guess; 24 percent chose the response saying they sometimes guessed and sometimes left the question blank; 5 percent chose the response saying they almost always left the question blank. Subjects' comments, then, suggest that guessing behavior was in accordance with test directions. We now consider the consequences of such behavior as revealed by patterns in the data.

The three-parameter IRC we have chosen for data analysis includes a lower asymptote parameter to account for guessing. The probability that a subject will respond correctly to a particular item is modelled as the sum of the probability that he will respond correctly as a result of his ability, plus the probability that he will respond correctly by guessing even if he cannot answer by virtue of ability; i.e., the lower asymptote of the item. (Because guessing behavior varies from one subject to another, these lower asymptotes were not included in the final estimation of subject abilities. Instead, the "guessing-resistent" biweight estimator described in Appendix B was used. Lower asymptotes do, however, indicate average levels across the population.)

It will be noted that the lower asymptotes of all items were set to .10 during the estimation of item thresholds and dispersions. This value was indicated by analyses of pretest data, collected under conditions identical to those of the present data. This lower asymptote is substantially lower than .25, the probability of answering an item correctly by selecting a distractor at random from the four offered. Two reasons explain this result. First, most subjects were not prone to random guessing. Second, subjects of low ability will have a lower chance of answering an item correctly when they try their best than when they guess, if the distractors have been constructed well.

Certain misfitting items were found to depart from this pattern. Such an item is shown as Figure 1. The dotted line is the theoretical item response curve, and the X's represent groups of about 100 subjects each at various points along the ability scale. When an item fits well, the X's lie near the theoretical curve; the model does a good job predicting the probabilities of correct response at all points along the scale. The pictured item departs from theory at the lower end of the scale, where the proportions of correct response actually begin to increase as ability decreases.

What has happened is this: Subjects of <u>moderately</u> low ability are just able enough to be deceived by well-written item distractors. They respond correctly with probabilities around, say,

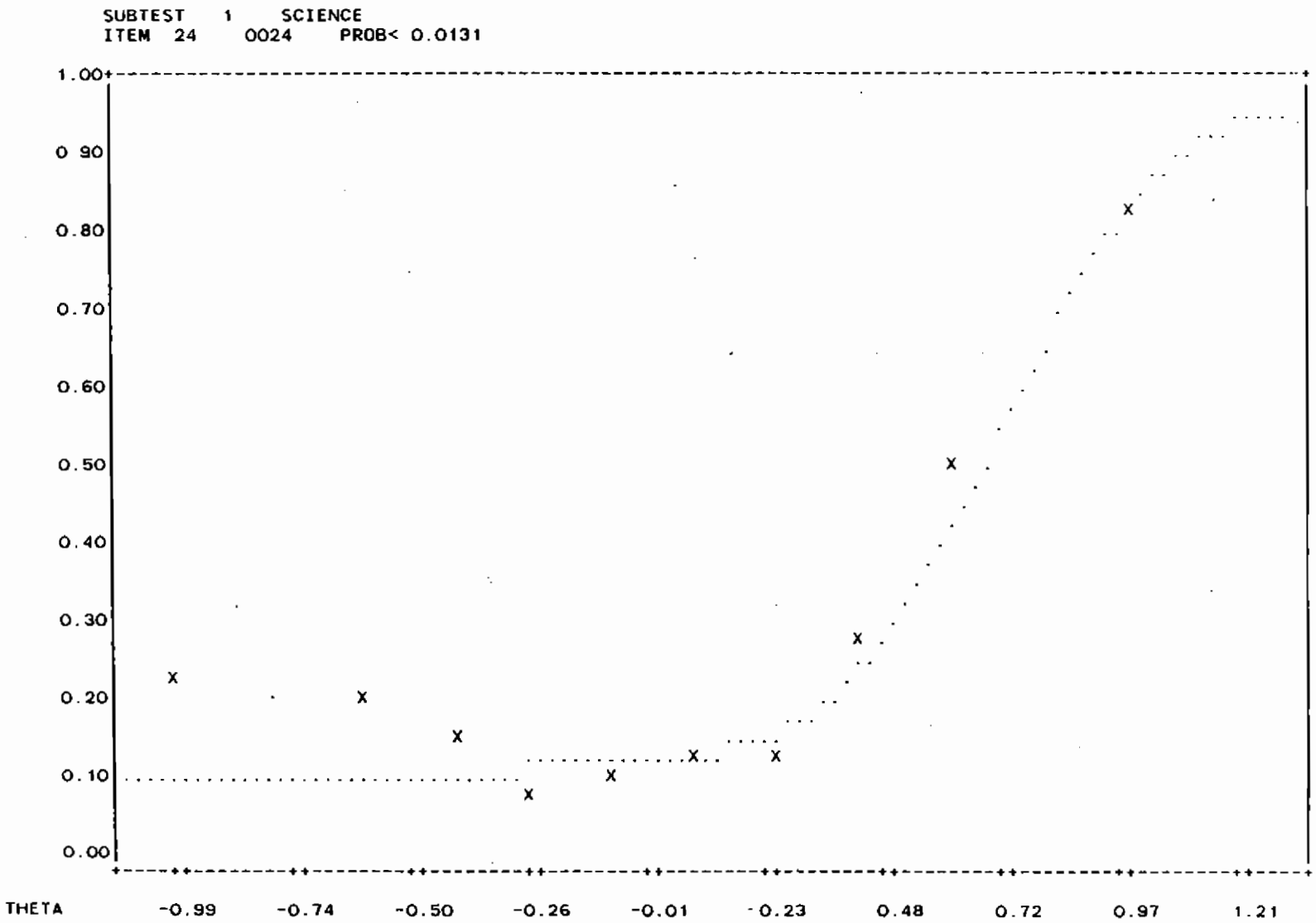



FIGURE 1
EMPIRICAL VS THEORETIC ITEM RESPONSE CURVE

.10. Subjects of very low ability are not sufficiently able to be tempted rationally by any of the distractors; some of these subjects guess, and a fourth of them guess correctly.

This pattern, peculiar to multiple-choice items, is modelled explicitly in the approach of Samejima (1980). This model has not been utilized due to its inability to account for omitted responses and to a lack of a time-tested algorithm for parameter estimation. (The marginal maximum likelihood methods introduced by Bock and Aitken (1981) would be applicable.) Fortunately, the pattern appears only for difficult ASVAB items, and then only for a few. The General Science subtest, for example, has two items that show this pattern; Arithmetic Reasoning has three.

## HIGHLIGHTS OF ITEM-LEVEL RESULTS

Among the items in General Science which best discriminate high from low ability subjects are Item 6 (the use of insulin), Item 18 (a prism refracts light), and Item 24 (definition of absolute zero). The poorest discriminating items are Item 12 (ohm is a unit of resistance) and Item 15 (the lighest metal used for cooking utensils). Among the items which appear to provoke guessing from subjects of low ability are 18, 24, and 25. The content matter and the distractors of such items may be all so unfamiliar to a low ability subject that he is not interacting meaningfully with the item.

The single most discriminating item in Arithmetic Reasoning, which may be taken as the best representative of the variable being measured, is Item 11: "A truck and its load of boxes weighs 9000 pounds. The truck alone weighs 7950 pounds. Each box weighs 15 pounds. How many boxes were in the truck?" Among those which discriminate poorly are 1 and 7. Items which appeared to provoke guessing from low-ability subjects were 18, 24, and 30.

Word Knowledge is measured particularly well by an easy vocabulary word, a moderately difficult word, and a relatively hard word: "ignited," "confine," and "inkling." Words which did not discriminate well included "antidote," "credible," "incantation," and "endorse." Guessing seemed to be provoked by the words "inaccessible," "initial," and "divulge." (It must be recalled that these item characteristics depend inextricibly upon the alternatives offered for the definition. The same vocabulary word may become an easy item if the incorrect alternatives are far from its meaning, or a difficult item if the incorrect alternatives differ only slightly.)

Paragraph Comprehension is measured most accurately by Item 3, which asks the subject to recognize the metaphor on which a short written passage is based. Other items with low dispersion (high reliability) are Items 5, 9, 10, and 13. The content of these items suggests that in general, items addressing the main idea of a passage are more informative than those addressing specifics or details.

The last item in the test has a very high dispersion, suggesting that it is not discriminating high from low ability subjects. Unexpectedly often, subjects who otherwise perform well miss Item 15. This may be traced to the placement of Item 15 in the test booklet: on a nearly blank page by itself, after all the rest of the items, after another item referring to the same passage. Based on the proportion of high ability subjects who missed this item, it seems that forty percent of the subjects were simply not aware that Paragraph Comprehension even had an Item 15.

Auto & Shop Information is measured most accurately by Items 6 (knowing the definition of "power train") and Item 23 (use of a sleeve in auto mechanics). It is not as well measured by Item 8 (purpose of shock absorbers) and Item 9 (properties contributing to the strength of plywood). It may be inferred that by and large, the subtest generally indicates direct experience in the shop; the more esoteric items are better indicators, while items answered by general knowledge or reasoning ability are not as informative.

Mathematics Knowledge is particularly well measured by Item 13 [(0.2)(2.0 + 0.2 + 0.22)=?], Item 18 [if a=2 and b=5, then $4ab^2$=?] and Item 22 (a simple quadratic expansion). It is not as well measured by Item 5 (a protractor is used to measure an angle) or Item 15 (definitions of complementary and supplementary angles). It would appear that the simplest applications of the algebraic rules presented in higher-level high school math classes are most informative here.

Mechanical Comprehension is best measured by Item 2 (rollers reduce friction), Item 8 (center of gravity), and Item 10 (relative speeds of different sized pulleys). Less informative are Item 6 (a cam and rod in a rather complicated diagram), Item 11 (which asks which of four wheels is best suited for rough terain), and Item 19 (rpm's of wheels on a shaft). Guessing seems prevalent on Items 18, 19, and 21, all relatively hard and all with complex illustrations.

Electronics Information is measured almost equally well by all items except three: Item 11 (differences between fuses and circuit breakers), Item 14 (what does the color of a light swith mean?), and Item 17 (the purpose of an output transformer). For one reason or another, even subjects who generally performed well on other items had a particularly hard time with Item 17.

## CONCLUSIONS

The overall level of guessing behavior was about 10-percent; that is, even the lowest ability subjects answered the hardest items correctly about one time of every ten, probably with a lucky guess. This level is substantially lower than the 25-percent level of correct responses that these subjects would obtain by always guessing. It may be inferred that subjects were, on the whole, interacting meaningfully with the content of the items. With less noise from guessing than typical in multiple-choice tests, the ASVAB results will provide more reliable measures of

ability than might be ordinarily expected at the lower ends of the distribution.

A detailed analysis of item parameter estimates and item fit statistics revealed only one badly flawed item, the final item on the Paragraph Comprehension test which, because of its location in the test booklet, was carelessly omitted by nearly half of the subjects.

## RECOMMENDATIONS

Final versions of test forms should avoid "widowed" items, i.e., items at the end of a test appearing on a page by themselves.

Writers of Paragraph Comprehension items should concentrate on questions addressing main ideas and generalizations rather than details.

Technical subtests such as Auto & Shop Information, Electronics Information, and Mathematics Knowledge should stress applications of basic principles in the field.

# PART V

# ANALYSIS OF SUBJECT FIT

The IRC models represent an ideal of the interaction between subjects and test items. Indices of subject fit indicate the degree to which the response pattern of a particular subject is in accord with the ideal. Atypical patterns of response signal carelessness, random response, inattentiveness, or malingering.

This part of the report is an analysis of indices of subject fit. When possible, observed distributions of subject fit indices are compared with a theoretical standard. In all cases, distributions are compared across sex and racial/ethnic subgroups.

## POWER SUBTESTS

### METHODOLOGY

Two indices of subject fit are examined for the power subtests of the ASVAB. The first is a count of "unexpected" responses; that is, correct responses to hard items from low-ability subjects, and incorrect responses to easy items from high-ability subjects. For this index, responses are compared to an ideal that assumes the absence of guessing behavior. The proportions of

responses that are attributable to lucky guesses can be estimated with this index, and compared across subtests and demographic subgroups. The second index is a modified Wright-Mead Z-statistic (Mead, 1976; Wright, 1977). For this index, the ideal assumes a constant level of guessing probabilities of .10, then evaluates the likelihood of a subject's response pattern in terms of an approximate standard normal deviate. This index allows a comparison across subgroups of the relative quality of the data, given the population average of guessing behavior. The two indices are defined as follows:

Counts of unexpected responses model the probability that Subject i will answer Item j correctly as

$$P_{ij} = \exp[(\theta_i - b_j)/s_j]/\{1 + \exp[\theta_i - b_j)/s_j]\}.$$

The observed response $X_{ij}$ is counted as unexpected if its modelled probability is less than .01. This occurs when either a low-ability subject answers a difficult item correctly (probably by guessing) or when a high-ability subject answers an easy item incorrectly (probably through carelessness).

Contrary to appearances, the theoretical proportion of unexpected responses is not necessarily .01, because an unexpected response can occur only when a low-ability subject is presented a difficult item or a high-ability subject is presented an easy item. The theoretical expectation would be a complicated function of the distributions of subject ability, item thresholds, and item dispersions. The .01 level does represent an upper limit of the expectation, however.

Our modified Wright-Mead Z index models the probability that Subject i will answer Item j correctly as

$$P_{ij} = .10 + .90 \exp[(\theta_i - b_j)/s_j]/\{1 + \exp[\theta_i - b_j)/s_j]\}.$$

The observed reponse $X_{ij}$ is assigned the value 1 if it is correct and 0 if it is not. The observed response is compared with its modelled probability through a psuedo-Chi-square statistic:

$$V_{ij} = (X_{ij} - P_{ij})^2 / [P_{ij}(1 - P_{ij})].$$

Such values are summed for Subject i over all the items in a subtest to provide an approximate Chi-square statistic, with degrees of freedom equal to the number of items minus one. For convenience of comparisons across subtests, these values have been converted to standard normal deviates via Fisher's transformation (see Bishop, Fienberg & Holland, 1975, page 527). The expected average Wright-Mead Z in any group of subjects, then, is zero and the expected standard deviation is one.

## RESULTS

Table 4 presents the percentages of unexpected ($p<.01$) responses in each of the ASVAB power subtests; Table 5 presents average Wright-Mead Z-statistics. In both tables, the classifications of low, medium, and high ability are based on a comparison to the overall distribution of ability estimates in the youth population: low abilities are -.75 and below, medium abilities range from -.75 to +.75, and high abilities are those above +.75.

TABLE 4
PERCENTAGES OF UNEXPECTED RESPONSES IN POWER SUBTESTS*

| | SCIE | ARTH | WORD | PARA | AUTO | MATH | MECH | ELEC |
|---|---|---|---|---|---|---|---|---|
| EXPECTATION** | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| OVERALL | .28 | .09 | .23 | .05 | .23 | .45 | .01 | .02 |
| MALES | .31 | .10 | .30 | .07 | .25 | .46 | .02 | .03 |
| FEMALES | .25 | .09 | .16 | .03 | .22 | .44 | .01 | .01 |
| LOW ABILITY | .90 | .31 | .36 | .00 | 1.43 | 2.04 | .00 | .05 |
| MEDIUM ABILITY | .10 | .00 | .05 | .08 | .08 | .05 | .00 | .00 |
| HIGH ABILITY | .12 | .17 | .48 | .00 | .19 | .19 | .10 | .05 |
| LOW WHITE | .66 | .21 | .45 | .00 | 1.20 | 1.59 | .00 | .00 |
| LOW BLACK | .86 | .35 | .28 | .00 | 1.65 | 2.34 | .00 | .11 |
| LOW HISPANIC | 1.19 | .34 | .40 | .00 | 1.14 | 2.29 | .00 | .00 |
| MEDIUM WHITE | .05 | .01 | .08 | .06 | .05 | .03 | .00 | .00 |
| MEDIUM BLACK | .20 | .00 | .02 | .17 | .11 | .06 | .00 | .00 |
| MEDIUM HISPANIC | .16 | .00 | .03 | .07 | .12 | .13 | .00 | .00 |
| HIGH WHITE | .09 | .18 | .40 | .00 | .02 | .13 | .07 | .04 |
| HIGH BLACK | .00 | .00 | .63 | .00 | .00 | .57 | .00 | .00 |
| HIGH HISPANIC | .05 | .00 | 1.27 | .00 | .00 | .44 | .23 | .22 |
| WHITE | .16 | .07 | .22 | .04 | .09 | .25 | .02 | .01 |
| BLACK | .47 | .13 | .19 | .09 | .52 | .77 | .00 | .03 |
| HISPANIC | .60 | .11 | .31 | .04 | .30 | .73 | .02 | .03 |

* BASED ON A 10% RANDOM SAMPLE OF NLS PANEL SUBJECTS
** UPPER LIMIT TO EXPECTATION IN ANY GROUP OF SUBJECTS

TABLE 5
AVERAGE WRIGHT-MEAD Z-STATISTIC IN POWER SUBTESTS*

| | SCIE | ARTH | WORD | PARA | AUTO | MATH | MECH | ELEC |
|---|---|---|---|---|---|---|---|---|
| EXPECTATION** | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| OVERALL | .47 | .13 | .40 | .02 | .02 | .28 | .13 | .23 |
| MALES | .63 | .43 | .36 | .20 | -.02 | .09 | .12 | .16 |
| FEMALES | .32 | -.15 | .44 | -.15 | .06 | .45 | .13 | .30 |
| LOW ABILITY | .41 | .00 | .21 | .19 | -.10 | .01 | -.31 | -.03 |
| MEDIUM ABILITY | .50 | .21 | .59 | .43 | .17 | .36 | .24 | .31 |
| HIGH ABILITY | .44 | .01 | .19 | -2.58 | -.55 | .28 | .23 | .18 |
| LOW WHITE | .22 | .00 | .30 | .24 | .18 | -.20 | -.35 | -.06 |
| LOW BLACK | .43 | .05 | .06 | .11 | -.02 | .21 | -.24 | .06 |
| LOW HISPANIC | .55 | -.10 | .36 | .26 | -.54 | .01 | -.38 | -.15 |
| MEDIUM WHITE | .26 | .25 | .50 | .30 | .16 | .19 | .22 | .32 |
| MEDIUM BLACK | .82 | -.03 | .56 | .78 | .19 | .59 | .24 | .25 |
| MEDIUM HISPANIC | 1.14 | .45 | 1.01 | .55 | .16 | .61 | .30 | .39 |
| HIGH WHITE | .06 | .04 | -.20 | -2.64 | -.50 | -.35 | .20 | .12 |
| HIGH BLACK | .77 | .54 | 1.01 | -3.11 | .22 | 3.43 | .67 | -.08 |
| HIGH HISPANIC | 4.41 | -.63 | 3.73 | -1.48 | -1.15 | 4.05 | .31 | .81 |
| WHITE | .21 | .17 | .27 | -.21 | .00 | .01 | .16 | .24 |
| BLACK | .65 | .01 | .33 | .41 | .13 | .61 | .06 | .17 |
| HISPANIC | 1.19 | .21 | 1.01 | .32 | -.09 | .78 | .09 | .29 |

* BASED ON A 10% RANDOM SAMPLE OF NLS PANEL SUBJECTS
** VALUES BELOW ZERO INDICATE BETTER FIT THAN EXPECTED;
VALUES ABOVE ZERO INDICATE WORSE FIT THAN EXPECTED.

The divisions break the population into roughly the lowest quartile, the middle half, and the highest quartile.

The ASVAB power subtests show low proportions of unexpected responses. There are practically none in Mechanical Comprehension and Electronics Information, and few in Paragraph Comprehension and Arithmetic Reasoning. Somewhat higher (though still low in absolute terms) levels are found in General Science, Word Knowledge, and Auto & Shop Information. Of the eight ASVAB power subtests, Mathematical Knowledge shows the highest proportion of unexpected responses, about one-half of one percent.

Females show slightly but consistently lower percentages of unexpected responses than males. The magnitude of this effect is inconsequential.

From the table of percentages of unexpected responses, it can be seen that the bulk of unexpected responses are correct answers to hard items by otherwise low-scoring subjects. Most of these responses are guesses, although some do represent actual ability. This may be particularly true in the Auto & Shop subtest, where a low-scoring subject may have experience with a particular concept or tool.

An exception to the pattern is Word Knowledge, where more than half of the unexpected responses are incorrect responses to easy items from high-scoring subjects--often Black and Hispanic high-scoring subjects. (An explanation for these anomolies is suggested in Part VI, where a heterogeneity of item content with

respect to racial/ethnic groups is discovered.) In General Science and Mathematics Knowledge also, high-scoring subjects from the Black and Hispanic subgroups show higher proportions of incorrect answers to easier items than high-scoring subjects from the white group.

The classification of subjects into low, medium, and high ability strata can be considered a crude control for ability levels on the quality of data. With the exception noted above of high-scoring Black and Hispanic carelessness(?) in General Science, Word Knowledge, and Mathematics Knowledge, the differences among racial/ethnic groups within ability stata are minimal. The proportions of unexpected responses are lowest among the subjects of medium ability in all racial/ethnic groups, next lowest among the subjects of high ability, then highest among subjects of low ability. Because the Black and Hispanic groups have more subjects of low ability, however, the combined data for these groups contains proportionally more unexpected responses than the data for the White group.

## SPEEDED SUBTESTS

### METHODOLOGY

Two indices of subject fit are also examined for the speeded tests of the ASVAB battery, Numerical Operations and Coding Speed. The first is a count of unexpected runs of incorrect responses before the subject's last non-omitted response. Included in such

a count would be an instance of marking the wrong column in the answer sheet. The second index is a test of the equality of a subject's proportions of correct response in the first, middle, and final thirds of the test before his last non-omitted response. Instances of initial difficulties and of random guessing as the time limit neared would be flagged by this index. The indices are defined as follows:

Counts of unexpected runs of incorrect responses are defined uniquely for each subject. Let $P_i$ be the proportion of correct responses made by Subject i to items before and including his last non-omitted response. Let $K_i$ be the smallest integer such that

$$(1-P_i)^{K_i} < .005.$$

If the incorrect responses of Subject i were distributed randomly throughout his attempts, strings of $K_i$ incorrect responses would appear about 5 times in 1000 strings of $K_i$ responses. An occurance of such a string suggests that the subject may have had a problem with the test booklet or the answer sheet.

Because the probability of observing a run of $K_i$ incorrect responses varies from one subject to the next with their percent-correct scores and their numbers-of-items attempted, no theoretical standard of comparison is available for this index. Lower counts are better counts nonetheless, so comparisons can be made across sex and racial/ethnic groups.

The serial-thirds Z-statistic expresses as a standard normal deviate the likelihood that a subject's probabability of respond-

ing correctly was the same over the first, middle, and final thirds of the items he responded to. For its computation define the following quantities:

$N_{i+}$ = number of items up to and including the last non-omitted response of Subject i,

$R_{i+}$ = number of correct responses by Subject i,

$N_{ij}$ = number of items attempted by Subject i in his j-th third of attempts, and

$R_{ij}$ = number of correct responses by Subject i in his j-th third of attempts.

A Chi-square value testing the equality of the probability of a correct response in the serial thirds of Subject i's responses is given by

$$V_i = \sum_{j=1}^{3} [R_{ij} - N_{ij}(R_{i+}/N_{i+})]^2 / (R_{i+}/N_{i+}),$$

which has two degrees of freedom. These quantities have been converted to standard normal deviates by Fisher's transformation. A high value suggests that a subject may have had difficulties with instructions or response formats at the beginning of the test, or may have become careless or disinterested as the time limit neared.

## RESULTS

Table 6 presents the average counts of unexpected runs of incorrect responses observed in Numerical Operations and Coding Speed; Table 7 presents average serial-thirds Z-statistics.

TABLE 6
AVERAGE NUMBERS OF UNEXPECTED RUNS OF INCORRECT RESPONSES*

| | NUMERICAL OPERATIONS | CODING SPEED |
|---|---|---|
| EXPECTATION | N.A. | N.A. |
| OVERALL | .31 | .59 |
| MALES | .31 | .57 |
| FEMALES | .31 | .61 |
| LOW ABILITY | .46 | 1.03 |
| MEDIUM ABILITY | .29 | .47 |
| HIGH ABILITY | .16 | .30 |
| LOW WHITE | .43 | 1.07 |
| LOW BLACK | .45 | .97 |
| LOW HISPANIC | .54 | 1.07 |
| MEDIUM WHITE | .23 | .47 |
| MEDIUM BLACK | .42 | .42 |
| MEDIUM HISPANIC | .35 | .50 |
| HIGH WHITE | .14 | .30 |
| HIGH BLACK | .22 | .26 |
| HIGH HISPANIC | .26 | .36 |
| WHITE | .24 | .53 |
| BLACK | .42 | .66 |
| HISPANIC | .42 | .66 |

* BASED ON A 10% RANDOM SAMPLE OF NLS PANEL SUBJECTS

TABLE 7
AVERAGE VALUES OF SERIAL-THIRDS Z-STATISTICS*

| | NUMERICAL OPERATIONS | CODING SPEED |
|---|---|---|
| EXPECTATION** | 0.00 | 0.00 |
| OVERALL | -1.28 | -1.11 |
| MALES | -1.25 | -1.11 |
| FEMALES | -1.31 | -1.11 |
| LOW ABILITY | -0.99 | -0.57 |
| MEDIUM ABILITY | -1.37 | -1.26 |
| HIGH ABILITY | -1.51 | -1.44 |
| LOW WHITE | -1.05 | -0.64 |
| LOW BLACK | -0.95 | -0.53 |
| LOW HISPANIC | -0.96 | -0.51 |
| MIDDLE WHITE | -1.41 | -1.28 |
| MIDDLE BLACK | -1.28 | -1.30 |
| MIDDLE HISPANIC | -1.35 | -1.15 |
| HIGH WHITE | -1.52 | -1.45 |
| HIGH BLACK | -1.49 | -1.45 |
| HIGH HISPANIC | -1.47 | -1.34 |
| WHITE | -1.37 | -1.21 |
| BLACK | -1.12 | -0.96 |
| HISPANIC | -1.19 | -0.97 |

* BASED ON A 10% RANDOM SAMPLE OF NLS PANEL SUBJECTS
** VALUES BELOW ZERO INDICATE BETTER FIT THAN EXPECTED; VALUES ABOVE ZERO INDICATE WORSE FIT THAN EXPECTED.

The counts of unexpected runs are higher for Coding Speed, which is to be expected because it is a longer test. In neither subtest do males differ from females. In both subtests, there are more runs of unexpected incorrect responses from low-scoring subjects. This pattern is consistent across racial/ethnic groups in Coding Speed, but favors Whites in Numerical Operations.

The serial-thirds averages are excellent for both subtests and all subject groups. Any initial difficulties or final guessing sprees that do exist would appear to be confined to the lowest-scoring groups, evenly across sex and racial/ethnic divisions.

## CONCLUSIONS

The incidence of "unexpected" patterns of response in the ASVAB responses of the NLS panel was low. For the most part, the subjects of the NLS panel provide evidence of meaningful interaction with the test items, rather than haphazard guessing, carelessness, or malingering. Given the limitations of their mulitiple-choice formats, the time available for testing, and the lengths of the subtests, the ASVAB subtests have evoked responses of a quality equal to or surpassing that of typical measures of educational achievement and aptitude.

Evidence of guessing behavior, namely correct responses to hard items from low-scoring subjects, was almost nil in Electronics Information, Mechanical Comprehension, and Paragraph

Comprehension. It was higher, though still negligible, in Arithmetic Reasoning and Word Knowledge. Of the ASVAB power subtests, it was highest in General Science, Auto & Shop Information, and Mathematics Knowledge, where some 1- to 2-percent of the responses from subjects in the lowest quartile may have been correct guesses. Overall, it may be estimated that even the lowest-ability subjects answered the hardest items correctly about one time out of ten with a correct guess.

This level of chance success is substantially less than the 25-percent level that would correspond to random guessing among the four response alternatives, a level often observed and occasionally encouraged in commercially-published tests of aptitude (see Samejima, 1980). Norms derived from the responses of the NLS panel, then, will be comparatively cleaner in this respect. To the extent that it does exist, guessing clouds precision at the lower ends of the norms. (This degradation of information is discussed in Part III, Subtest Information and Reliability.) While the degree of guessing is fairly consistent across racial/ethnic subgroups, those subgroups with lower average levels of ability will suffer more effect.

The opposite kind of unexpected response, namely an incorrect answer to an easy item by a high-scoring subject, occurred more rarely. Particular problems surfaced in General Science, Word Knowledge, and Mathematics Knowledge, where high-scoring subjects from the Black and/or Hispanic groups were more likely to miss

easy items than high-scoring subjects from the White group. The worst case is that of Word Knowledge, where about 1-percent of the responses of high-scoring Hispanics may be incorrect responses to easy items. Part VI below discusses the possiblility that these anomolous responses may be related systematically to item content.

# PART VI

# INVESTIGATION OF CULTURAL ITEM BIAS

Mental test scores are ambiguous and poorly-suited for the comparison of subjects if the test items have different meanings to subjects from different cultural backgrounds. Group-by-item interactions of this type may cause a uniform system of ability estimation to systematically over-estimate the abilities of one group and under-estimate the abilities of another.

This type of cultural item bias may be distinguished from another type of test bias, in which a particular use of a test score leads to the over- or under-prediction of some criterion variable. The relationship between these two types of bias, and the implications for the use of ASVAB subtest scores, are discussed in the final section of Part VI. Our primary concern, however, is whether a given score from an ASVAB power subtest has essentially the same meaning across sex and racial/ethnic groups.

Altogether, seven pairwise comparisons of demographic subgroups have been made for each ASVAB power test, for both thresholds and dispersions. These comparisons are the following:

(1) Males versus females

(2) Non-disadvantaged whites versus disadvantaged whites

(3) Non-disadvantaged whites versus Blacks

(4) Non-disadvantaged whites versus Hispanics

(5) Disadvantaged whites versus Blacks

(6) Disadvantaged whites versus Hispanics

(7) Blacks versus Hispanics

## METHODOLOGY

The theory of item response curve models provides a method for defining and detecting cultural item bias (see Lord, 1977; Wright, 1977; Mislevy, 1981). The reasoning is as follows:

If the items of a test are in fact measuring the same underlying variable for all subjects, then item parameters estimated from the responses of any subgroup of subjects will be statistically equivalent. This should be true even when the subgroups are chosen according to, say, a suspect cultural background variable.

Under the null hypothesis of equivalent item parameters, subjects from any subgroup are being measured on the same variable and comparisons across groups are justified. Under the alternative hypothesis of non-equivalence, the test items are defining qualitatively different variables in different groups, and comparisons of subjects across group boundaries are less trustworthy.

In this latter case, item parameter estimates from the various groups are examined to determine which items appear to operate differently from one group to another. Certain items may be consistently easier for members of some groups than for others; certain items may distinguish higher- from lower-ability subjects better in some groups than in others. Such items offer clues for subsequent test revision, perhaps by their omission in future versions of the subtest or their expansion into new subtests.

Our investigation of cultural bias in the power subtests of the ASVAB focuses on sex groups and racial/ethnic groups. The NLS sampling design identifies two sex groups (males and females) and four racial/ethnic groups (non-disadvantaged whites, disadvantaged whites, non-Hispanic Blacks, and Hispanics). Each power test was calibrated separately within each cell of the 2-by-4 design, based on the responses of a random sample of 500 subjects from that cell.

Aside from a linear transformation, the sets of item parameter estimates from each subpopulation should not differ by more than sampling error. A general linear model for testing such hypotheses was introduced by Mislevy (1980), with the matrix of standardized item parameter estimates expressed as the product of three matrices plus sampling error in the following manner:

$$\underset{nxl}{D} = \underset{nxs}{K}\ \underset{sxr}{\Gamma}\ \underset{rxl}{T} + \underset{nxl}{E},$$

where

D is a matrix of item parameter estimates for the l items

of a test, as calibrated in n subject groups,

K is a design matrix embodying s contrasts among groups,

T is a design matrix embodying r contrasts among items,

$\Gamma$ is a matrix of parameters expressing interactions between salient features of groups and salient features of items, and

E is a matrix of errors of estimation of item parameters.

This model is particularly powerful for testing for the presence of suspected interactions between particular groups of subjects with particular groups of items. When item thresholds are the object of study, the parameter matrix $\Gamma$ may be interpreted as the magnitudes by which subtest scores over- or under-estimate specific abilities of subjects in particular groups. For the subtests of the ASVAB, however, an a priori structuring of item content in a manner likely to produce item-by-group interactions was found only in the General Science subtest. An analysis of Sex-by-Content interactions is discussed in a following section.

When no clear structure is proposed a priori, the Mislevy model simplifies to the tests for item bias suggested by Lord (1977) and others. In particular, a Chi-square test for the equality of item thresholds or dispersions the equality of item parameters across two subject groups is obtained as follows:

Step 1

Rescale the item parameter estimates within each cell to the

same metric, requiring the arithmetic mean of the threshold estimates to be zero and the geometric mean of the dispersions to be one.

Step 2

Obtain parameter estimates for each margin by averaging corresponding cell estimates. Because the estimates within each cell are independent, the standard error of estimation of, say, the Males estimate of a particular item threshold is the square root of the sum of the squared standard errors of the estimates of that threshold in each of the racial/ethnic groups, divided by four. The resulting item parameters from this step are presented as Appendix F.

Step 3

To compare, say, the item threshold estimate of a particular item for males and females, a Chi-square index is obtained as the squared difference of the two estimates, divided by the sum of the squares of the corresponding standard errors; e.g.,

$$X^2_{b_j} = \frac{(b_{j1} - b_{j2})^2}{SE^2(b_{j1}) + SE^2(b_{j2})}$$

Step 4

Under the null hypothesis of equivalent item parameters across two demographic groups, the sum of the item Chi-squares in a particular subtest will itself be a Chi-square, with degrees of freedom one less than the number of test items.

## RESULTS

The Chi-square values for the equality of thresholds across demographic groups are shown as Table 8-A, and the values for dispersions, as Table 8-B. When these Chi-squares are considered in the light of a design effect of about two, few of these values cast doubt on the hypothesis of equality. It may be concluded that except for the cases discussed in the following paragraphs, (1) item parameters are essentially equivalent across sex and racial/ethnic groups, and (2) comparisons of subjects across subgroups in terms of either number-right or latent trait scores are justifiable.

The item thresholds in General Science are equivalent across racial/ethnic groups, but not across sex groups. Figure 2 plots thresholds estimated from males against those estimated from females. Items dealing with physical sciences--chemistry and physics--tend to be relatively easier for males, while items dealing with health sciences--medicine and nutrition--tend to be relatively easier for females. Biology items favor neither sex.

The magnitude of the effect may best be illustrated by an example. Consider a male and a female who perform at the same level in the General Science test. In a test of biology items only, both subjects would still perform at about the same level; if only one of them answered a particular item correctly, the odds

TABLE 8
CHI-SQUARE STATISTICS FOR EQUALITY OF WITHIN-GROUP ITEM PARAMETERS

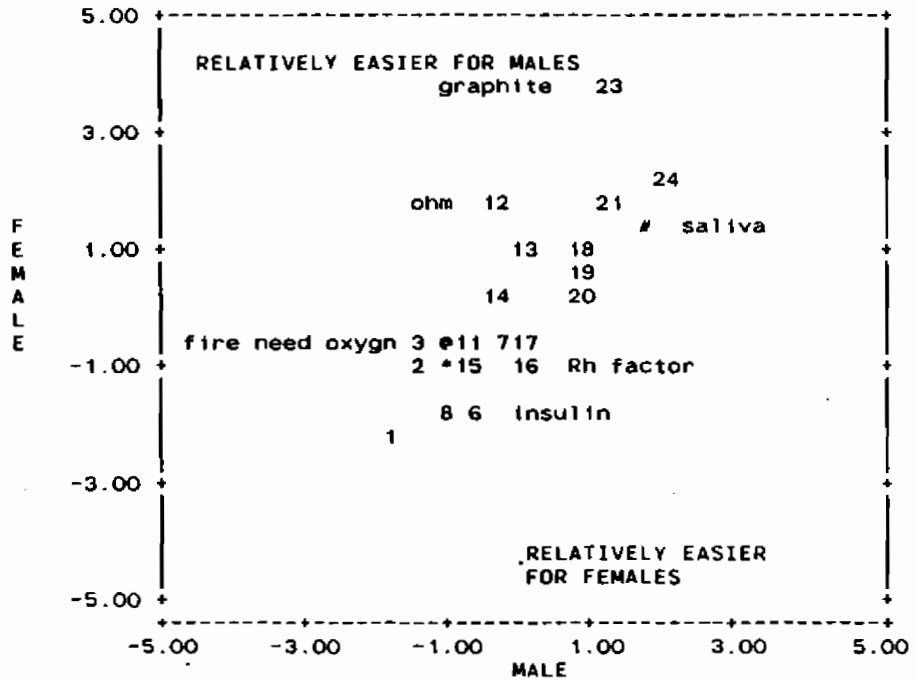
A. THRESHOLDS

| TEST | DF | MALE V FEMALE | WHITE V POOR | WHITE V BLACK | WHITE V HISPANIC | POOR V BLACK | POOR V HISPANIC | BLACK V HISPANIC |
|---|---|---|---|---|---|---|---|---|
| SCIE | 24 | 406.773 | 15.089 | 56.463 | 56.215 | 50.830 | 52.352 | 32.271 |
| ARTH | 29 | 83.926 | 25.472 | 40.679 | 47.878 | 25.945 | 29.585 | 18.536 |
| WORD | 34 | 360.663 | 50.848 | 160.789 | 299.997 | 101.654 | 205.937 | 118.465 |
| PARA | 14 | 20.624 | 5.829 | 43.651 | 26.504 | 11.525 | 4.416 | 9.607 |
| AUTO | 24 | 355.838 | 18.742 | 58.311 | 42.777 | 64.109 | 43.184 | 28.910 |
| MATH | 24 | 26.385 | 34.145 | 118.124 | 15.237 | 85.910 | 8.071 | 5.032 |
| MECH | 24 | 129.099 | 21.758 | 52.087 | 30.458 | 24.895 | 21.648 | 21.390 |
| ELEC | 19 | 126.817 | 11.637 | 33.155 | 35.618 | 30.103 | 43.461 | 14.380 |

B. DISPERSIONS

| TEST | DF | MALE V FEMALE | WHITE V POOR | WHITE V BLACK | WHITE V HISPANIC | POOR V BLACK | POOR V HISPANIC | BLACK V HISPANIC |
|---|---|---|---|---|---|---|---|---|
| SCIE | 24 | 69.730 | 35.556 | 69.760 | 76.760 | 40.331 | 54.184 | 22.773 |
| ARTH | 29 | 79.975 | 58.375 | 97.554 | 67.970 | 108.145 | 34.387 | 84.421 |
| WORD | 34 | 102.197 | 84.078 | 119.215 | 91.794 | 137.211 | 98.356 | 77.207 |
| PARA | 14 | 60.709 | 123.608 | 34.333 | 105.228 | 59.310 | 15.160 | 39.911 |
| AUTO | 24 | 185.700 | 43.162 | 25.224 | 89.485 | 36.654 | 45.546 | 53.497 |
| MATH | 24 | 50.034 | 45.773 | 102.440 | 144.993 | 112.233 | 108.061 | 246.146 |
| MECH | 24 | 70.687 | 44.315 | 53.198 | 70.477 | 24.839 | 28.843 | 32.415 |
| ELEC | 19 | 100.984 | 15.652 | 28.916 | 53.970 | 14.478 | 32.051 | 30.375 |

NOTE: SAMPLING DESIGN EFFECT OF ABOUT TWO SHOULD BE TAKEN INTO ACCOUNT.




ITEMS WITH SHARED LOCATIONS

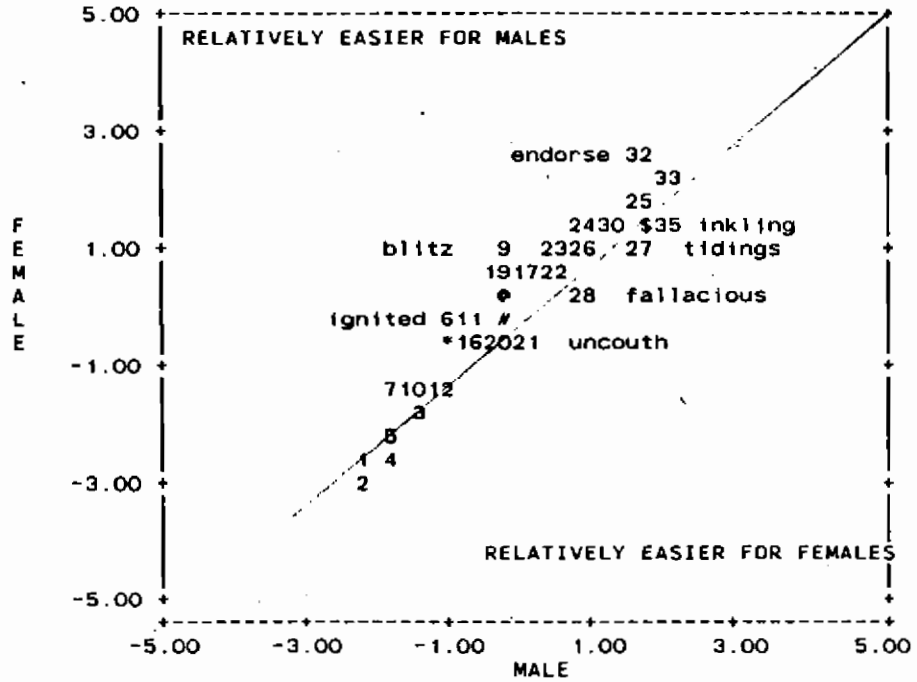
* : 5, 9,
@ : 4, 10,
\# : 22, 25,

FIGURE 2
SCIENCE TEST THRESHOLDS: MALE VS FEMALE

are 47:53 that was the male. In a test of health science items only, the female would probably perform better; if only one of them answered an item correctly, the odds that it was the male are only 22:78. Finally, a test of physical sciences items would favor the male, with item-level odds of about 56:44.

The thresholds of the items in Word Knowledge differ significantly between males and females, and between Hispanics and the other racial/ethnic groups. Figures 3 and 4 are plots of male versus female thresholds, and Hispanic versus White thresholds. The differences are not as readily explained in terms of categories of items as were the differences in General Science. Post hoc explanations for male-female differences are available for certain words; "blitz," a term used in football, is relatively easier for males. As for Hispanic-White differences, it has been hypothesized that the words relatively easier for Whites tend to be literary while the words relatively easier for Hispanics tend to be "media" words. This hypothesis suggests the distinction of literary vocabulary from media vocabulary; the direction of difference suggests that if a White and a Hispanic scored similarly on the Word Knowledge test, the White has more likely done better with literary words and the Hispanic, with media words. The magnitude of the difference is about 2:1 at the item level for the affected words.

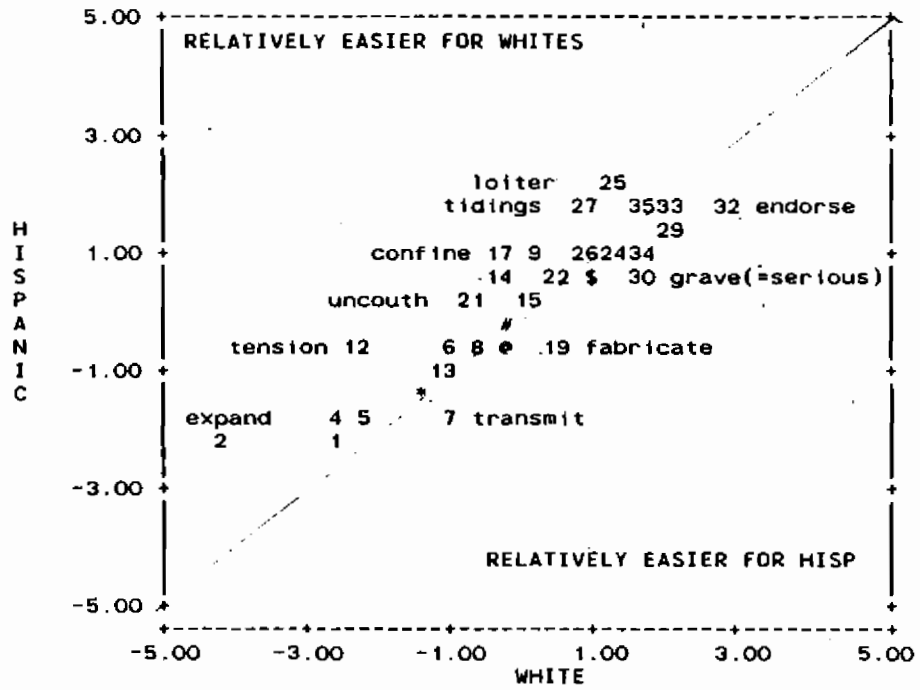
Auto & Shop Information is the final instance of threshold differences across groups. Male thresholds are plotted against




ITEMS WITH SHARED LOCATIONS

* : 8, 13,
@ : 14, 15,
\# : 18, 31,
$ : 29, 34,

FIGURE 3
WORD KNOWLEDGE THRESHOLDS: MALE VS FEMALE




ITEMS WITH SHARED LOCATIONS

* : 3, 10,
@ : 11, 16, 31,
# : 18, 20,
$ : 23, 28,

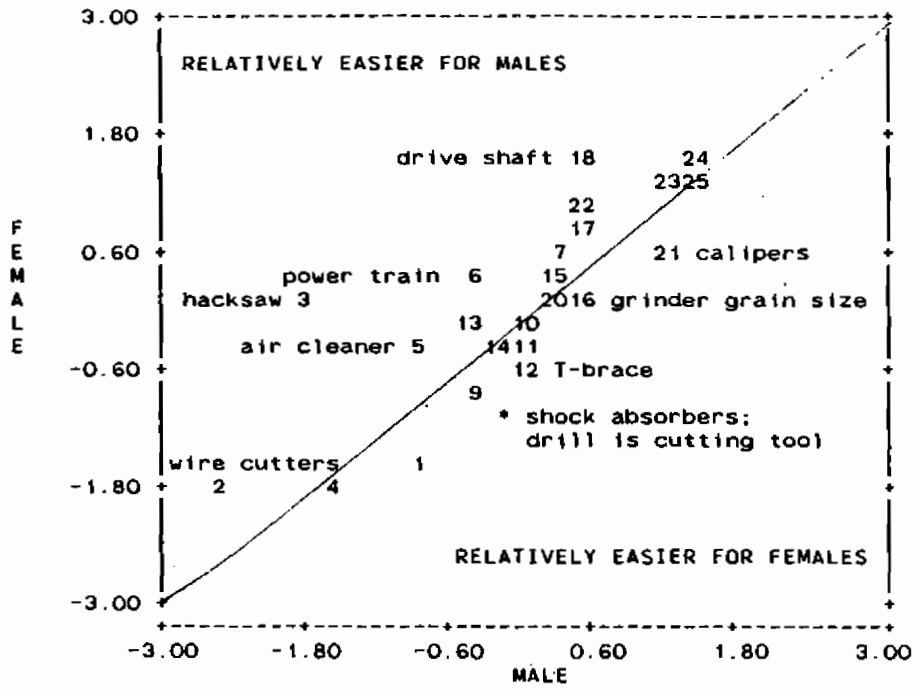
FIGURE 4
WORD KNOWLEDGE THRESHOLDS: WHITE VS HISPANIC

female thresholds in Figure 5. It appears that some items requiring general knowledge or reasoning ability rather than shop-specific knowledge are relatively easier for females, on the average.

Differences among demographic groups as to item dispersion are not great. Figure 6 illustrates one of greatest differences, that of Blacks and Whites in Word Knowledge. Certain items appear relatively more informative about Whites (e.g., "credible" and "transmit") while other items appear relatively more informative about Blacks (e.g., "expand" and "antidote"). These differences do not merit detailed discussion in our opinion, for a number of reasons, including (1) dispersions vary considerably from one sample of subjects to the next, and (2) even wide variations in discrimination parameters have little effect on the estimation of subject abilities, assuming that knowing a word is always more suggestive of ability than not knowing it.

## CONCLUSIONS

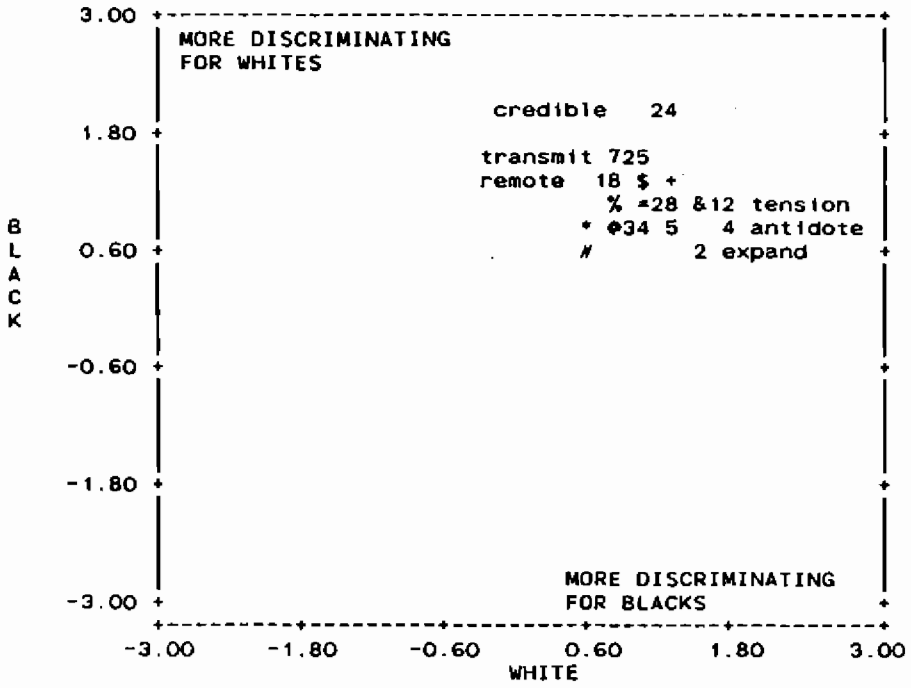
Test item cultural bias, with regard to sex and racial/ethnic groups, was not apparent in the ASVAB power subtests, with two minor exceptions: (1) The General Science subtest is a composite of three distinct variables, Physical Science, Health Science, and Biology. The use of a single General Science score tends to over-estimate the Health Science ability of males but underestimate their Physical Science abilities, while the reverse is




ITEMS WITH SHARED LOCATIONS

* : 8, 19.

FIGURE 5
AUTO & SHOP THRESHOLDS: MALE VS FEMALE




ITEMS WITH SHARED LOCATIONS

* : 6, 8, 10, 17,
@ : 1, 11, 13, 20, 26,
\# : 3, 15, 31, 35,
$ : 14, 19, 21, 30,
% : 22, 23,
& : 9, 27,
\+ : 29, 32,
= : 16, 33,

FIGURE 6
WORD KNOWLEDGE DISPERSIONS: WHITE VS BLACK

true for females. (2) Word Knowledge shows certain items that are slightly but systematically easier for Whites, and others that are slightly but systematically easier for Blacks and Hispanics. The words that were relatively easier for Whites tended to be more "literary." As indicated by the analyses of subject-fit indices, the disturbances caused by these group-by-item interactions are not severe.

## RECOMMENDATIONS

Whether the group-by-item interactions noted above will cause errors in selection or classification depends on the specific use of the scores. The determining factors will be the relationships between the criterion being predicted and the several abilities being collapsed into the subtest score in question. It is difficult to imagine a criterion for which the noted racial/ethnic group differences in Word Knowledge would matter, but certain uses of the General Science scores would merit a validation study:

General Science scores would predict equally well the performance of males and females in a training course that required some science aptitude but no particular experience. However, they might lead to biased predictions of performance in a course that depended more heavily on one component of General Science. Performance in training for medics, as an example, might be over-predicted for males and under-predicted for females.

## PART VII

## AREA SUPERVISOR EFFECTS

In order for norms derived from these data to be representative, it is necessary that test administration closely adhere to standardized procedures. Systematic departures from standardized procedures could produce over- or under-estimates of the ability levels in the youth population. The purpose of Part VII is to examine the test data as grouped by test-administration area supervisors, seeking evidence of systematic measurement disturbances.

It is clearly not appropriate simply to compare average scores across area supervisor groups. For any of a hundred reasons, average scores would be expected to vary from one area to the next, even without the slightest departure from standardized conditions; one area may have more younger subjects than another, or more disadvantaged subjects, or more subjects with lower motivation.

An ideal analysis would compare subjects' observed scores with their "true" scores, producing residuals that could be compared across area supervisor groups. In the absence of systematic area-supervisor measurement disturbances, theese residuals should average essentially zero in each group in each subtest.

This analysis is not possible. As a proxy for "true" scores, we have substituted the scores predicted by a subject's standing on five demographic variables: (1) sex, (2) racial/ethnic group, (3) age, (4) region of the country at age 14, and (5) mother's education, as a rough indication of socio-economic status.

As mentioned above, scores may differ across area supervisors for a hundred different valid reasons; we are able to account for five of them. The logic of the analysis, then, is extremely conservative. If no differences in our residuals are found across area supervisors, then there are no significant area supervisor effects. If differences are found, then area supervisor effects are one of the possible explanations for them.

## METHODOLOGY

Subject scores were computed for all subjects using the IRC methods described in Part I and Appendix B of this report. Using the MULTIVARIANCE computer program (Finn, 1977), we computed average scores in each subtest for the 480 cells of a five-way demographic classification scheme. The classificatory variables were as follows:

(1) Sex (male and female)

(2) Racial/ethnic group (Non-disadvantaged white, disadvantaged white, Black, Hispanic)

(3) Age (under 18, 18 to 20, over 20)

(4) Region of the country at age 14 (Southeast, West, Middle

West, Northeast)

(5) Mother's Education, in years of schooling (1-8, 9-11, 12, 13-15, 16+)

For each subject, a vector of residuals from his cell means was computed. These residuals were gathered by area supervisors. Only the 32 area supervisors with more than 100 subjects were retained, to insure the stability of cell residual averages. The residuals were rescaled within each subtest, to make the pooled within-supervisor variance of the residuals one in each subtest. This standardization facilitates comparisons across subtests as well as across supervisors.

The average residuals in each supervisor group in each subtest should be essentially zero if there are no supervisor effects and if there are no effects other than those explicitly accounted for in the model which are correlated with area supervisor groups. (This would include all other demographic effects, psychological effects, and environmental effects.) The expected variance of the mean of the residuals of a given supervisor group is obtained simply after the standardization described above: if there are N subjects in a group, the expected variance of the mean of the residuals in that group is 1/N. By use of the Central Limit Theorom, it is possible to use the normal distribution to obtain the .01, .001, and .0001 points of the expected distribution of the cell averages.

## RESULTS

Table 9 shows the average residuals of each area supervisor group in each subtest. It may be seen that the average residuals over all these supervisor groups--those containing at least 100 subjects--are essentially zero, implying that there are no systematic effects differentiating the subjects in large supervisor group from those in small supervisor groups. The variance of the supervisor-group residuals is about .02. In light of the standardization, we may conclude that about 2% of the variation among subjects not accounted for by the five demographic variables included in the analysis can be explained by supervisor effects or by effects correlated with supervisor groups. These effects cannot be separated with the available data.

It may be noted in particular that the variation from one supervisor group to the next is not appreciably greater in Numerical Operations and Coding Speed, the two speeded tests, than in the other subtests. Because test conditions and timing may be more critical in the speeded tests than in the power tests, this result tends to suggest that apparent differences are more likely explained by unspecified demographic and psychological effects than by test administration differences.

Table 10 highlights the particular subtests in particular supervisor groups where residuals are not statistically zero.

TABLE 9
AVERAGE RESIDUALS BY TEST ADMINISTRATION SUPERVISORS

| SUPERVISOR | N | SCIE | ARTH | WORD | PARA | NUMR | CODE | AUTO | MATH | MECH | ELEC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 15262 | 297 | -.218 | -.119 | -.139 | -.150 | -.027 | -.140 | -.318 | -.137 | -.255 | -.214 |
| 70317 | 213 | .280 | .210 | .208 | .203 | .020 | .123 | -.010 | .327 | .106 | .156 |
| 70626 | 224 | .011 | -.031 | -.031 | .028 | -.154 | -.096 | -.119 | -.023 | -.050 | -.013 |
| 75191 | 325 | -.154 | -.150 | -.291 | -.205 | -.307 | -.240 | -.142 | -.061 | -.137 | -.205 |
| 75299 | 514 | -.083 | -.060 | -.084 | -.095 | .012 | -.088 | .154 | -.152 | -.025 | -.051 |
| 75435 | 250 | .040 | .029 | -.015 | .042 | .040 | .061 | .030 | .130 | .059 | -.062 |
| 75499 | 381 | -.152 | -.126 | -.176 | -.161 | -.147 | -.214 | -.095 | -.048 | -.156 | -.147 |
| 75876 | 283 | .017 | .022 | .095 | .023 | -.026 | -.031 | -.223 | .066 | -.119 | -.042 |
| 76111 | 222 | -.043 | -.130 | -.095 | -.155 | -.099 | .030 | -.053 | -.122 | -.063 | .000 |
| 76208 | 525 | -.060 | -.072 | -.124 | -.045 | -.049 | -.082 | -.042 | -.064 | -.005 | -.076 |
| 76657 | 506 | -.115 | -.105 | -.195 | -.161 | -.075 | -.093 | -.101 | -.086 | -.135 | -.132 |
| 77132 | 504 | -.115 | -.074 | -.071 | -.021 | -.123 | -.118 | .051 | -.099 | .034 | -.011 |
| 78553 | 681 | .129 | .233 | .113 | .110 | .250 | .209 | .105 | .173 | .219 | .112 |
| 79135 | 207 | -.042 | .059 | .058 | .030 | -.005 | .079 | -.334 | .036 | -.219 | -.098 |
| 79529 | 401 | -.025 | -.143 | -.153 | -.032 | -.125 | -.069 | -.030 | -.127 | -.124 | -.053 |
| 79634 | 472 | -.080 | .003 | -.024 | .004 | -.060 | -.170 | -.461 | .099 | -.248 | -.089 |
| 80106 | 136 | -.046 | -.213 | -.079 | -.057 | .105 | .036 | -.253 | -.073 | -.353 | -.042 |
| 81831 | 204 | .172 | -.068 | .123 | .129 | -.112 | -.080 | .399 | -.061 | .228 | .185 |
| 81883 | 242 | .049 | .035 | .087 | .032 | .045 | .022 | .042 | .132 | .020 | .089 |
| 81919 | 327 | -.068 | -.071 | .002 | -.061 | -.010 | .021 | .043 | -.099 | -.085 | -.040 |
| 82167 | 340 | -.306 | -.216 | -.236 | -.229 | -.156 | -.210 | -.154 | -.274 | -.182 | -.288 |
| 82918 | 393 | .069 | .079 | .053 | -.053 | -.001 | -.067 | .184 | .016 | .169 | .073 |
| 83523 | 385 | -.015 | .074 | -.094 | -.076 | .105 | .063 | .016 | .031 | .000 | .029 |
| 87467 | 298 | .025 | -.014 | .057 | -.011 | -.033 | -.043 | .184 | -.115 | .133 | -.087 |
| 87607 | 328 | .023 | .002 | -.097 | -.085 | -.013 | -.021 | -.051 | .003 | .011 | -.038 |
| 88103 | 519 | .008 | .147 | .136 | .042 | .047 | .095 | -.196 | .186 | -.029 | -.005 |
| 89563 | 300 | .220 | .200 | .173 | .140 | .167 | .141 | .109 | .170 | .155 | .172 |
| 89778 | 369 | -.032 | .074 | .021 | .081 | .004 | -.046 | -.012 | .058 | .000 | .025 |
| 89985 | 400 | -.175 | -.065 | -.213 | -.127 | -.190 | -.147 | -.111 | -.082 | -.122 | -.121 |
| 90626 | 118 | .137 | -.006 | .195 | .044 | .303 | .135 | .119 | -.038 | .167 | .076 |
| 96208 | 124 | .105 | .076 | .187 | .208 | .241 | .302 | .111 | .082 | .058 | .143 |
| 99997 | 115 | .088 | -.042 | .140 | -.017 | .390 | .413 | -.059 | -.041 | -.016 | .098 |
| MEAN | | -.011 | -.014 | -.015 | -.020 | .001 | -.007 | -.038 | -.006 | -.030 | -.020 |
| STD DEV | | .126 | .114 | .138 | .111 | .149 | .147 | .174 | .124 | .144 | .115 |
| VARIANCE | | .016 | .013 | .019 | .012 | .022 | .022 | .030 | .015 | .021 | .013 |

NOTE: WITHIN-SUPERVISOR RESIDUALS HAVE A POOLED VARIANCE OF ONE.

TABLE 10
SIGNIFICANT AVERAGE RESIDUALS BY TEST ADMINISTRATION SUPERVISORS

| SUPERVISOR | N | SCIE | ARTH | WORD | PARA | NUMR | CODE | AUTO | MATH | MECH | ELEC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 15262 | 297 | -- | | | - | | | --- | | --- | -- |
| 70317 | 213 | +++ | + | + | + | | | | +++ | | |
| 70626 | 224 | | | | | | | | | | |
| 75191 | 325 | - | - | --- | --- | --- | --- | | | | -- |
| 75299 | 514 | | | | | | | ++ | -- | | |
| 75435 | 250 | | | | | | | | | | |
| 75499 | 381 | - | | -- | - | - | --- | | | - | - |
| 75876 | 283 | | | | | | | -- | | | |
| 76111 | 222 | | | | | | | | | | |
| 76208 | 525 | | | - | | | | | | | |
| 76657 | 506 | | | --- | -- | | | | | - | - |
| 77132 | 504 | | | | | - | - | | | | |
| 78553 | 681 | ++ | +++ | + | + | +++ | +++ | + | +++ | +++ | + |
| 79135 | 207 | | | | | | | | | - | |
| 79529 | 401 | | - | - | | | | | | | |
| 79634 | 472 | | | | | | -- | --- | | --- | |
| 80106 | 136 | | | | | | | - | | --- | |
| 81831 | 204 | | | | | | | +++ | | + | + |
| 81883 | 242 | | | | | | | | | | |
| 81919 | 327 | | | | | | | | | | |
| 82167 | 340 | --- | --- | --- | --- | - | -- | - | --- | -- | --- |
| 82918 | 393 | | | | | | | ++ | | ++ | |
| 83523 | 385 | | | | | | | | | | |
| 87467 | 298 | | | | | | | + | | | |
| 87607 | 328 | | | | | | | | | | |
| 88103 | 519 | | ++ | + | | | | --- | +++ | | |
| 89563 | 300 | ++ | ++ | + | | + | | | + | + | + |
| 89778 | 369 | | | | | | | | | | |
| 89985 | 400 | -- | | --- | | -- | - | | | | |
| 90626 | 118 | | | | | ++ | | | | | |
| 96208 | 124 | | | | | + | ++ | | | | |
| 99997 | 115 | | | | | +++ | +++ | | | | |

NOTE: +++ OR --- DENOTES SIGNIFICANCE AT P<.0001 IN INDICATED DIRECTION;
++ OR -- DENOTES SIGNIFICANCE AT P<.001 IN INDICATED DIRECTION;
+ OR - DENOTES SIGNIFICANCE AT P<.01 IN INDICATED DIRECTION.

Most of the entries are blank. Of the non-blank entries, attention may be focused on two specific groups, those of supervisors 78553 and 82167.

Given the demographic characteristics of the subjects in the group of Supervisor 78553, scores are systematically higher than expected; for the subjects in the group of supervisor 82167, scores are lower than expected. Depatures from standardized testing conditions are a possibility, but psychological effects are a stronger possibility. The area corresponding to Supervisor 78553 is the heartland of the Middle West, containing all or parts of Minnesota, South Dakota, Iowa, Wisconsin, and Nebraska. The area corresponding to Supervisor 82167 is Los Angeles, including the inner city with oversampling of Blacks, Hispanics, and disadvantaged whites. It is likely that both motivation and familiarity with testing are higher for subjects from the heartland of the Middle West than from the barrios of the Far West.

## CONCLUSIONS

A conservative test of area supervisor effects shows little evidence of variations from one test administration area to another. Those differences that do exist are more plausibly explained by differences in motivation and demography than by departures from standard test conditions.

## REFERENCES


Andersen, E.B. Discrete Statistical Models with Social Science Applications. Amsterdam: North Holland, 1980.

Birnbaum, A. "Some latent trait models and their use in inferring an examinee's ability." In F.M. Lord & M.R. Novick, Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Bock, R.D. "Basic issues in the measurement of change." In D.N.M. DeGruijter & L.J. van der Kamp,(Eds.), Advances in Psychological and Educational Measurement. London: Wiley, 1976.

Bock, R.D. & Aitkin, M. "Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm." Psychometrika (in press).

Bock, R.D., & Mislevy, R.J. BILOG: Maximum Likelihood Item Analysis and Test Scoring with Logistic Ogive Models. Chicago: International Educational Services, 1981.

Frankel, M.R. The Profile of American Youth: Technical Sampling Report. Chicago: National Opinion Research Center, 1981.

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

Lord, F.M. & Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

McWilliams, H. The Profile of American Youth: Field Report. Chicago: National Opinion Research Center, 1980.

Mead, R.J. Assessing the Fit of the Rasch Model through the Analysis of Residuals. Doctoral dissertation, University of Chicago, 1976.

Mislevy, R.J. A General Linear Model for the Analysis of Rasch Item Threshold Estimates. Doctoral dissertation, University of Chicago, 1981.

Mislevy, R.J. & Bock, R.D. "Biweight Estimates of Latent Ability." Educational and Psychological Measurement, 1982 (in press).

Rasch, G. Probabalistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danmarks Paedogogiske Institut, 1960; Chicago: University of Chicago Press, 1980.

Samejima, F. "Research on the multiple-choice test item in Japan: Toward the validation of mathematical models." Scientific Monograph ONRT-M3. Tokyo: Department of the Navy, Office of Naval Research Tokyo, 1980.

Wright, B.D. "Solving measurement problems with the Rasch model." Journal of Educational Measurement, 1977, 14, 96-116.

# APPENDIX A

# CALIBRATION OF POWER SUBTESTS

The items in each ASVAB power subtest were calibrated with the BILOG computer program (Bock & Mislevy, 1981) using an adaptation of the fixed-effects solution introduced in Bock (1976). Item parameter estimates were based on the responses of a 10-percent random sample (1200 subjects) of the NLS data, excluding those subjects not tested under standard administration procedures. This appendix outlines the calibration algorithm.

According to the assumptions of item response curve theory, item parameters are invariant across subjects and could be estimated from any sample, regardless of its ability or demographic features. This assumption is never satisfied perfectly in practice, however, so precautions were taken to guard against biases in item parameter estimates caused by the oversampling of disadvantaged subjects. Rather than weight each subject in the calibration sample equally, we have weighted them in proportion to their NLS sampling weights. The weights have been rescaled to make the sum of subject weights add to 1200, the actual number of subjects. The data from a given subject will be weighted inversely to his probability of being selected. In the following discussion, we use the following terms for a subject's item attempts and correct responses:

$N_{ij}$ = the weighted number of attempts by Subject i to Item j

= $W_{ij}$, and

$R_{ij}$ = the weighted number of correct responses by Subject i to Item j

= $W_{ij}$ if the response is correct and 0 otherwise.

As a consequence of this weighting, the item-fit and test-fit Chi-square statistics will not be strictly correct. It may be appropriate to adjust them in accordance with a design effect, probabably a value around two like the design effects for many other variables in the NLS survey. The resulting item parameter estimates will, however, correspond more closely to those that would be obtained from the responses of a true, simple random sample of the youth population.

## CALIBRATION ALGORITHM

### Step 1

An initial estimate of the ability of each subject in the calibration sample is obtained as the logit of correct response:

$$\hat{\theta}_i = \ln(\Sigma R_{ij}/(\Sigma N_{ij}-\Sigma R_{ij}).$$

If all the responses of Subject i are correct or all are incorrect, $W_{ij}$ is added to the number of attempts and $W_{ij}/2$ to the number correct.

### Step 2

Based on the provisional ability estimates, the calibration sample is partitioned into ten intervals as follows: The lowest and highest scoring five-percent of subjects are assigned to the

lowest and highest intervals, then the attainment scale between these extremes is divided into ten intervals of equal length.

It is assumed that the abilities of subjects in each interval are sufficiently similar to be approximated by a single interval ability. Numbers-tried and numbers-correct of Interval $\ell$ are defined by

$N_{\ell j}$ = the weighted number of attempts to Item j by subjects in Interval $\ell$, and

$R_{\ell j}$ = the weighted number of correct responses to Item j by subjects in Interval $\ell$.

Step 3

The probability of a correct response to Item j from a subject in Interval $\ell$ is assumed to be given by the logistic ogive:

$$P_{\ell j} = .10 + .90 \ \Psi \ (Z_{\ell j})$$
$$= .10 + .90 \ \exp(Z_{\ell j})/(1 + \exp(Z_{\ell j})),$$

where

$$Z_{\ell j} = (\theta_\ell - b_j)/s_j$$
$$= (1.0/s_j)\ \theta_\ell - (b_j/s_j)$$
$$= a_j\ \theta_\ell - c_j.$$

The item parameters $b_j$ and $s_j$ are the item threshold and dispersion discussed above. The re-expression in terms of the item parameters $a_j$ and $c_j$, the item slope and intercept, simplifies computation.

Assuming the local independence of responses to test items, the probability of observing $R_{\ell j}$ correct responses to Item j from the subjects in Interval $\ell$ is given by

$$P_{\ell j} = \text{Prob}(R_{\ell j} \mid N_{\ell j}, \theta_\ell)$$

$$= \frac{N_{\ell j}!}{R_{\ell j}!(N_{\ell j}-R_{\ell j})!} P_{\ell j}^{R_{\ell j}} (1-P_{\ell j})^{N_{\ell j}-R_{\ell j}}$$

and the probability of the entire calibration data matrix becomes

$$P = \prod_{1}^{10} \prod_{j}^{n} P_{\ell j}$$

Estimates of the $a_j$'s, $c_j$'s, and $\theta_\ell$'s are chosen to maximize this probability. The log likelihood function is

$$L = \prod_{1}^{10} \prod_{j}^{n} C + R_{\ell j} \ln(P_{\ell j}) + (N_{\ell j}-R_{\ell j}) \ln(P_{\ell j}),$$

where C does not depend on the parameters. The likelihood equations for $\ell = 1, 2, \ldots, 10$ and $j = 1, 2, \ldots, n$ are

$$c_j: \sum_{1} (R_{\ell j} - N_{\ell j} P_{\ell j}) = 0$$

$$a_j: \sum_{1} (R_{\ell j} - N_{\ell j} P_{\ell j}) \theta_\ell = 0$$

$$\theta_\ell: \sum_{j} (R_{\ell j} - N_{\ell j} P_{\ell j}) a_j = 0$$

In order to fix the size and origin of the provisional scale units, the highest and lowest intervals are assigned scores of plus one and minus one respectively. BILOG solves the reduced equations by means of Newton-Raphson iteration.

Step 4

From the provisional item parameters estimated in the preceding step, each subject's scale score is estimated. The appropriate likelihood equation, under the assumption of local independence, is given by

$$\theta_i : \sum (R_{ij} - N_{ij} P_{ij}) a_j = 0,$$

where

$$P_{ij} = .10 + .90 \exp(Z_{ij}) / (1 + \exp(Z_{ij}))$$

with

$$Z_{ij} = a_j \theta_i - c_j.$$

This equation has no solution if all of the responses of Subject i are correct or all are incorrect. In the former case, $W_{ij}$ is added to the number of attempts to the item with the highest threshold and $W_{ij}/2$ is added to his number correct; in the latter case, the same procedure is applied to the item with the lowest threshold.

Step 5

Step 2 is repeated with the improved subject score estimates.

Step 6

Step 3 is repeated with the improved interval boundaries.

Step 7

Standard errors of estimation for the item parameters are obtained in the final Newton iteration of Step 6, as the square roots of the negative reciprocals of the second derivatives of the log likelihood at the final solution.

Item fit is indicated by a Pearsonian chi-square over intervals:

$$X_j^2 = \Sigma \ (R_{\ell j} - N_{\ell j} P_{\ell j})^2 / [N_{\ell j} P_{\ell j} (1-P_{\ell j})].$$

If the expected number of either correct or incorrect responses to Item j in Interval $\ell$ is less than 5, the Interval $\ell$ is collapsed into an adjacent interval for the purpose of the item-fit index. The number of degrees of freedom associated with the value is two less than the number of intervals after collapsing.

Overall test fit is indicated by the sum of the item-fit

chi-squares, with degrees of freedom similarly summed but reduced by 8 to account for the estimation of interval scores.

As noted above, it may be appropriate to divide the resulting item and test fit Chi-squares by two to account for the stratified sampling design.

# DATA QUALITY ANALYSIS OF THE ARMED SERVICES VOCATIONAL APTITUDE BATTERY

R. Darrell Bock
University of Chicago

and

Robert J. Mislevy
National Opinion Reseach Center

May, 1981

# APPENDIX B

# BIWEIGHT ESTIMATES OF LATENT ABILITY

Maximum likelihood estimates of subjects' abilities in item response curve models are overly senstive to disturbances that are common in educational measurement, such as guessing and carelessness. The biweight solution described in this appendix, introduced and illustrated in Mislevy and Bock (1980), is highly resistent to these disturbances. It effectively discounts suspicious responses, and agrees with the maximum likelihood estimator when none are present.

We first review the form of maximum likelihood estimation of latent ability. Our final estimates of subject abilities in the power subtests of the ASVAB employ two variations on this basic theme, namely, the use of a prior distribution and biweighting.

## MAXIMUM LIKELIHOOD ESTIMATES OF LATENT ABILITY

Suppose that the item parameter of n test items are known. Let $b_j$ be the threshold of Item j and $s_j$ be the dispersion. We observe the responses of Subject i to these items. Let $X_{ij}$ be one if the response to Item j is correct and zero if it is not.

Under the assumptions of the 2-parameter logistic item response curve model, the probability that Subject i will respond correctly to Item j is given by

$$P_{ij} = \exp(Z_{ij}) \; / \; (1 + \exp(Z_{ij})), \qquad \text{(B.1)}$$

where

$$Z_{ij} = (\theta_i - b_j)/s_j$$

and $\theta_i$ denotes the ability of Subject i. Assuming the responses of Subject i are independent, given $\theta_i$, the probability of his vector of responses is given by the product of expressions like (B.1) over all the items:

$$P_i = \prod_{j=1}^{n} P_{ij}^{X_{ij}} \, (1-P_{ij})^{1-X_{ij}}. \qquad \text{(B.2)}$$

If the item parameters are known but the ability is not, then (B.2) is the likelihood function of $\theta_i$ given the vector of responses. The maximum likelihood estimate of the ability, $\hat{\theta}_i$, is the value which maximizes (B.2) with respect to the observed responses.

In practice the log of the likelihood is maximized. The log likelihood is given by

$$\ell n \; L = \sum_{j=1}^{n} C + X_{ij} \, \ell n(P_{ij}) + (1-X_{ij}) \, \ell n(1-P_{ij}).$$

where C does not depend on the parameters. The first derivative of the log likelihood function is given by

$$\frac{d\ell nL}{d\theta_i} = \sum_{j=1}^{n} (X_{ij}-P_{ij})/s_j. \qquad \text{(B.3)}$$

and its second derivative, by

$$\frac{d\,\ell nL}{d\theta_i} = - \sum_{j=1}^{n} P_{ij} \, (1-P_{ij})/ \, s_j^{\,2}.$$

As long as not all of the responses are correct and not all are incorrect, the is a unique and finite value for which the

first derivative is zero. Since the second derivative is always negative, the zero is a maximum of the log likelihood. A large-sample standard error for the estimate may be obtained as the negative reciprocal of the square root of the second derivative of $\ell n$ L, evaluated at the maximum.

## BAYES MODAL ESTIMATES

As noted in Part I of this report, the scale of the item parameters was fixed by requiring the mean of the youth population to be zero and the true-score variation to be one. Under the assumption that the underlying distribution is normal, it is possible to use Bayes Theorom to obtain estimates of subject abilities with lower mean-squared errors than maximum likelihood estimates.

Under this scheme, the prior density of $\theta_i$ is Normal (0,1). The posterior density, F, is proportional to the prior density times the likelihood (B.2):

$$F = \{ \prod_{j=1}^{n} P_{ij}^{X_{ij}} (1-P_{ij})^{1-X_{ij}} \} \times \{ \frac{-1}{2\Pi} \exp(-\theta_i^2/2) \}$$

where $P_{ij}$ is as defined in (B.1). The value of $\theta_i$ that minimizes this expression is the Bayes modal estimate of $\tilde{\theta}_i$, the highest value with the highest posterior density. The log of F and its first and second derivatives are nearly the same as those for L, except for additional terms:

$$\ell n\ F = \sum_{j=1}^{n} K + X_{ij}\, \ell n(P_{ij}) + (1-X_{ij})\, \ell n(1-P_{ij}) - \theta_i^2/2.$$

$$\frac{d\ell nF}{d\theta} = \theta_i + \sum_{j=1}^{n} (X_{ij}-P_{ij})/s_j.$$

$$\frac{d^2\ell nF}{d\theta^2} = -1 - \sum_{j=1}^{n} P_{ij}\,(1-P_{ij})/\,s_j^{\,2}.$$

It is typical to take as an indication of the precision of estimation the negative reciprocal of the square root of the second derivative at the maximum; i.e., the curvature of the posterior distribution at its highest point.

## BIWEIGHT ESTIMATES

In theory, a subject's responses to items with thresholds far above or far below his level of ability provide little information about his ability. In practice, they may provide misinformation. An incorrect response to an easy item from a subject who otherwise appears quite able is probably a careless error; a correct response to a hard item from a subject who otherwise appears unable is probably a lucky guess. Inasmuch as a subject's responses to items far from his apparent ability contain least information and most potential for misinformation, it would be desirable to weight a subject's responses accordingly. The biweight estimator described in this section does just that.

Based on the principal of Tukey's biweight estimate of location, the biweight estimate of ability responds sensitively to information from items in the neighborhood of the subject's apparent ability, while effectively discounting responses to items far above or below this level. The response of Subject i to Item j is

assigned the weight $W_{ij}$ in accordance with the distance of the subject from the item, in units of the item's dispersion:

$$W_{ij} = \begin{cases} (1 - U_{ij}^2)^2 & \text{if } |U_{ij}| < 1 \\ 0 & \text{otherwise} \end{cases}$$

with

$$U_{ij} = \frac{b_j - \hat{\theta}_i}{3\, s_j} .$$

In this last expression, $\hat{\theta}_i$ represents the biweight estimate of the ability of Subject i. (The biweight estimate depends on the weights and the weights depend on the estimate; together they must be computed iteratively.)

The fitting function used in the computation of the biweight estimate is a modification of (B.3), the likelihood equation:

$$G' = \sum_{j=1}^{n} W_{ij} (X_{ij} - P_{ij})/s_j .$$

As an indication of the precision of estimation, one may use the negative reciprocal of the square root of a facsimile of a second derivative:

$$G'' = - \sum_{j=1}^{n} W_{ij} P_{ij} (1-P_{ij})^2 / s_j .$$

This quantity would be the second derivative of a log likelihood if the weights $W_{ij}$ had been specified in advance rather than in response to the data.

## BIWEIGHTED BAYES ESTIMATES

Final estimates of subject abilities in the ASVAB power tests use both the standard normal prior, to provide lower mean-squared errors, and biweighting, to trim potentially misleading responses to extreme items. The fitting function incorporates aspects of the Bayes and the biweight estimates:

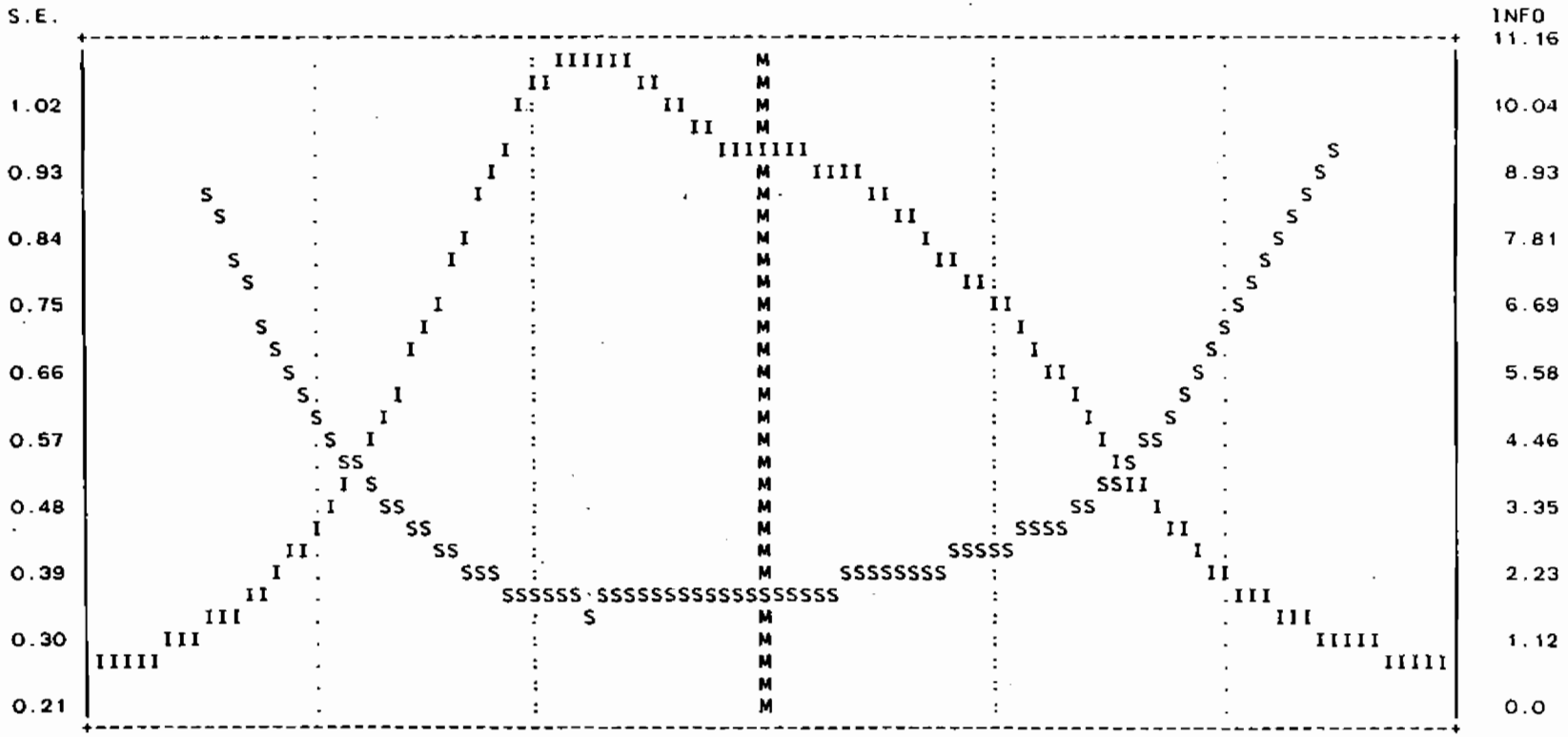
$$H' = \theta_i + \sum_{j=1}^{n} W_{ij} (X_{ij}-P_{ij})/s_j,$$

where $W_{ij}$ is the biweight. As a standard error of estimation, we use the negative reciprocal of the square root of a facsimile of a a second derivative:

$$H'' = -1 - \sum_{j=1}^{n} W_{ij} P_{ij} (1-P_{ij})^2/ s_j .$$

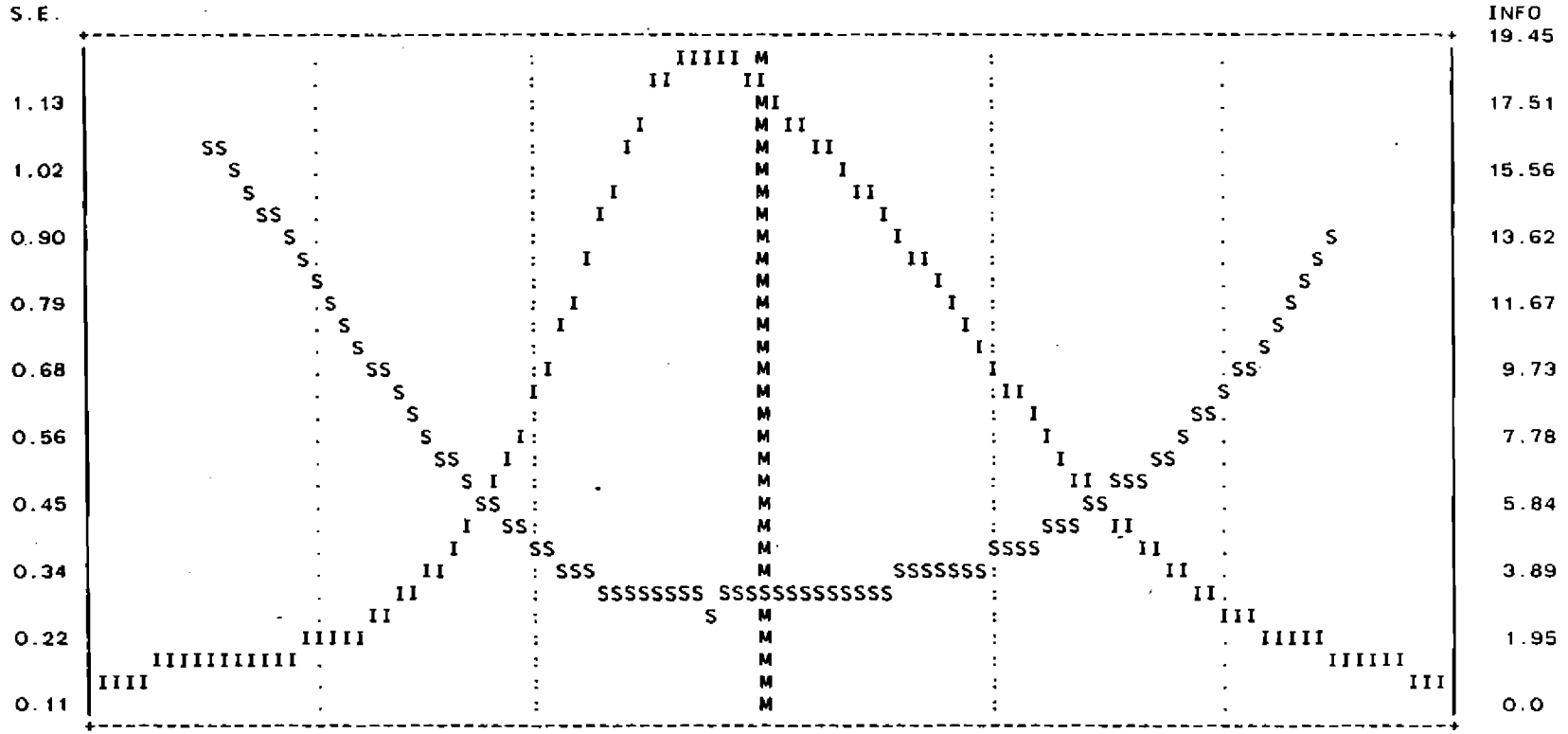
# APPENDIX C

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: GENERAL SCIENCE



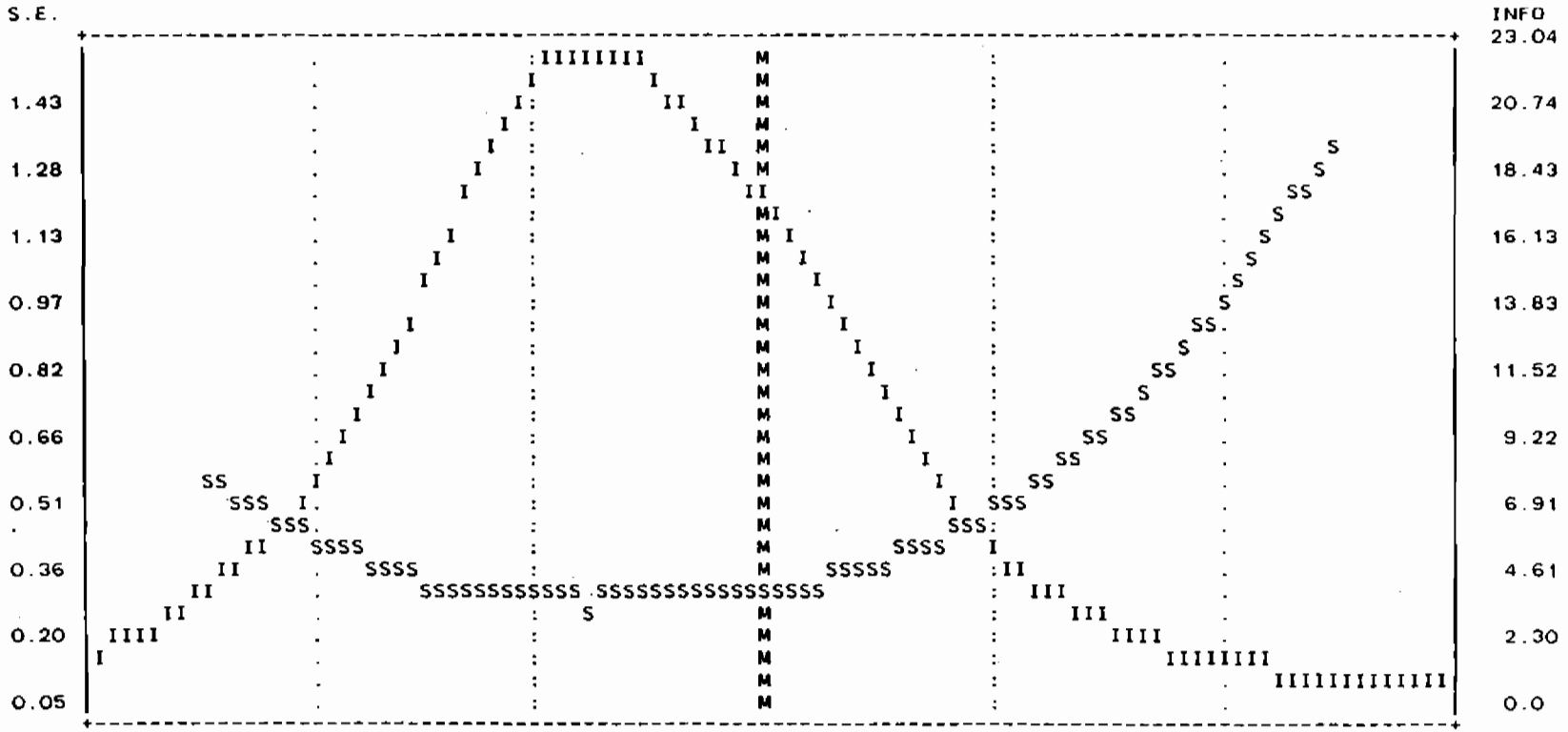

| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 3.78 | 4.87 | 6.79 | 9.69 | 13.05 | 16.17 | 18.97 | 21.26 | 22.90 | 23.87 | 24.39 |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: ARITHMETIC REASONING



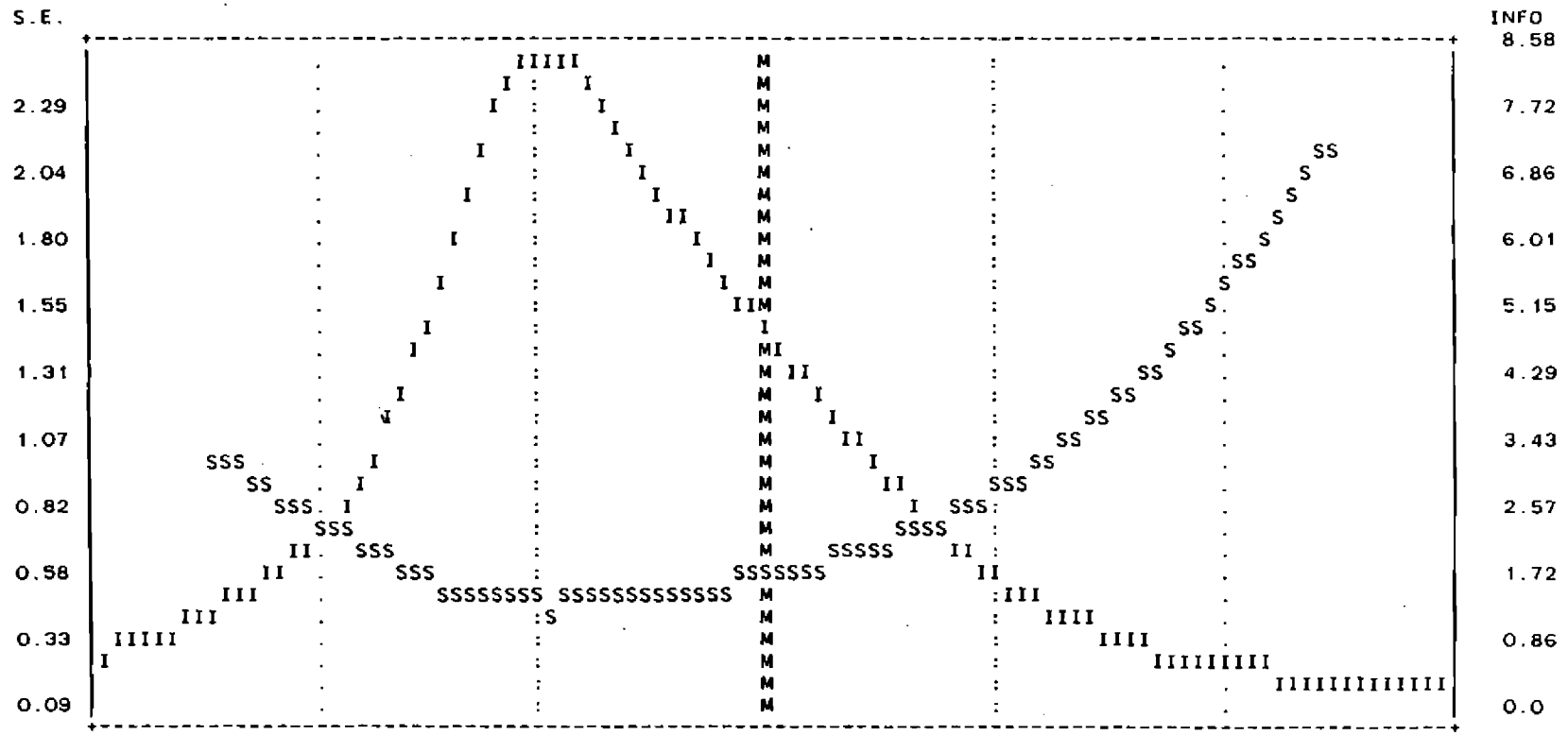

| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 4.48 | 5.25 | 6.55 | 8.96 | 12.98 | 17.88 | 22.29 | 25.68 | 27.83 | 28.99 | 29.54 |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: WORD KNOWLEDGE



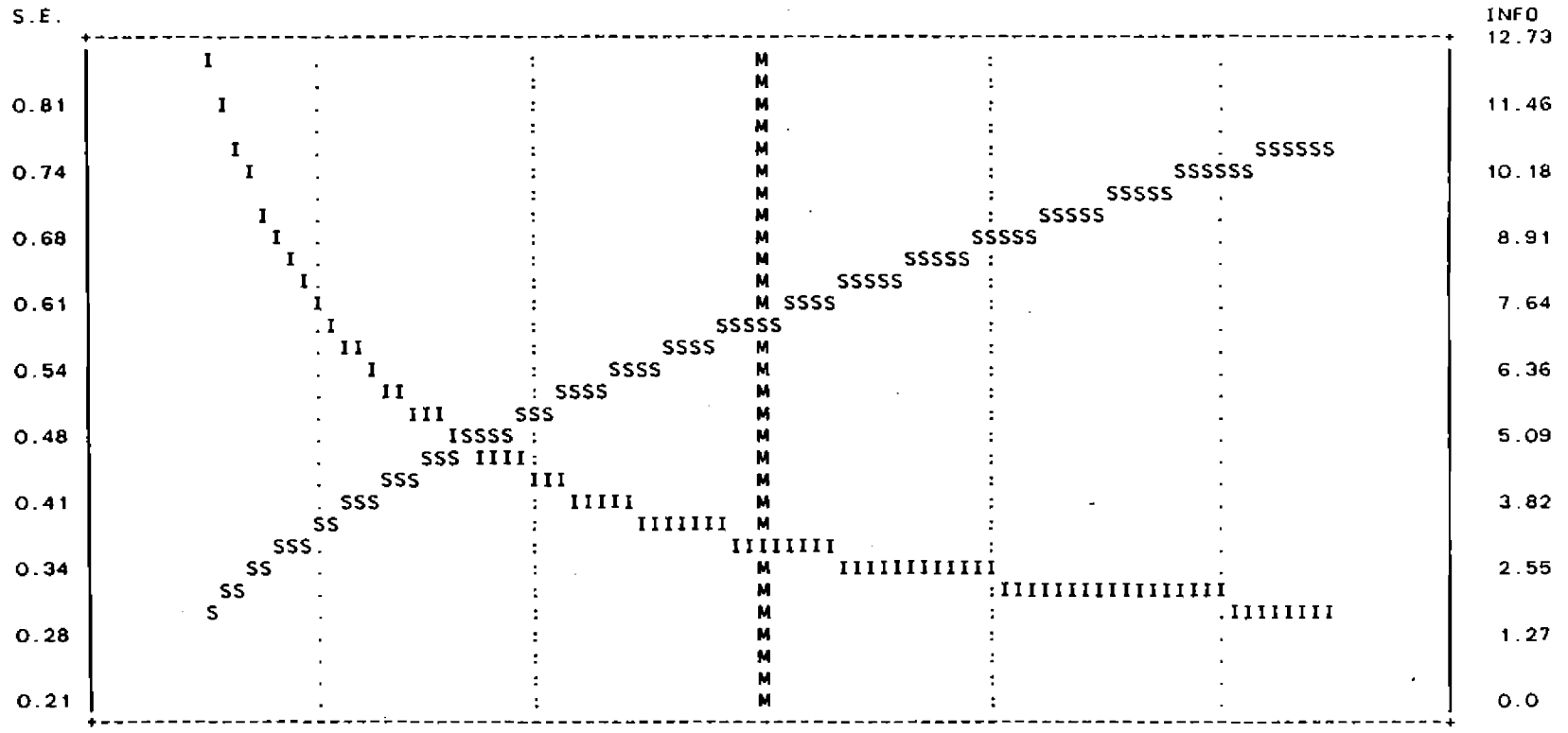

| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 5.82 | 8.00 | 11.61 | 16.78 | 22.47 | 27.33 | 30.81 | 32.83 | 33.89 | 34.42 | 34.70 |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: PARAGRAPH COMPREHENSION



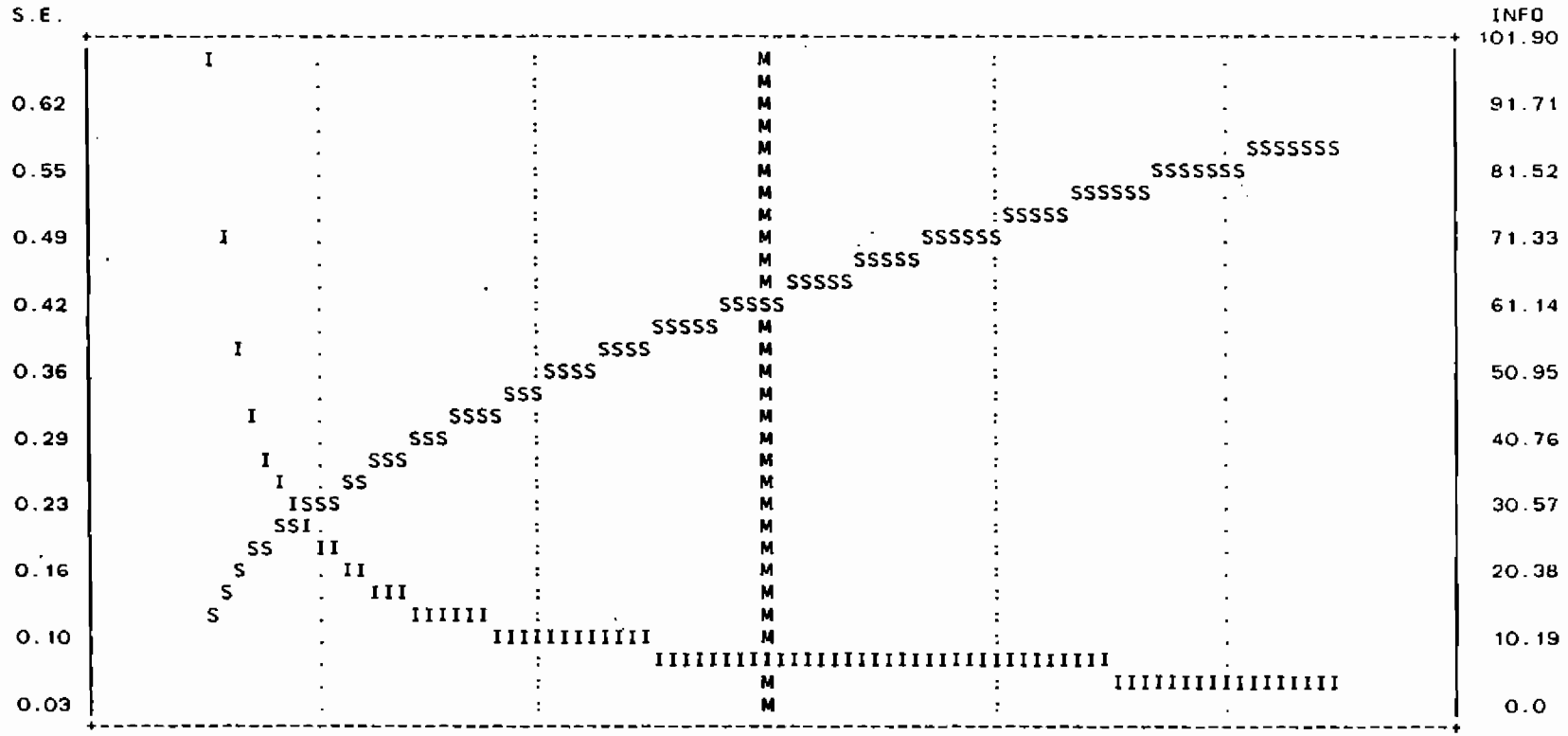

| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 3.09 | 4.00 | 5.44 | 7.60 | 9.86 | 11.67 | 12.92 | 13.68 | 14.11 | 14.34 | 14.48 |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: NUMERICAL OPERATIONS



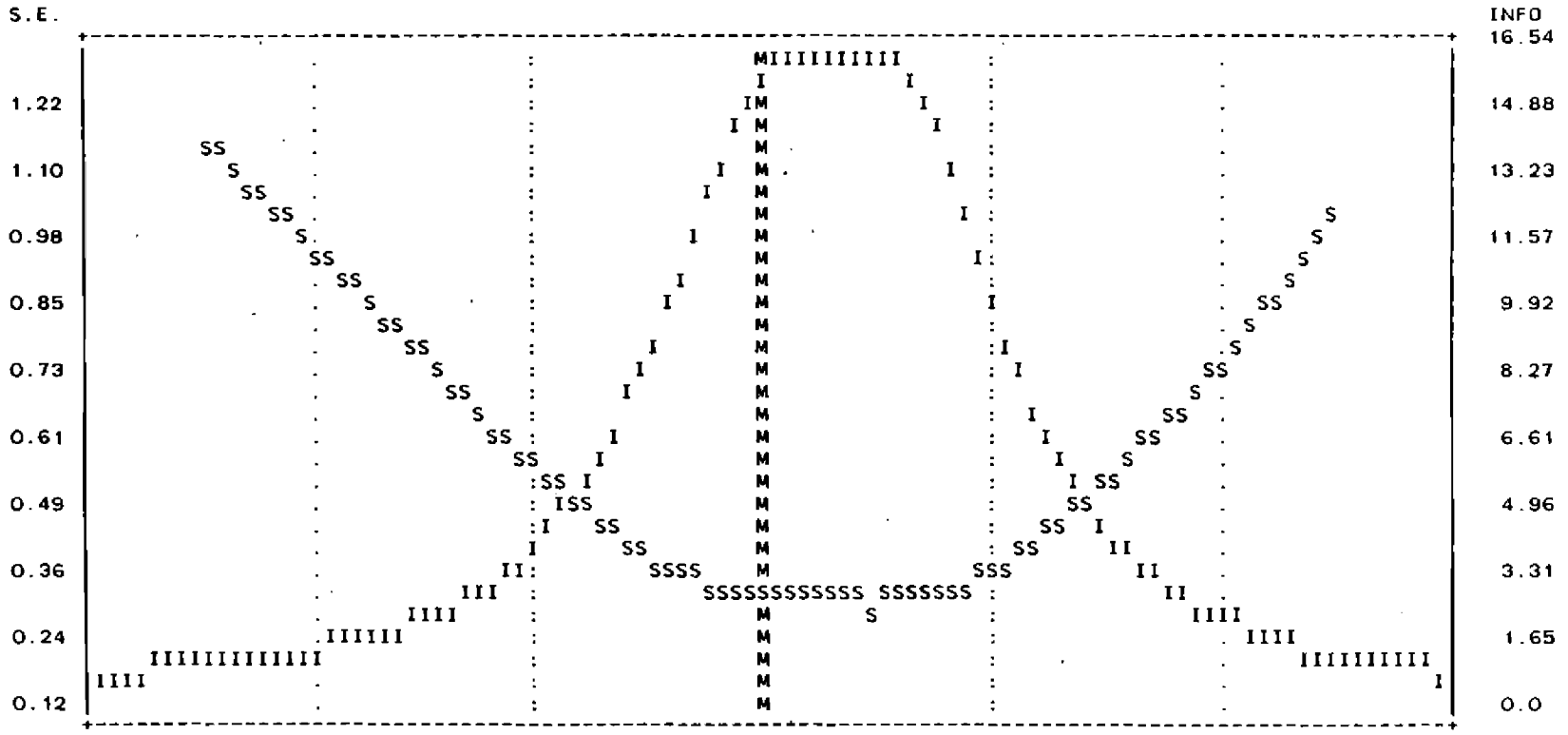

| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 10.78 | 15.41 | 20.03 | 24.65 | 29.28 | 33.90 | 38.52 | 43.14 | 47.77 | 50.00+ | |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: CODING SPEED



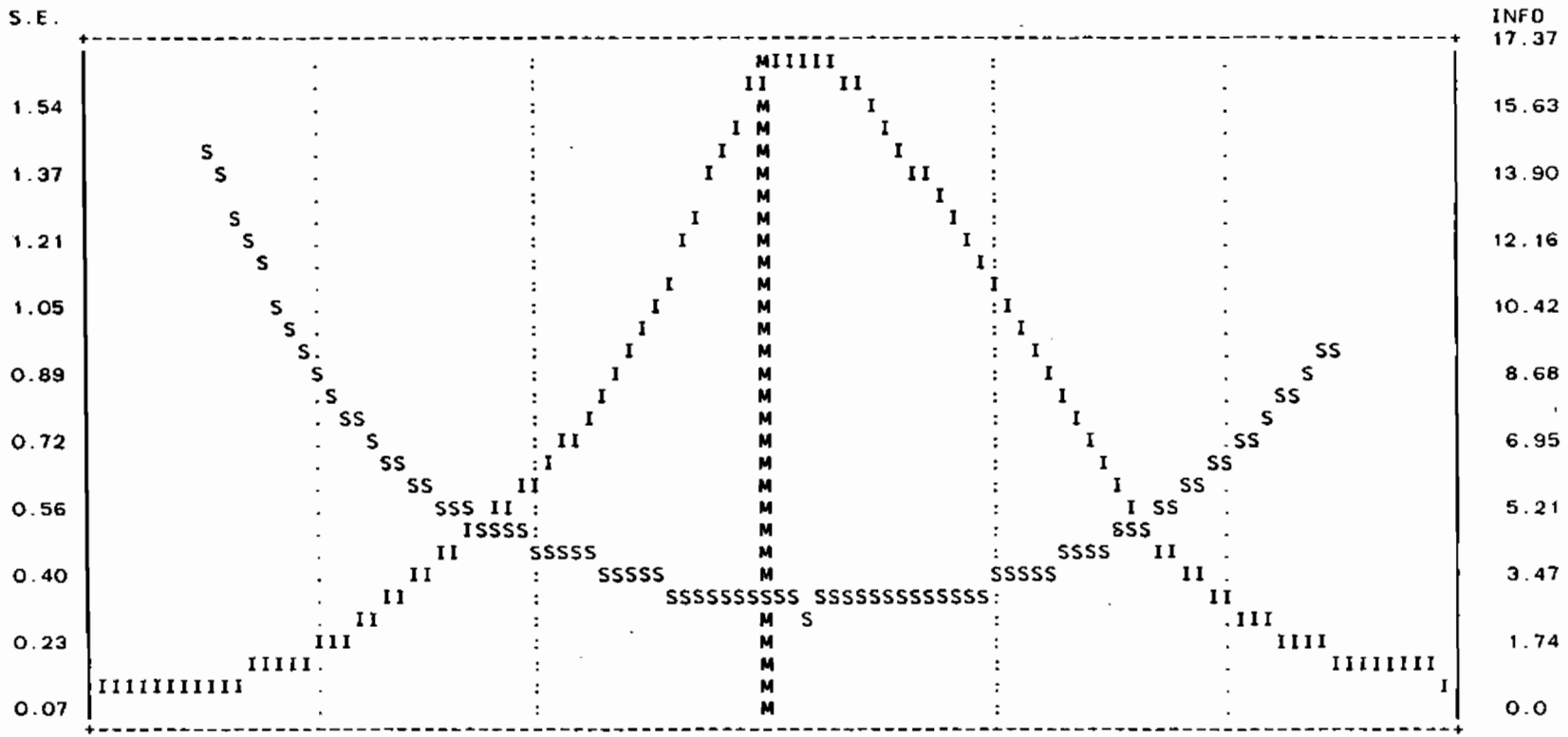

| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 8.51 | 15.80 | 23.09 | 30.38 | 37.68 | 44.97 | 52.26 | 59.56 | 66.85 | 74.14 | 81.43 |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: AUTO & SHOP INFORMATION



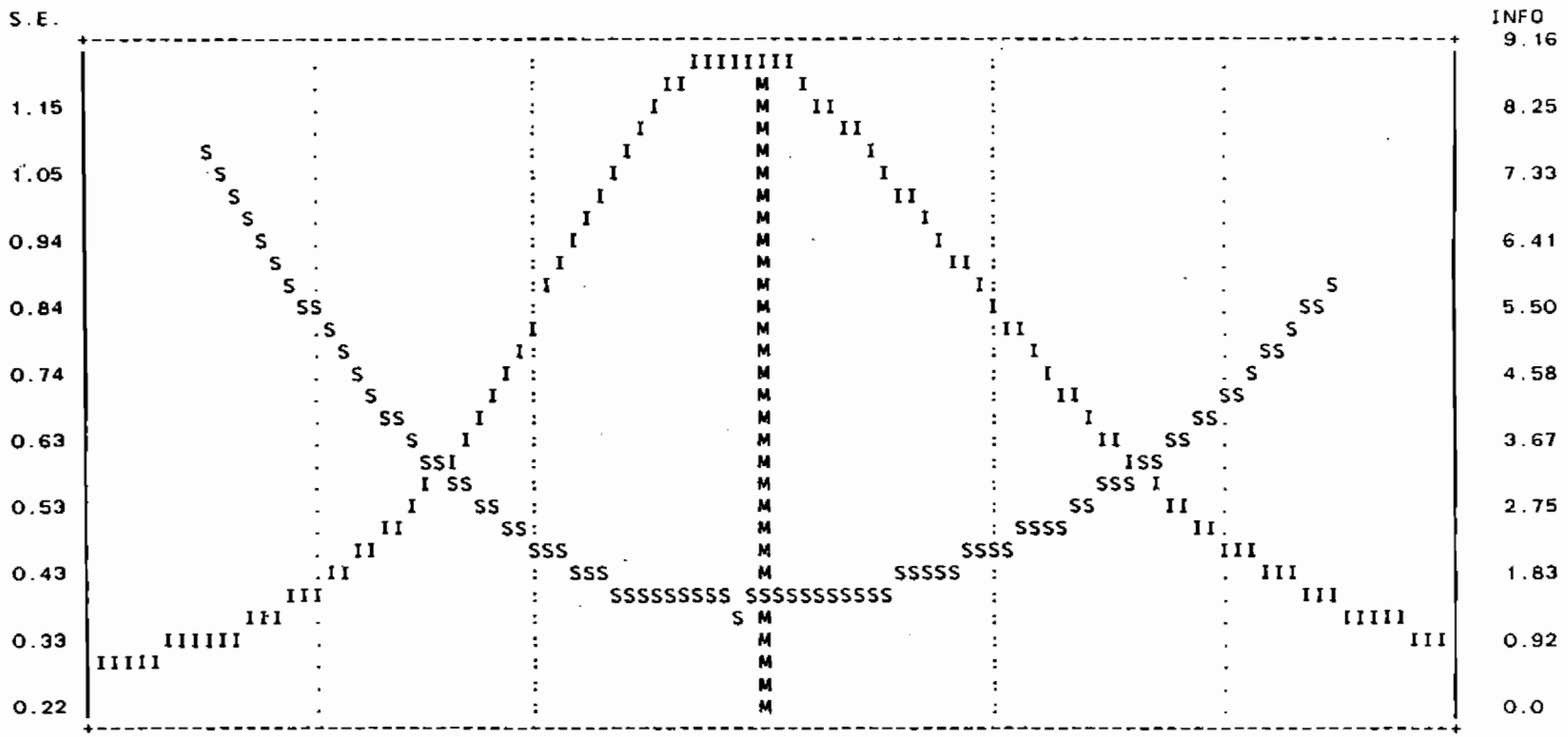

| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 4.35 | 5.07 | 6.10 | 7.63 | 10.04 | 13.59 | 17.55 | 20.75 | 22.54 | 23.50 | 24.05 |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: MATHEMATICS KNOWLEDGE



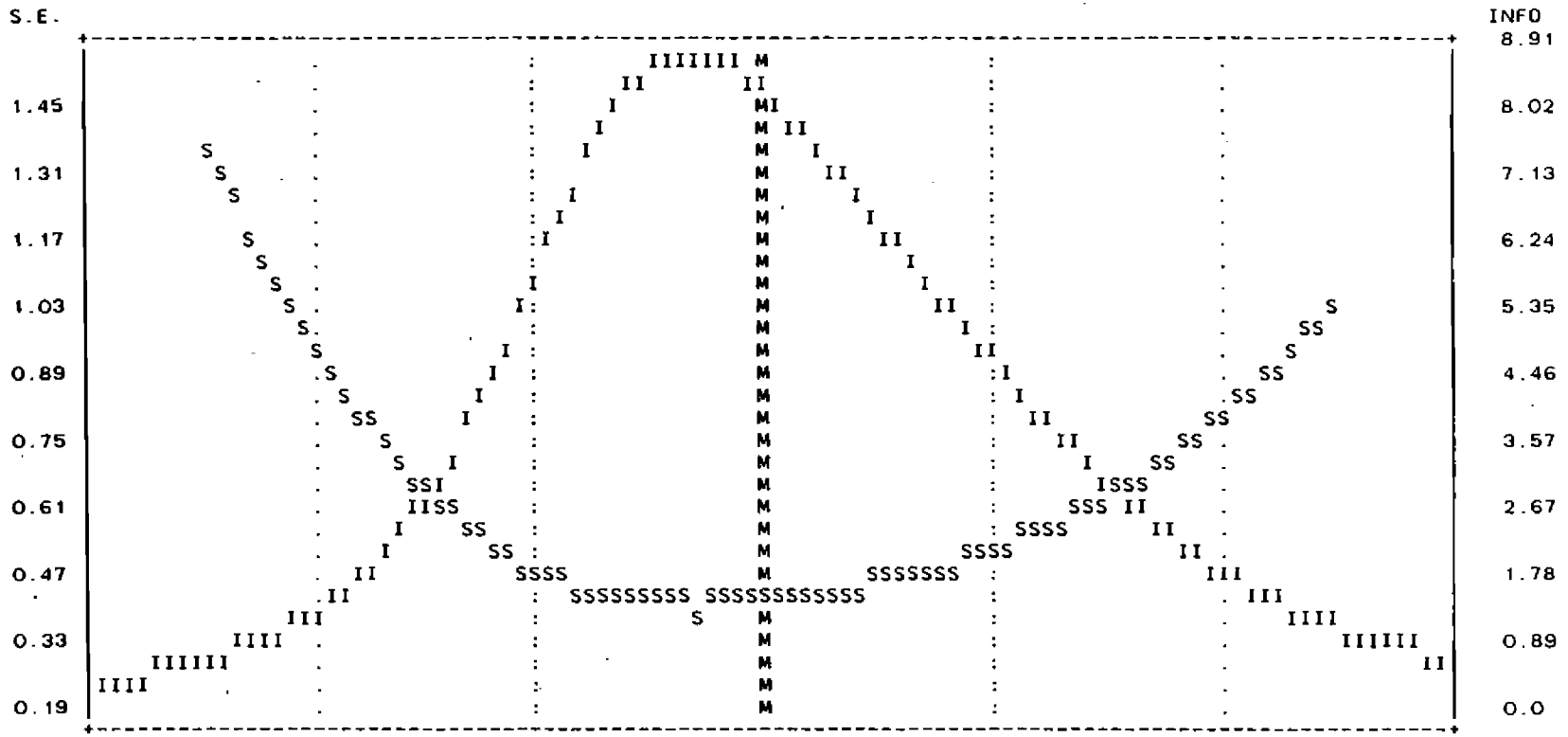

| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 3.15 | 3.71 | 4.78 | 6.60 | 9.33 | 13.15 | 17.28 | 20.58 | 22.71 | 23.83 | 24.38 |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: MECHANICAL COMPREHENSION




| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 4.10 | 4.97 | 6.32 | 8.36 | 11.19 | 14.45 | 17.50 | 19.96 | 21.75 | 22.94 | 23.70 |

TEST INFORMATION CURVE AND STANDARD ERRORS
TEST NAME: ELECTRONICS INFORMATION




| SCALE SCORE | -2.50 | -2.00 | -1.50 | -1.00 | -0.50 | 0.0 | 0.50 | 1.00 | 1.50 | 2.00 | 2.50 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| EXPECTED NUMBER-CORRECT SCORE | 2.83 | 3.43 | 4.49 | 6.27 | 8.82 | 11.61 | 14.05 | 15.96 | 17.30 | 18.17 | 18.70 |

# APPENDIX D

TABLE D-1
ITEM PARAMETER ESTIMATES FOR GENERAL SCIENCE
TOTAL CHI-SQUARE = 315.97 WITH 171.00 DF

| | PARAMETER ESTIMATES | | | | | ITEM FIT STATISTICS | | |
|---|---|---|---|---|---|---|---|---|
| ITEM | THRESHOLD | (S.E.) | DISPERSION | (S.E.) | GUESSING | CHI-SQ | D.F. | PROB |
| 1 | -1.48 | ( 0.11) | 0.42 | ( 0.04) | 0.10 | 1.91 | 4.80 | 0.844 |
| 2 | -1.03 | ( 0.10) | 0.54 | ( 0.05) | 0.10 | 5.64 | 5.80 | 0.435 |
| 3 | -0.81 | ( 0.09) | 0.36 | ( 0.03) | 0.10 | 4.68 | 4.80 | 0.429 |
| 4 | -0.85 | ( 0.10) | 0.69 | ( 0.06) | 0.10 | 16.46 | 6.70 | 0.018 |
| 5 | -1.00 | ( 0.11) | 0.63 | ( 0.06) | 0.10 | 7.94 | 6.70 | 0.311 |
| 6 | -0.92 | ( 0.09) | 0.30 | ( 0.03) | 0.10 | 1.21 | 4.80 | 0.934 |
| 7 | -0.73 | ( 0.10) | 0.63 | ( 0.05) | 0.10 | 19.42 | 6.70 | 0.006 |
| 8 | -1.56 | ( 0.14) | 0.82 | ( 0.08) | 0.10 | 23.12 | 6.70 | 0.001 |
| 9 | -0.93 | ( 0.10) | 0.53 | ( 0.05) | 0.10 | 6.70 | 5.80 | 0.324 |
| 10 | -0.93 | ( 0.10) | 0.61 | ( 0.05) | 0.10 | 4.39 | 6.70 | 0.705 |
| 11 | -0.70 | ( 0.09) | 0.55 | ( 0.05) | 0.10 | 13.74 | 6.70 | 0.049 |
| 12 | 0.13 | ( 0.09) | 1.03 | ( 0.10) | 0.10 | 25.61 | 7.70 | 0.001 |
| 13 | 0.05 | ( 0.07) | 0.45 | ( 0.04) | 0.10 | 17.66 | 6.70 | 0.011 |
| 14 | -0.53 | ( 0.10) | 0.79 | ( 0.07) | 0.10 | 8.47 | 7.70 | 0.357 |
| 15 | -0.92 | ( 0.13) | 1.05 | ( 0.11) | 0.10 | 3.85 | 7.70 | 0.849 |
| 16 | -0.72 | ( 0.11) | 0.84 | ( 0.08) | 0.10 | 30.94 | 7.70 | 0.0 |
| 17 | -0.56 | ( 0.10) | 0.93 | ( 0.09) | 0.10 | 2.79 | 7.70 | 0.936 |
| 18 | 0.33 | ( 0.07) | 0.31 | ( 0.03) | 0.10 | 25.64 | 6.70 | 0.0 |
| 19 | 0.25 | ( 0.08) | 0.75 | ( 0.07) | 0.10 | 16.53 | 7.70 | 0.030 |
| 20 | 0.10 | ( 0.08) | 0.64 | ( 0.05) | 0.10 | 5.95 | 7.70 | 0.620 |
| 21 | 0.72 | ( 0.08) | 0.72 | ( 0.07) | 0.10 | 7.86 | 7.70 | 0.414 |
| 22 | 0.83 | ( 0.09) | 0.79 | ( 0.08) | 0.10 | 3.06 | 7.70 | 0.916 |
| 23 | 0.91 | ( 0.09) | 0.80 | ( 0.08) | 0.10 | 6.09 | 7.70 | 0.604 |
| 24 | 1.12 | ( 0.06) | 0.33 | ( 0.04) | 0.10 | 18.84 | 7.70 | 0.013 |
| 25 | 0.75 | ( 0.07) | 0.53 | ( 0.05) | 0.10 | 37.45 | 7.70 | 0.0 |

NOTES: GUESSING PARAMETERS FIXED DURING ESTIMATION.
DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR IN CHI-SQUARES.

TABLE D-2
ITEM PARAMETER ESTIMATES FOR ARITHMETIC REASONING
TOTAL CHI-SQUARE = 476.45 WITH 208.00 DF

| | PARAMETER ESTIMATES | | | | | ITEM FIT STATISTICS | | |
|---|---|---|---|---|---|---|---|---|
| ITEM | THRESHOLD | (S.E.) | DISPERSION | (S.E.) | GUESSING | CHI-SQ | D.F. | PROB |
| 1 | -2.74 | ( 0.30) | 1.07 | ( 0.14) | 0.10 | 7.04 | 6.80 | 0.399 |
| 2 | -2.28 | ( 0.19) | 0.77 | ( 0.08) | 0.10 | 21.75 | 4.80 | 0.001 |
| 3 | -0.79 | ( 0.12) | 0.35 | ( 0.03) | 0.10 | 4.98 | 4.80 | 0.396 |
| 4 | -0.64 | ( 0.11) | 0.33 | ( 0.03) | 0.10 | 8.80 | 4.80 | 0.108 |
| 5 | -0.41 | ( 0.11) | 0.48 | ( 0.04) | 0.10 | 10.88 | 6.80 | 0.130 |
| 6 | -0.69 | ( 0.12) | 0.57 | ( 0.05) | 0.10 | 17.77 | 6.80 | 0.011 |
| 7 | -1.41 | ( 0.17) | 1.16 | ( 0.12) | 0.10 | 6.08 | 7.70 | 0.611 |
| 8 | -0.79 | ( 0.12) | 0.59 | ( 0.05) | 0.10 | 10.42 | 6.80 | 0.151 |
| 9 | -0.65 | ( 0.12) | 0.56 | ( 0.05) | 0.10 | 12.30 | 6.80 | 0.082 |
| 10 | -0.50 | ( 0.11) | 0.40 | ( 0.04) | 0.10 | 11.82 | 5.80 | 0.059 |
| 11 | -0.35 | ( 0.10) | 0.23 | ( 0.02) | 0.10 | 11.32 | 4.80 | 0.041 |
| 12 | -0.18 | ( 0.10) | 0.50 | ( 0.04) | 0.10 | 8.24 | 6.80 | 0.290 |
| 13 | -0.25 | ( 0.11) | 0.69 | ( 0.06) | 0.10 | 12.80 | 7.70 | 0.107 |
| 14 | -0.18 | ( 0.10) | 0.40 | ( 0.04) | 0.10 | 6.60 | 6.80 | 0.446 |
| 15 | -0.31 | ( 0.10) | 0.41 | ( 0.04) | 0.10 | 7.61 | 5.80 | 0.249 |
| 16 | -0.18 | ( 0.10) | 0.73 | ( 0.06) | 0.10 | 16.05 | 7.70 | 0.037 |
| 17 | 0.13 | ( 0.09) | 0.70 | ( 0.06) | 0.10 | 35.05 | 7.70 | 0.0 |
| 18 | 0.21 | ( 0.09) | 0.64 | ( 0.05) | 0.10 | 49.56 | 7.70 | 0.0 |
| 19 | -0.05 | ( 0.10) | 0.69 | ( 0.06) | 0.10 | 18.52 | 7.70 | 0.015 |
| 20 | 0.61 | ( 0.08) | 0.48 | ( 0.04) | 0.10 | 26.36 | 7.70 | 0.001 |
| 21 | 0.51 | ( 0.08) | 0.50 | ( 0.04) | 0.10 | 24.21 | 7.70 | 0.002 |
| 22 | 0.01 | ( 0.09) | 0.53 | ( 0.05) | 0.10 | 9.01 | 7.70 | 0.316 |
| 23 | 0.13 | ( 0.09) | 0.70 | ( 0.06) | 0.10 | 14.69 | 7.70 | 0.058 |
| 24 | 0.41 | ( 0.08) | 0.34 | ( 0.03) | 0.10 | 34.82 | 6.80 | 0.0 |
| 25 | 0.34 | ( 0.08) | 0.52 | ( 0.04) | 0.10 | 24.35 | 7.70 | 0.002 |
| 26 | 0.36 | ( 0.08) | 0.59 | ( 0.05) | 0.10 | 9.45 | 7.70 | 0.282 |
| 27 | 0.53 | ( 0.08) | 0.44 | ( 0.04) | 0.10 | 12.41 | 7.70 | 0.120 |
| 28 | 1.06 | ( 0.07) | 0.49 | ( 0.05) | 0.10 | 10.39 | 7.70 | 0.219 |
| 29 | 1.16 | ( 0.07) | 0.51 | ( 0.05) | 0.10 | 4.37 | 7.70 | 0.802 |
| 30 | 0.69 | ( 0.08) | 0.70 | ( 0.06) | 0.10 | 28.78 | 7.70 | 0.0 |

NOTES: GUESSING PARAMETERS FIXED DURING ESTIMATION.
DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR IN CHI-SQUARES.

TABLE D-3
ITEM PARAMETER ESTIMATES FOR WORD KNOWLEDGE
TOTAL CHI-SQUARE = 499.87 WITH 220.00 DF

| | PARAMETER ESTIMATES | | | | | ITEM FIT STATISTICS | | |
|---|---|---|---|---|---|---|---|---|
| ITEM | THRESHOLD | (S.E.) | DISPERSION | (S.E.) | GUESSING | CHI-SQ | D.F. | PROB |
| 1 | -2.04 | ( 0.12) | 0.54 | ( 0.05) | 0.10 | 8.22 | 4.90 | 0.135 |
| 2 | -2.05 | ( 0.12) | 0.53 | ( 0.05) | 0.10 | 8.14 | 4.90 | 0.139 |
| 3 | -1.57 | ( 0.09) | 0.43 | ( 0.04) | 0.10 | 8.04 | 4.90 | 0.144 |
| 4 | -1.95 | ( 0.15) | 0.86 | ( 0.08) | 0.10 | 12.49 | 6.80 | 0.078 |
| 5 | -1.96 | ( 0.11) | 0.52 | ( 0.05) | 0.10 | 22.50 | 4.90 | 0.0 |
| 6 | -1.21 | ( 0.08) | 0.32 | ( 0.03) | 0.10 | 4.55 | 3.90 | 0.321 |
| 7 | -1.68 | ( 0.12) | 0.72 | ( 0.06) | 0.10 | 7.00 | 6.80 | 0.407 |
| 8 | -1.14 | ( 0.08) | 0.32 | ( 0.03) | 0.10 | 15.33 | 4.90 | 0.008 |
| 9 | -0.64 | ( 0.08) | 0.64 | ( 0.05) | 0.10 | 20.03 | 7.80 | 0.009 |
| 10 | -1.63 | ( 0.09) | 0.42 | ( 0.04) | 0.10 | 16.76 | 3.90 | 0.002 |
| 11 | -0.91 | ( 0.08) | 0.44 | ( 0.04) | 0.10 | 16.99 | 5.80 | 0.008 |
| 12 | -1.49 | ( 0.11) | 0.71 | ( 0.06) | 0.10 | 10.92 | 6.80 | 0.131 |
| 13 | -1.31 | ( 0.09) | 0.50 | ( 0.04) | 0.10 | 14.06 | 5.80 | 0.026 |
| 14 | -0.84 | ( 0.08) | 0.48 | ( 0.04) | 0.10 | 8.94 | 6.80 | 0.240 |
| 15 | -0.86 | ( 0.07) | 0.35 | ( 0.03) | 0.10 | 1.20 | 5.80 | 0.973 |
| 16 | -1.01 | ( 0.09) | 0.62 | ( 0.05) | 0.10 | 18.64 | 6.80 | 0.008 |
| 17 | -0.69 | ( 0.07) | 0.30 | ( 0.02) | 0.10 | 5.62 | 5.80 | 0.446 |
| 18 | -0.90 | ( 0.08) | 0.48 | ( 0.04) | 0.10 | 6.12 | 5.80 | 0.390 |
| 19 | -0.54 | ( 0.08) | 0.63 | ( 0.05) | 0.10 | 9.24 | 7.80 | 0.302 |
| 20 | -0.93 | ( 0.08) | 0.36 | ( 0.03) | 0.10 | 8.69 | 4.90 | 0.113 |
| 21 | -0.90 | ( 0.08) | 0.57 | ( 0.05) | 0.10 | 21.01 | 6.80 | 0.003 |
| 22 | -0.42 | ( 0.07) | 0.35 | ( 0.03) | 0.10 | 34.27 | 5.80 | 0.0 |
| 23 | -0.13 | ( 0.06) | 0.36 | ( 0.03) | 0.10 | 28.04 | 6.80 | 0.0 |
| 24 | -0.05 | ( 0.08) | 0.83 | ( 0.07) | 0.10 | 26.19 | 7.80 | 0.001 |
| 25 | -0.19 | ( 0.08) | 0.67 | ( 0.05) | 0.10 | 13.50 | 7.80 | 0.087 |
| 26 | -0.26 | ( 0.07) | 0.35 | ( 0.03) | 0.10 | 11.23 | 6.80 | 0.119 |
| 27 | -0.16 | ( 0.08) | 0.72 | ( 0.06) | 0.10 | 14.06 | 7.80 | 0.073 |
| 28 | -0.42 | ( 0.08) | 0.59 | ( 0.05) | 0.10 | 9.55 | 7.80 | 0.278 |
| 29 | 0.39 | ( 0.08) | 0.85 | ( 0.07) | 0.10 | 10.77 | 7.80 | 0.199 |
| 30 | 0.26 | ( 0.07) | 0.62 | ( 0.05) | 0.10 | 13.80 | 7.80 | 0.079 |
| 31 | -0.93 | ( 0.08) | 0.35 | ( 0.03) | 0.10 | 4.13 | 4.90 | 0.511 |
| 32 | 0.38 | ( 0.08) | 0.90 | ( 0.08) | 0.10 | 20.13 | 7.80 | 0.009 |
| 33 | 0.51 | ( 0.06) | 0.39 | ( 0.03) | 0.10 | 46.62 | 6.80 | 0.0 |
| 34 | 0.04 | ( 0.07) | 0.53 | ( 0.04) | 0.10 | 17.68 | 7.80 | 0.021 |
| 35 | 0.14 | ( 0.06) | 0.28 | ( 0.02) | 0.10 | 5.42 | 5.80 | 0.470 |

NOTES: GUESSING PARAMETERS FIXED DURING ESTIMATION.
DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR IN CHI-SQUARES.

TABLE D-4
ITEM PARAMETER ESTIMATES FOR PARAGRAPH COMPREHENSION
TOTAL CHI-SQUARE = 173.92 WITH 79.00 DF

| | PARAMETER ESTIMATES | | | | | ITEM FIT STATISTICS | | |
|---|---|---|---|---|---|---|---|---|
| ITEM | THRESHOLD | (S.E.) | DISPERSION | (S.E.) | GUESSING | CHI-SQ | D.F. | PROB |
| 1 | -1.12 | ( 0.13) | 0.63 | ( 0.06) | 0.10 | 4.95 | 5.60 | 0.500 |
| 2 | -1.97 | ( 0.15) | 0.72 | ( 0.07) | 0.10 | 15.61 | 4.70 | 0.006 |
| 3 | -1.11 | ( 0.13) | 0.25 | ( 0.03) | 0.10 | 10.85 | 2.80 | 0.010 |
| 4 | -1.00 | ( 0.14) | 0.89 | ( 0.08) | 0.10 | 12.33 | 5.60 | 0.044 |
| 5 | -0.68 | ( 0.12) | 0.45 | ( 0.04) | 0.10 | 18.48 | 5.60 | 0.004 |
| 6 | -0.90 | ( 0.13) | 0.78 | ( 0.07) | 0.10 | 5.69 | 5.60 | 0.411 |
| 7 | -0.99 | ( 0.13) | 0.79 | ( 0.07) | 0.10 | 9.99 | 5.60 | 0.103 |
| 8 | -1.45 | ( 0.15) | 0.99 | ( 0.09) | 0.10 | 3.69 | 5.60 | 0.672 |
| 9 | -0.87 | ( 0.13) | 0.48 | ( 0.04) | 0.10 | 21.23 | 4.70 | 0.001 |
| 10 | 0.01 | ( 0.12) | 0.48 | ( 0.05) | 0.10 | 8.25 | 5.60 | 0.188 |
| 11 | -1.27 | ( 0.17) | 1.24 | ( 0.12) | 0.10 | 30.23 | 6.50 | 0.0 |
| 12 | -0.20 | ( 0.12) | 0.64 | ( 0.06) | 0.10 | 1.98 | 5.60 | 0.898 |
| 13 | -0.94 | ( 0.13) | 0.51 | ( 0.05) | 0.10 | 6.92 | 4.70 | 0.196 |
| 14 | -0.45 | ( 0.12) | 0.68 | ( 0.06) | 0.10 | 12.35 | 5.60 | 0.044 |
| 15 | 1.50 | ( 0.36) | 3.60 | ( 0.70) | 0.10 | 11.38 | 6.50 | 0.100 |

NOTES: GUESSING PARAMETERS FIXED DURING ESTIMATION.
DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR IN CHI-SQUARES.

TABLE D-5
ITEM PARAMETER ESTIMATES FOR AUTO & SHOP INFORMATION
TOTAL CHI-SQUARE = 360.56 WITH 172.00 DF

| | PARAMETER ESTIMATES | | | | | ITEM FIT STATISTICS | | |
|---|---|---|---|---|---|---|---|---|
| ITEM | THRESHOLD | (S.E.) | DISPERSION | (S.E.) | GUESSING | CHI-SQ | D.F. | PROB |
| 1 | -1.81 | ( 0.26) | 1.67 | ( 0.21) | 0.10 | 21.72 | 7.70 | 0.004 |
| 2 | -1.45 | ( 0.18) | 0.83 | ( 0.08) | 0.10 | 6.89 | 6.70 | 0.410 |
| 3 | -0.27 | ( 0.12) | 0.46 | ( 0.04) | 0.10 | 6.35 | 5.80 | 0.358 |
| 4 | -1.76 | ( 0.19) | 0.95 | ( 0.09) | 0.10 | 9.70 | 6.70 | 0.186 |
| 5 | -0.34 | ( 0.12) | 0.32 | ( 0.03) | 0.10 | 5.71 | 4.80 | 0.312 |
| 6 | 0.00 | ( 0.11) | 0.22 | ( 0.02) | 0.10 | 6.98 | 4.80 | 0.203 |
| 7 | 0.17 | ( 0.11) | 0.53 | ( 0.05) | 0.10 | 19.38 | 6.70 | 0.006 |
| 8 | -0.70 | ( 0.18) | 1.58 | ( 0.19) | 0.10 | 9.04 | 7.70 | 0.310 |
| 9 | -1.12 | ( 0.22) | 1.90 | ( 0.25) | 0.10 | 20.52 | 7.70 | 0.007 |
| 10 | -0.25 | ( 0.13) | 0.64 | ( 0.06) | 0.10 | 16.86 | 6.70 | 0.016 |
| 11 | -0.09 | ( 0.13) | 1.03 | ( 0.11) | 0.10 | 14.20 | 7.70 | 0.066 |
| 12 | -0.37 | ( 0.15) | 1.30 | ( 0.14) | 0.10 | 15.94 | 7.70 | 0.037 |
| 13 | -0.24 | ( 0.12) | 0.58 | ( 0.06) | 0.10 | 27.14 | 6.70 | 0.0 |
| 14 | -0.27 | ( 0.12) | 0.51 | ( 0.05) | 0.10 | 23.08 | 6.70 | 0.001 |
| 15 | 0.14 | ( 0.11) | 0.54 | ( 0.05) | 0.10 | 10.89 | 6.70 | 0.128 |
| 16 | 0.10 | ( 0.11) | 0.45 | ( 0.04) | 0.10 | 2.73 | 6.70 | 0.893 |
| 17 | 0.44 | ( 0.10) | 0.36 | ( 0.04) | 0.10 | 5.41 | 6.70 | 0.578 |
| 18 | 0.43 | ( 0.10) | 0.37 | ( 0.04) | 0.10 | 14.84 | 6.70 | 0.033 |
| 19 | -0.57 | ( 0.15) | 0.96 | ( 0.10) | 0.10 | 26.97 | 7.70 | 0.001 |
| 20 | 0.17 | ( 0.11) | 0.71 | ( 0.07) | 0.10 | 15.56 | 7.70 | 0.042 |
| 21 | 0.73 | ( 0.11) | 0.85 | ( 0.09) | 0.10 | 25.49 | 7.70 | 0.001 |
| 22 | 0.46 | ( 0.10) | 0.36 | ( 0.04) | 0.10 | 15.69 | 6.70 | 0.024 |
| 23 | 0.67 | ( 0.09) | 0.21 | ( 0.03) | 0.10 | 14.89 | 6.70 | 0.032 |
| 24 | 1.13 | ( 0.08) | 0.52 | ( 0.05) | 0.10 | 17.10 | 7.70 | 0.025 |
| 25 | 1.01 | ( 0.13) | 1.20 | ( 0.14) | 0.10 | 7.50 | 7.70 | 0.451 |

NOTES: GUESSING PARAMETERS FIXED DURING ESTIMATION.
DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR IN CHI-SQUARES.

TABLE D-6
ITEM PARAMETER ESTIMATES FOR MATHEMATICS KNOWLEDGE
TOTAL CHI-SQUARE = 418.00 WITH 173.00 DF

| | PARAMETER ESTIMATES | | | | | ITEM FIT STATISTICS | | |
|---|---|---|---|---|---|---|---|---|
| ITEM | THRESHOLD | (S.E.) | DISPERSION | (S.E.) | GUESSING | CHI-SQ | D.F. | PROB |
| 1 | -1.27 | ( 0.17) | 0.37 | ( 0.04) | 0.10 | 9.86 | 3.80 | 0.038 |
| 2 | -1.16 | ( 0.16) | 0.73 | ( 0.07) | 0.10 | 6.62 | 6.70 | 0.438 |
| 3 | -0.28 | ( 0.12) | 0.51 | ( 0.05) | 0.10 | 29.29 | 6.70 | 0.0 |
| 4 | -0.63 | ( 0.14) | 0.48 | ( 0.05) | 0.10 | 8.19 | 5.80 | 0.204 |
| 5 | -1.13 | ( 0.17) | 0.99 | ( 0.10) | 0.10 | 11.36 | 7.70 | 0.162 |
| 6 | -0.16 | ( 0.12) | 0.42 | ( 0.04) | 0.10 | 9.10 | 5.80 | 0.151 |
| 7 | -0.40 | ( 0.13) | 0.42 | ( 0.04) | 0.10 | 16.93 | 5.80 | 0.008 |
| 8 | -0.17 | ( 0.12) | 0.73 | ( 0.07) | 0.10 | 25.32 | 7.70 | 0.001 |
| 9 | -0.50 | ( 0.13) | 0.55 | ( 0.05) | 0.10 | 4.83 | 6.70 | 0.650 |
| 10 | -0.29 | ( 0.13) | 0.67 | ( 0.06) | 0.10 | 18.98 | 7.70 | 0.012 |
| 11 | -0.01 | ( 0.12) | 0.97 | ( 0.09) | 0.10 | 13.32 | 7.70 | 0.088 |
| 12 | 0.21 | ( 0.10) | 0.43 | ( 0.04) | 0.10 | 12.81 | 6.70 | 0.067 |
| 13 | -0.01 | ( 0.11) | 0.25 | ( 0.02) | 0.10 | 18.83 | 5.80 | 0.004 |
| 14 | 0.32 | ( 0.10) | 0.35 | ( 0.04) | 0.10 | 5.49 | 6.70 | 0.569 |
| 15 | 0.87 | ( 0.12) | 1.10 | ( 0.12) | 0.10 | 14.15 | 7.70 | 0.068 |
| 16 | 0.02 | ( 0.11) | 0.50 | ( 0.05) | 0.10 | 29.56 | 7.70 | 0.0 |
| 17 | 0.37 | ( 0.09) | 0.39 | ( 0.04) | 0.10 | 25.25 | 7.70 | 0.001 |
| 18 | 0.17 | ( 0.10) | 0.31 | ( 0.03) | 0.10 | 22.86 | 5.80 | 0.001 |
| 19 | -0.03 | ( 0.12) | 0.75 | ( 0.07) | 0.10 | 26.84 | 7.70 | 0.001 |
| 20 | 0.38 | ( 0.10) | 0.76 | ( 0.07) | 0.10 | 11.89 | 7.70 | 0.138 |
| 21 | 0.70 | ( 0.10) | 0.77 | ( 0.07) | 0.10 | 39.30 | 7.70 | 0.0 |
| 22 | 0.82 | ( 0.08) | 0.36 | ( 0.04) | 0.10 | 18.20 | 7.70 | 0.017 |
| 23 | 0.97 | ( 0.07) | 0.35 | ( 0.04) | 0.10 | 16.05 | 7.70 | 0.035 |
| 24 | 1.20 | ( 0.08) | 0.61 | ( 0.06) | 0.10 | 9.03 | 7.70 | 0.310 |
| 25 | 0.96 | ( 0.07) | 0.40 | ( 0.04) | 0.10 | 13.93 | 7.70 | 0.072 |

NOTES: GUESSING PARAMETERS FIXED DURING ESTIMATION.
DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR IN CHI-SQUARES.

TABLE D-7
ITEM PARAMETER ESTIMATES FOR MECHANICAL COMPREHENSION
TOTAL CHI-SQUARE = 360.78 WITH 178.00 DF

| | PARAMETER ESTIMATES | | | | | ITEM FIT STATISTICS | | |
|---|---|---|---|---|---|---|---|---|
| ITEM | THRESHOLD | (S.E.) | DISPERSION | (S.E.) | GUESSING | CHI-SQ | D.F. | PROB |
| 1 | -2.18 | ( 0.20) | 0.96 | ( 0.11) | 0.10 | 3.06 | 5.80 | 0.777 |
| 2 | -0.97 | ( 0.10) | 0.49 | ( 0.04) | 0.10 | 13.13 | 4.80 | 0.019 |
| 3 | -1.24 | ( 0.12) | 0.78 | ( 0.07) | 0.10 | 5.58 | 5.80 | 0.442 |
| 4 | 0.67 | ( 0.09) | 0.84 | ( 0.07) | 0.10 | 3.75 | 7.70 | 0.859 |
| 5 | -0.18 | ( 0.09) | 0.69 | ( 0.06) | 0.10 | 16.26 | 7.70 | 0.033 |
| 6 | -0.38 | ( 0.13) | 1.49 | ( 0.16) | 0.10 | 16.22 | 7.70 | 0.033 |
| 7 | -0.20 | ( 0.10) | 0.76 | ( 0.06) | 0.10 | 10.26 | 7.70 | 0.223 |
| 8 | -0.43 | ( 0.09) | 0.45 | ( 0.04) | 0.10 | 27.11 | 5.80 | 0.0 |
| 9 | -0.73 | ( 0.12) | 1.03 | ( 0.10) | 0.10 | 20.51 | 7.70 | 0.007 |
| 10 | -0.25 | ( 0.09) | 0.39 | ( 0.03) | 0.10 | 18.85 | 5.80 | 0.004 |
| 11 | -0.81 | ( 0.13) | 1.29 | ( 0.13) | 0.10 | 6.83 | 7.70 | 0.521 |
| 12 | -0.18 | ( 0.09) | 0.56 | ( 0.05) | 0.10 | 9.44 | 6.70 | 0.201 |
| 13 | -0.44 | ( 0.09) | 0.57 | ( 0.05) | 0.10 | 9.47 | 6.70 | 0.199 |
| 14 | -0.18 | ( 0.09) | 0.56 | ( 0.05) | 0.10 | 6.37 | 6.70 | 0.466 |
| 15 | -0.13 | ( 0.09) | 0.65 | ( 0.05) | 0.10 | 17.35 | 7.70 | 0.022 |
| 16 | -0.25 | ( 0.11) | 1.05 | ( 0.10) | 0.10 | 11.24 | 7.70 | 0.168 |
| 17 | 0.20 | ( 0.10) | 0.90 | ( 0.08) | 0.10 | 23.80 | 7.70 | 0.002 |
| 18 | 0.20 | ( 0.09) | 0.65 | ( 0.05) | 0.10 | 21.12 | 7.70 | 0.006 |
| 19 | 1.00 | ( 0.15) | 1.51 | ( 0.17) | 0.10 | 45.57 | 7.70 | 0.0 |
| 20 | 0.13 | ( 0.10) | 0.89 | ( 0.08) | 0.10 | 13.53 | 7.70 | 0.083 |
| 21 | 0.49 | ( 0.09) | 0.70 | ( 0.06) | 0.10 | 17.26 | 7.70 | 0.023 |
| 22 | 1.06 | ( 0.12) | 1.17 | ( 0.12) | 0.10 | 17.08 | 7.70 | 0.025 |
| 23 | 0.76 | ( 0.08) | 0.55 | ( 0.05) | 0.10 | 4.55 | 7.70 | 0.778 |
| 24 | 0.71 | ( 0.10) | 0.89 | ( 0.08) | 0.10 | 15.64 | 7.70 | 0.041 |
| 25 | 0.84 | ( 0.08) | 0.60 | ( 0.05) | 0.10 | 6.78 | 7.70 | 0.526 |

NOTES: GUESSING PARAMETERS FIXED DURING ESTIMATION.
DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR IN CHI-SQUARES.

TABLE D-8
ITEM PARAMETER ESTIMATES FOR ELECTRONICS INFORMATION
TOTAL CHI-SQUARE = 244.90 WITH 135.00 DF

| | PARAMETER ESTIMATES | | | | | ITEM FIT STATISTICS | | |
|---|---|---|---|---|---|---|---|---|
| ITEM | THRESHOLD | (S.E.) | DISPERSION | (S.E.) | GUESSING | CHI-SQ | D.F. | PROB |
| 1 | -1.24 | ( 0.16) | 0.68 | ( 0.06) | 0.10 | 4.37 | 5.70 | 0.590 |
| 2 | -0.69 | ( 0.14) | 0.50 | ( 0.05) | 0.10 | 13.79 | 5.70 | 0.027 |
| 3 | -0.57 | ( 0.14) | 0.40 | ( 0.04) | 0.10 | 2.90 | 5.70 | 0.794 |
| 4 | -0.59 | ( 0.14) | 0.41 | ( 0.04) | 0.10 | 8.29 | 5.70 | 0.193 |
| 5 | -1.03 | ( 0.16) | 0.68 | ( 0.06) | 0.10 | 2.65 | 5.70 | 0.826 |
| 6 | -0.45 | ( 0.13) | 0.40 | ( 0.04) | 0.10 | 3.25 | 5.70 | 0.745 |
| 7 | -0.55 | ( 0.14) | 0.58 | ( 0.05) | 0.10 | 11.52 | 6.60 | 0.101 |
| 8 | -0.51 | ( 0.14) | 0.69 | ( 0.06) | 0.10 | 11.24 | 6.60 | 0.110 |
| 9 | -0.83 | ( 0.16) | 0.96 | ( 0.09) | 0.10 | 15.84 | 7.60 | 0.037 |
| 10 | -0.08 | ( 0.12) | 0.63 | ( 0.06) | 0.10 | 16.06 | 6.60 | 0.020 |
| 11 | 0.56 | ( 0.13) | 1.14 | ( 0.12) | 0.10 | 17.97 | 7.60 | 0.017 |
| 12 | 0.24 | ( 0.11) | 0.53 | ( 0.05) | 0.10 | 7.53 | 6.60 | 0.340 |
| 13 | 0.48 | ( 0.11) | 0.66 | ( 0.06) | 0.10 | 12.48 | 7.60 | 0.112 |
| 14 | 0.31 | ( 0.15) | 1.71 | ( 0.21) | 0.10 | 28.19 | 7.60 | 0.0 |
| 15 | 0.37 | ( 0.11) | 0.50 | ( 0.05) | 0.10 | 9.15 | 6.60 | 0.214 |
| 16 | 0.53 | ( 0.11) | 0.81 | ( 0.08) | 0.10 | 18.51 | 7.60 | 0.014 |
| 17 | 3.78 | ( 0.80) | 2.46 | ( 0.56) | 0.10 | 12.65 | 7.60 | 0.106 |
| 18 | 0.55 | ( 0.11) | 0.80 | ( 0.08) | 0.10 | 17.94 | 7.60 | 0.017 |
| 19 | 0.44 | ( 0.11) | 0.68 | ( 0.06) | 0.10 | 9.45 | 7.60 | 0.271 |
| 20 | 1.10 | ( 0.09) | 0.58 | ( 0.06) | 0.10 | 21.13 | 7.60 | 0.005 |

NOTES: GUESSING PARAMETERS FIXED DURING ESTIMATION.
DESIGN EFFECT OF ABOUT 2.0 NOT ACCOUNTED FOR IN CHI-SQUARES.

# APPENDIX E

FIGURE E-1
ITEM THRESHOLDS FOR GENERAL SCIENCE

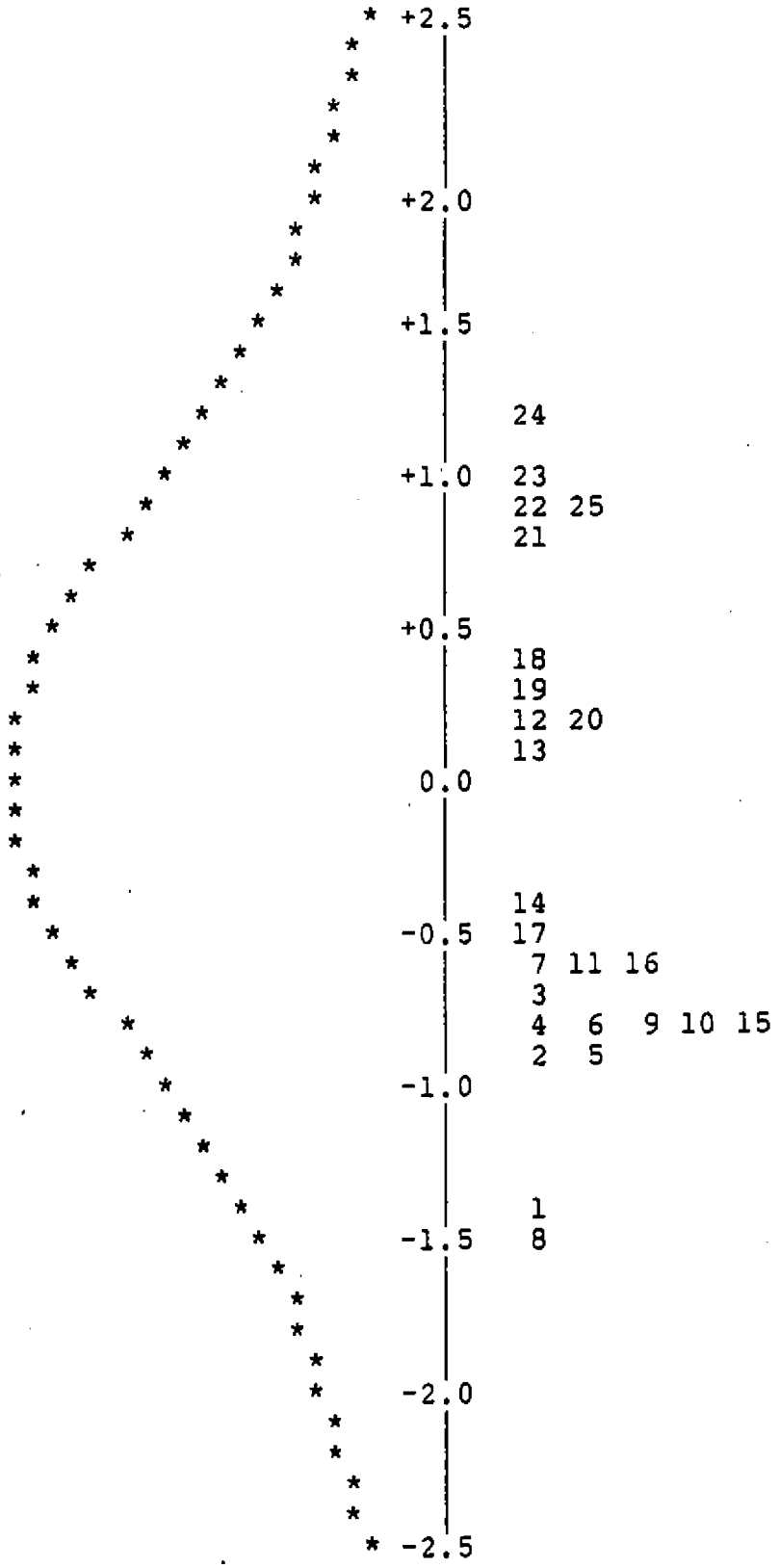


FIGURE E-2
ITEM THRESHOLDS FOR ARITHMETIC REASONING

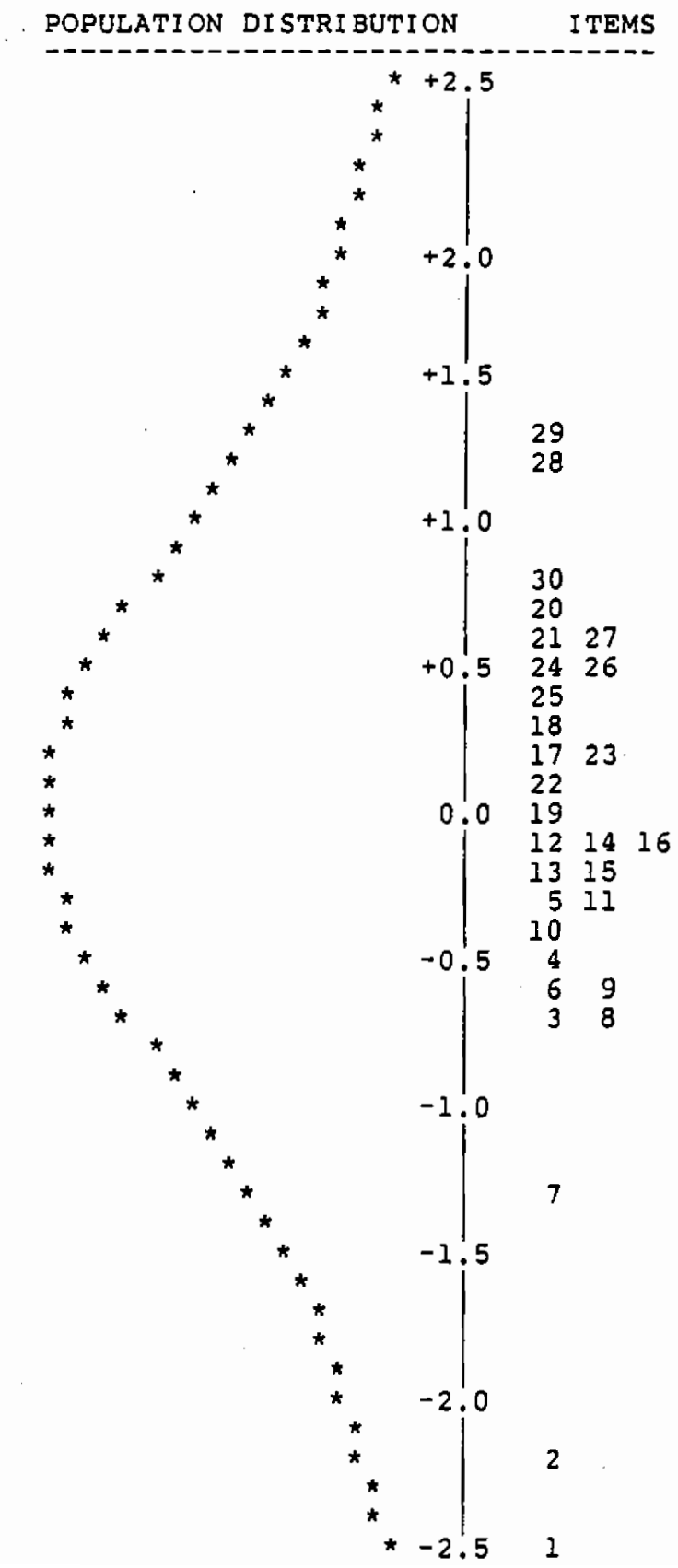


FIGURE E-3
ITEM THRESHOLDS FOR WORD KNOWLEDGE

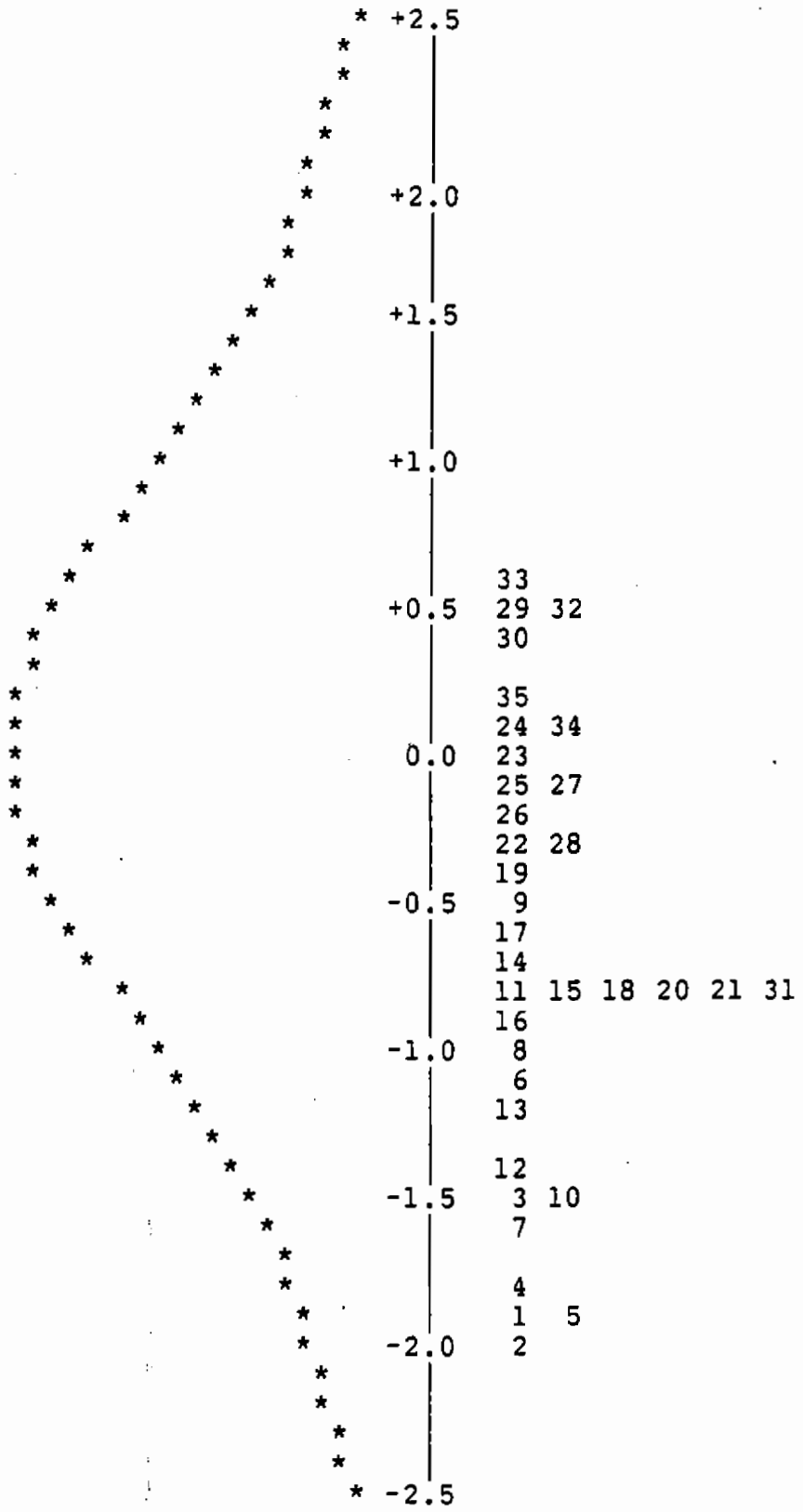


FIGURE E-4
ITEM THRESHOLDS FOR PARAGRAPH COMPREHENSION

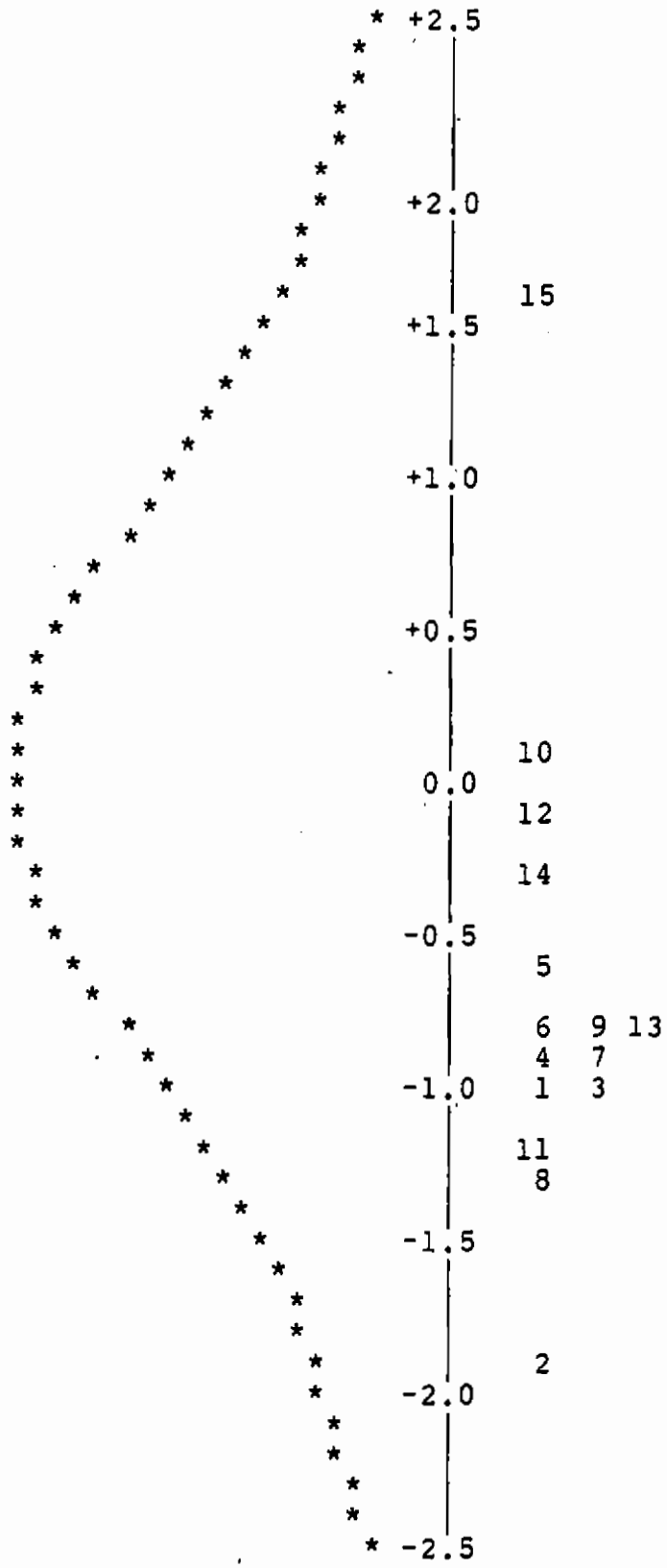


FIGURE E-5
ITEM THRESHOLDS FOR AUTO & SHOP INFORMATION

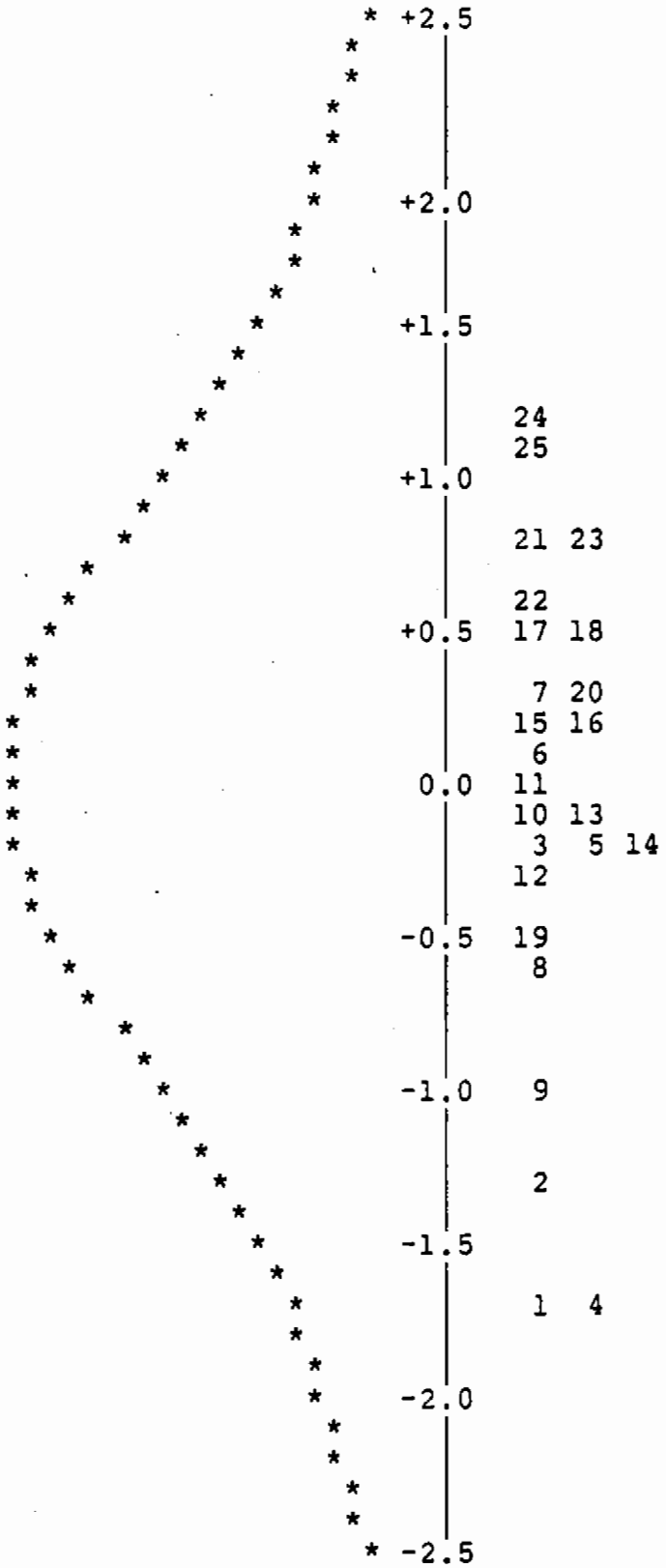


FIGURE E-6
ITEM THRESHOLDS FOR MATHEMATICS KNOWLEDGE

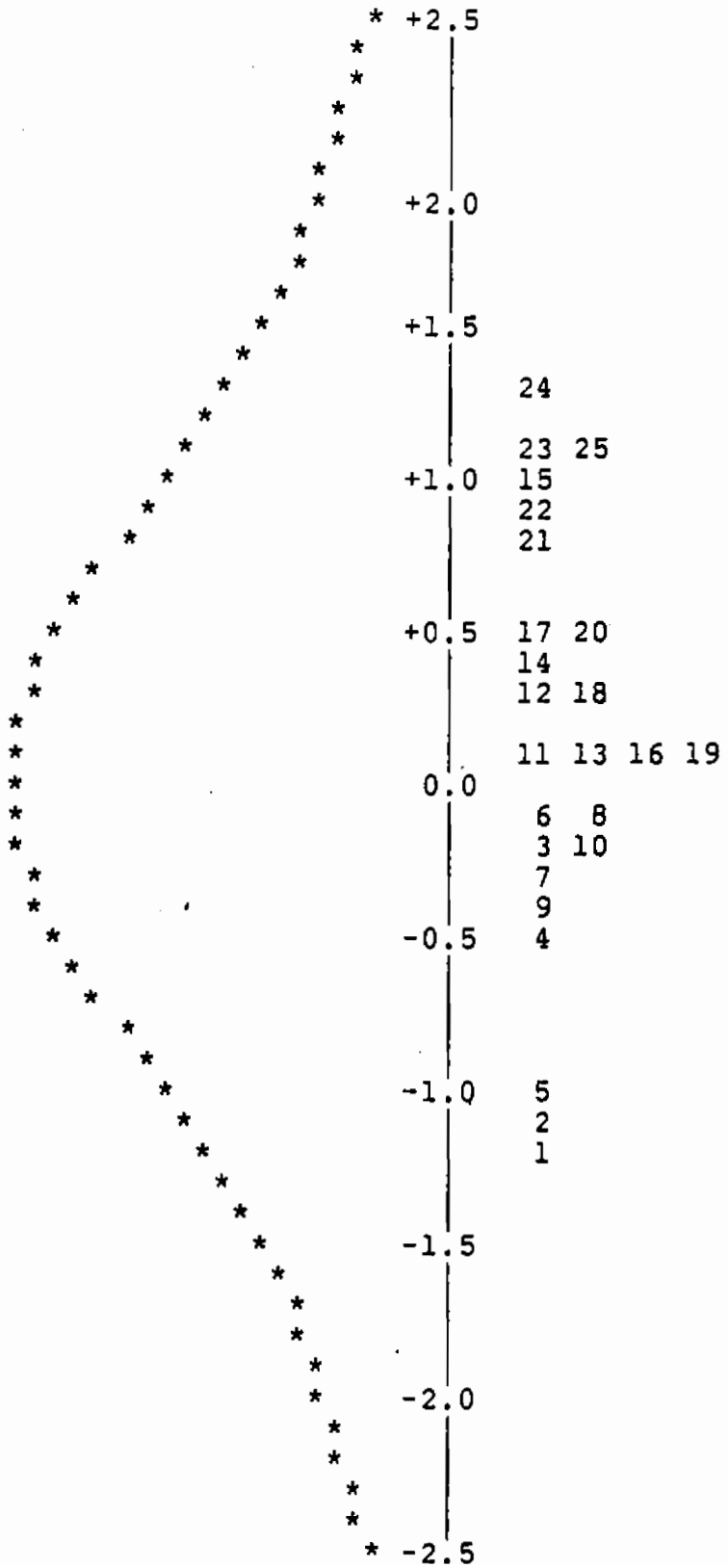


FIGURE E-7
ITEM THRESHOLDS FOR MECHANICAL COMPREHENSION

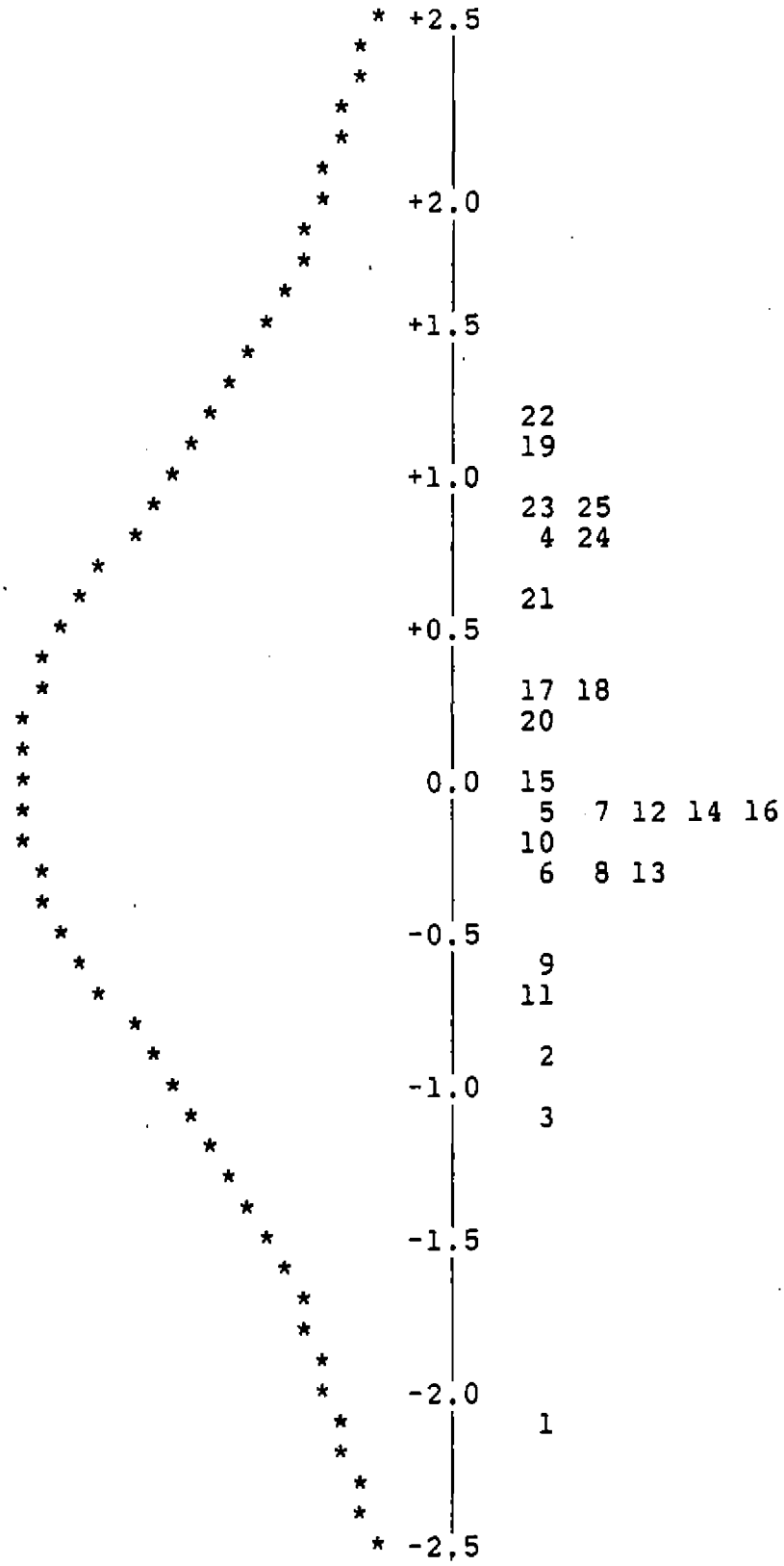
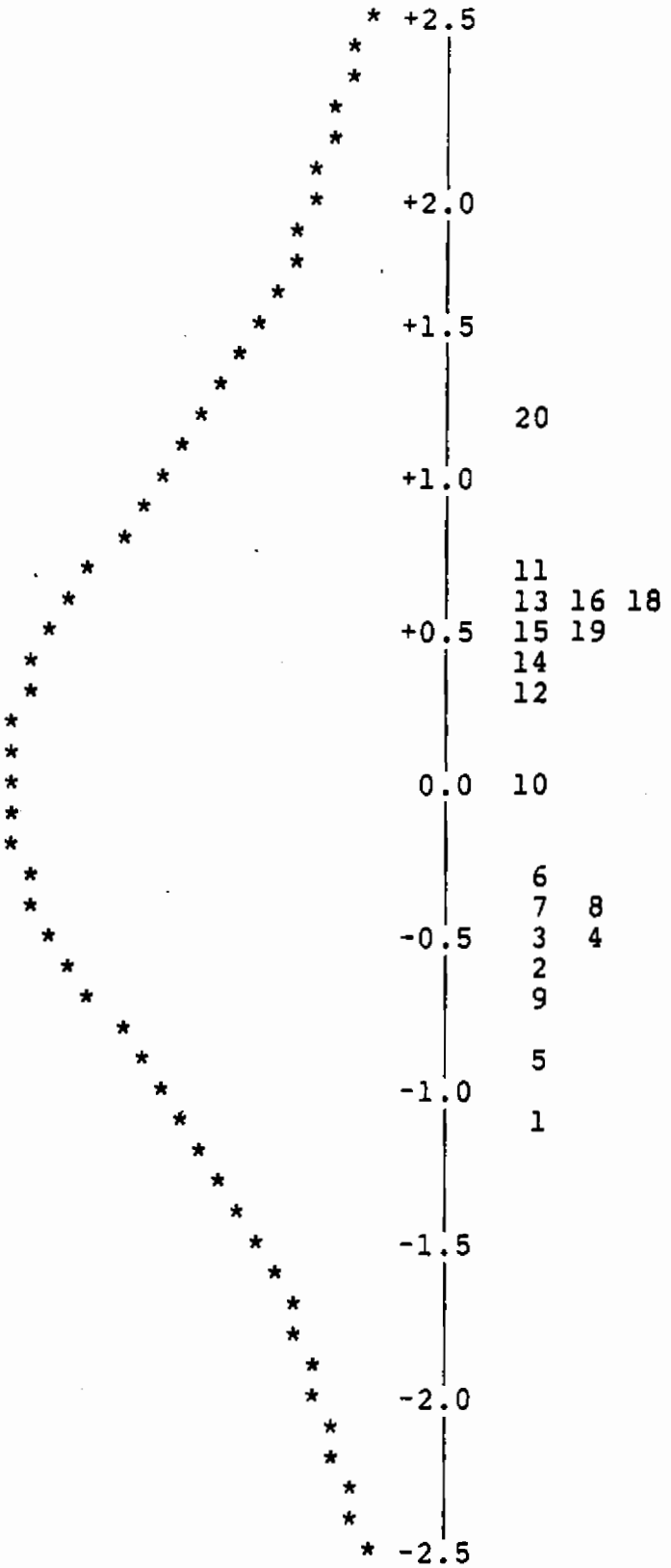


FIGURE E-8
ITEM THRESHOLDS FOR ELECTRONICS INFORMATION



# APPENDIX F

TABLE F-1
WITHIN-GROUP ITEM PARAMETER ESTIMATES FOR GENERAL SCIENCE

| | THRESHOLDS | | | | | | DISPERSIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | MALE | FEMALE | WHITE | POOR | BLACK | HISP | MALE | FEMALE | WHITE | POOR | BLACK | HISP |
| 1 | -1.75 | -2.12 | -1.99 | -2.21 | -1.95 | -1.60 | 0.88 | 0.80 | 0.69 | 0.91 | 0.81 | 0.96 |
| 2 | -1.50 | -1.10 | -1.56 | -1.56 | -0.94 | -1.16 | 0.94 | 0.84 | 0.94 | 1.06 | 0.82 | 0.74 |
| 3 | -1.52 | -0.83 | -1.44 | -1.47 | -0.95 | -0.84 | 0.81 | 0.80 | 0.92 | 0.90 | 0.73 | 0.67 |
| 4 | -1.13 | -0.69 | -0.97 | -1.01 | -0.59 | -1.07 | 1.25 | 1.20 | 1.46 | 1.26 | 1.18 | 1.00 |
| 5 | -1.16 | -1.14 | -1.64 | -1.27 | -0.95 | -0.75 | 0.90 | 1.41 | 1.64 | 1.14 | 0.88 | 0.95 |
| 6 | -0.74 | -1.79 | -1.49 | -1.13 | -1.46 | -0.98 | 0.80 | 0.72 | 0.78 | 0.63 | 0.71 | 0.92 |
| 7 | -0.41 | -0.73 | -0.49 | -0.44 | -0.77 | -0.59 | 0.91 | 0.84 | 0.83 | 0.87 | 0.95 | 0.84 |
| 8 | -0.97 | -1.88 | -1.20 | -1.46 | -1.67 | -1.39 | 1.07 | 0.91 | 0.89 | 0.85 | 1.07 | 1.16 |
| 9 | -1.08 | -1.13 | -1.09 | -1.05 | -1.14 | -1.15 | 0.94 | 0.93 | 1.08 | 0.96 | 0.88 | 0.82 |
| 10 | -1.12 | -0.74 | -1.15 | -1.01 | -0.97 | -0.59 | 1.09 | 1.24 | 1.19 | 1.29 | 1.04 | 1.14 |
| 11 | -0.44 | -0.71 | -0.41 | -0.51 | -0.76 | -0.63 | 0.91 | 0.84 | 0.89 | 0.86 | 0.82 | 0.92 |
| 12 | -0.04 | 2.00 | 0.91 | 1.45 | 1.15 | 0.41 | 1.58 | 2.44 | 1.34 | 2.20 | 2.59 | 1.91 |
| 13 | 0.05 | 1.23 | 0.45 | 0.46 | 1.31 | 0.35 | 0.94 | 1.16 | 0.83 | 0.88 | 1.43 | 1.06 |
| 14 | -0.36 | 0.09 | -0.12 | -0.09 | -0.07 | -0.27 | 1.03 | 1.38 | 1.09 | 1.16 | 1.31 | 1.24 |
| 15 | -0.59 | -1.11 | -0.69 | -0.65 | -1.12 | -0.94 | 1.19 | 1.31 | 1.31 | 1.18 | 1.21 | 1.30 |
| 16 | 0.18 | -1.13 | -0.34 | -0.64 | -0.47 | -0.44 | 1.05 | 0.93 | 1.06 | 1.00 | 0.90 | 1.02 |
| 17 | 0.40 | -0.78 | -0.02 | -0.09 | -0.41 | -0.23 | 1.12 | 0.93 | 0.86 | 0.98 | 1.32 | 0.94 |
| 18 | 1.07 | 0.84 | 0.94 | 1.02 | 0.88 | 0.97 | 0.69 | 0.50 | 0.59 | 0.55 | 0.67 | 0.56 |
| 19 | 1.00 | 0.56 | 1.01 | 0.84 | 0.75 | 0.53 | 1.48 | 1.25 | 1.05 | 1.15 | 1.62 | 1.63 |
| 20 | 0.98 | 0.41 | 0.74 | 0.66 | 0.73 | 0.65 | 1.14 | 1.01 | 1.24 | 1.27 | 0.93 | 0.87 |
| 21 | 1.45 | 1.73 | 1.56 | 1.59 | 1.46 | 1.75 | 0.99 | 1.31 | 0.93 | 1.13 | 1.42 | 1.14 |
| 22 | 1.95 | 1.39 | 1.89 | 1.70 | 1.20 | 1.90 | 1.08 | 1.11 | 0.98 | 1.11 | 1.06 | 1.23 |
| 23 | 1.39 | 4.05 | 2.65 | 2.40 | 3.24 | 2.60 | 1.22 | 2.02 | 1.66 | 1.27 | 1.53 | 2.02 |
| 24 | 2.47 | 2.29 | 2.68 | 2.52 | 2.13 | 2.17 | 0.61 | 0.43 | 0.70 | 0.55 | 0.41 | 0.44 |
| 25 | 1.89 | 1.31 | 1.77 | 1.92 | 1.39 | 1.31 | 1.15 | 1.17 | 1.10 | 1.07 | 1.27 | 1.19 |

TABLE F-2
WITHIN-GROUP ITEM PARAMETER ESTIMATES FOR ARITHMETIC REASONING

| | THRESHOLDS | | | | | | DISPERSIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | MALE | FEMALE | WHITE | POOR | BLACK | HISP | MALE | FEMALE | WHITE | POOR | BLACK | HISP |
| 1 | -4.03 | -4.53 | -5.70 | -4.44 | -3.43 | -3.56 | 1.57 | 1.71 | 2.42 | 1.84 | 1.05 | 1.24 |
| 2 | -3.19 | -3.56 | -3.65 | -3.41 | -3.41 | -3.03 | 1.09 | 1.17 | 1.32 | 1.12 | 1.07 | 1.01 |
| 3 | -1.35 | -1.09 | -1.23 | -1.26 | -1.47 | -0.90 | 0.88 | 0.76 | 0.80 | 0.85 | 0.71 | 0.91 |
| 4 | -0.80 | -0.81 | -0.74 | -0.85 | -0.77 | -0.87 | 0.59 | 0.61 | 0.58 | 0.69 | 0.48 | 0 65 |
| 5 | -0.56 | -0.67 | -0.49 | -0.64 | -0.60 | -0.74 | 1.06 | 0.89 | 1.06 | 1.10 | 0.78 | 0.97 |
| 6 | -0.69 | -1.23 | -0.73 | -0.79 | -1.19 | -1.13 | 1.07 | 0.76 | 1.19 | 0.82 | 0.70 | 0.96 |
| 7 | -1.61 | -1.86 | -2.09 | -1.64 | -1.73 | -1.50 | 1.27 | 1.63 | 1.61 | 1.29 | 1.35 | 1.55 |
| 8 | -1.35 | -1.07 | -1.44 | -1.09 | -1.18 | -1.13 | 1.00 | 1.05 | 1.19 | 0.88 | 0.90 | 1.13 |
| 9 | -0.63 | -0.73 | -0.58 | -0.49 | -0.75 | -0.90 | 0.88 | 0.77 | 0.86 | 0.94 | 0.71 | 0.78 |
| 10 | -0.62 | -0.49 | -0.44 | -0.52 | -0.70 | -0.56 | 0.61 | 0.63 | 0.64 | 0.60 | 0.62 | 0.64 |
| 11 | -0.20 | -0.55 | -0.30 | -0.29 | -0.40 | -0.52 | 0.54 | 0.56 | 0.59 | 0.57 | 0.52 | 0.52 |
| 12 | -0.33 | -0.19 | -0.18 | -0.15 | -0.33 | -0.38 | 1.00 | 0.93 | 1.04 | 0.98 | 0.94 | 0.89 |
| 13 | -0.12 | -0.25 | 0.03 | -0.13 | -0.26 | -0.40 | 0.96 | 1.29 | 1.04 | 1.01 | 1.45 | 0.99 |
| 14 | 0.12 | -0.09 | 0.13 | 0.07 | -0.07 | -0.08 | 0.84 | 0.75 | 0.75 | 0.84 | 0.71 | 0.87 |
| 15 | -0.37 | -0.45 | -0.32 | -0.32 | -0.49 | -0.49 | 0.99 | 0.80 | 1.00 | 0.92 | 0.82 | 0.84 |
| 16 | -0.05 | -0.32 | -0.04 | -0.13 | -0.34 | -0.22 | 1.19 | 1.38 | 1.17 | 1.29 | 1.36 | 1.34 |
| 17 | 0.62 | 0.31 | 0.61 | 0.45 | 0.49 | 0.31 | 1.79 | 1.30 | 1.41 | 1.59 | 1.76 | 1.41 |
| 18 | 0.74 | 0.84 | 1.06 | 0.77 | 0.62 | 0.72 | 1.18 | 1.29 | 0.93 | 1.05 | 1.55 | 1.41 |
| 19 | 0.16 | 0.20 | 0.27 | 0.31 | 0.14 | -0.01 | 1.24 | 1.19 | 1.29 | 1.41 | 1.04 | 1.12 |
| 20 | 1.29 | 2.00 | 1.61 | 1.67 | 1.49 | 1.79 | 0.87 | 0.92 | 0.82 | 0.96 | 0.87 | 0.94 |
| 21 | 1.18 | 1.59 | 1.47 | 1.10 | 1.75 | 1.23 | 1.26 | 1.39 | 0.86 | 1.07 | 1.95 | 1.41 |
| 22 | 0.48 | 0.65 | 0.43 | 0.61 | 0.67 | 0.54 | 1.05 | 1.04 | 1.09 | 0.90 | 1.21 | 0.98 |
| 23 | 0.64 | 0.69 | 0.98 | 0.80 | 0.45 | 0.44 | 1.12 | 1.54 | 1.23 | 1.19 | 1.54 | 1.37 |
| 24 | 0.92 | 1.17 | 1.25 | 0.96 | 1.01 | 0.96 | 0.88 | 1.16 | 0.70 | 0.90 | 1.47 | 1.01 |
| 25 | 1.09 | 1.24 | 1.17 | 1.06 | 1.25 | 1.17 | 0.92 | 0.88 | 0.84 | 0.90 | 1.02 | 0.84 |
| 26 | 1.14 | 0.90 | 1.23 | 1.02 | 0.81 | 1.03 | 1.17 | 1.08 | 1.17 | 1.04 | 1.06 | 1.22 |
| 27 | 1.22 | 1.68 | 1.46 | 1.32 | 1.45 | 1.56 | 0.89 | 0.79 | 0.97 | 1.00 | 0.61 | 0.78 |
| 28 | 2.25 | 2.66 | 2.15 | 2.25 | 3.05 | 2.37 | 1.11 | 1.02 | 0.83 | 1.12 | 1.39 | 0.92 |
| 29 | 2.53 | 2.44 | 2.56 | 2.47 | 2.41 | 2.49 | 1.02 | 1.23 | 1.07 | 1.02 | 1.23 | 1.19 |
| 30 | 1.51 | 1.54 | 1.50 | 1.28 | 1.53 | 1.80 | 1.30 | 1.23 | 1.10 | 1.32 | 1.31 | 1.33 |

TABLE F-3
WITHIN-GROUP ITEM PARAMETER ESTIMATES FOR WORD KNOWLEDGE

| | THRESHOLDS | | | | | | DISPERSIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | MALE | FEMALE | WHITE | POOR | BLACK | HISP | MALE | FEMALE | WHITE | POOR | BLACK | HISP |
| 1 | -2.25 | -2.52 | -2.56 | -2.24 | -2.49 | -2.25 | 0.88 | 0.93 | 0.95 | 0.99 | 0.84 | 0.85 |
| 2 | -2.43 | -3.10 | -4.31 | -2.31 | -2.16 | -2.28 | 0.97 | 1.12 | 1.67 | 0.96 | 0.64 | 0.91 |
| 3 | -1.37 | -1.71 | -1.26 | -1.59 | -1.81 | -1.51 | 0.86 | 0.71 | 0.74 | 0.85 | 0.74 | 0.81 |
| 4 | -1.69 | -2.90 | -2.52 | -2.62 | -1.99 | -2.04 | 0.99 | 1.85 | 1.84 | 1.67 | 0.91 | 1.27 |
| 5 | -1.75 | -2.09 | -2.37 | -1.62 | -1.86 | -1.81 | 0.99 | 1.03 | 1.37 | 0.90 | 0.80 | 0.96 |
| 6 | -1.25 | -0.36 | -1.08 | -1.13 | -0.46 | -0.54 | 0.75 | 0.71 | 0.60 | 0.79 | 0.79 | 0.73 |
| 7 | -1.74 | -1.66 | -0.95 | -1.77 | -2.32 | -1.76 | 1.21 | 1.27 | 0.97 | 1.43 | 1.53 | 1.04 |
| 8 | -0.87 | -0.70 | -0.60 | -0.91 | -0.90 | -0.73 | 0.90 | 0.74 | 0.65 | 0.85 | 0.89 | 0.88 |
| 9 | -0.14 | 1.22 | 0.23 | 0.55 | 0.44 | 0.95 | 1.00 | 1.59 | 1.65 | 1.16 | 1.16 | 1.22 |
| 10 | -1.48 | -1.35 | -1.25 | -1.48 | -1.26 | -1.65 | 0.81 | 0.67 | 0.68 | 0.78 | 0.77 | 0.73 |
| 11 | -0.78 | -0.19 | -0.35 | -0.51 | -0.34 | -0.75 | 0.91 | 0.79 | 0.81 | 0.87 | 0.86 | 0.85 |
| 12 | -0.92 | -1.65 | -2.32 | -1.20 | -0.88 | -0.75 | 1.15 | 1.44 | 1.84 | 1.17 | 1.21 | 0.97 |
| 13 | -1.02 | -0.65 | -0.95 | -0.82 | -0.72 | -0.87 | 0.85 | 0.67 | 0.85 | 0.84 | 0.77 | 0.59 |
| 14 | -0.19 | 0.34 | -0.29 | -0.46 | 0.49 | 0.56 | 1.18 | 1.33 | 1.02 | 1.20 | 1.27 | 1.54 |
| 15 | -0.25 | 0.41 | 0.02 | 0.20 | 0.09 | 0.01 | 0.77 | 0.68 | 0.72 | 0.76 | 0.63 | 0.78 |
| 16 | -0.52 | -0.74 | -0.40 | -0.45 | -0.92 | -0.75 | 1.12 | 1.03 | 1.10 | 0.92 | 1.10 | 1.17 |
| 17 | 0.26 | 0.43 | -0.08 | 0.02 | 0.46 | 0.97 | 0.73 | 0.72 | 0.73 | 0.72 | 0.76 | 0.69 |
| 18 | -0.41 | -0.04 | -0.03 | -0.26 | -0.35 | -0.26 | 1.08 | 1.02 | 0.89 | 1.07 | 1.34 | 0.90 |
| 19 | -0.10 | 0.74 | 0.78 | 0.59 | 0.49 | -0.57 | 1.24 | 1.15 | 1.12 | 1.10 | 1.34 | 1.21 |
| 20 | -0.24 | -0.56 | -0.02 | -0.61 | -0.56 | -0.41 | 0.85 | 0.86 | 0.81 | 0.84 | 0.93 | 0.85 |
| 21 | 0.02 | -0.43 | -0.54 | -0.16 | -0.35 | 0.24 | 1.22 | 1.29 | 1.19 | 1.15 | 1.32 | 1.35 |
| 22 | 0.62 | 0.74 | 0.63 | 0.69 | 0.90 | 0.51 | 0.82 | 1.01 | 0.83 | 0.73 | 1.14 | 0.95 |
| 23 | 0.78 | 1.00 | 1.21 | 1.07 | 0.54 | 0.74 | 1.03 | 0.95 | 0.78 | 0.97 | 1.04 | 1.17 |
| 24 | 0.95 | 1.30 | 1.43 | 0.95 | 1.26 | 0.86 | 1.64 | 1.85 | 1.31 | 1.86 | 2.16 | 1.65 |
| 25 | 1.75 | 1.70 | 1.46 | 1.41 | 1.57 | 2.45 | 1.49 | 1.24 | 1.16 | 1.16 | 1.53 | 1.60 |
| 26 | 1.23 | 1.03 | 1.07 | 0.98 | 1.34 | 1.13 | 0.86 | 0.76 | 0.77 | 0.66 | 0.91 | 0.89 |
| 27 | 1.73 | 0.98 | 1.20 | 1.23 | 1.25 | 1.75 | 1.36 | 1.26 | 1.55 | 1.21 | 1.20 | 1.30 |
| 28 | 0.84 | 0.30 | 0.88 | 0.47 | 0.42 | 0.51 | 1.02 | 1.24 | 1.29 | 1.28 | 1.05 | 0.90 |
| 29 | 1.89 | 1.61 | 2.23 | 1.93 | 1.55 | 1.29 | 1.40 | 1.25 | 1.39 | 1.36 | 1.28 | 1.27 |
| 30 | 1.38 | 1.39 | 1.97 | 1.68 | 1.39 | 0.50 | 1.27 | 1.28 | 1.05 | 1.43 | 1.28 | 1.35 |
| 31 | -0.19 | -0.40 | -0.13 | -0.29 | -0.34 | -0.42 | 0.61 | 0.69 | 0.70 | 0.73 | 0.62 | 0.54 |
| 32 | 2.01 | 2.75 | 2.94 | 2.82 | 1.88 | 1.88 | 1.33 | 1.66 | 1.45 | 1.53 | 1.34 | 1.67 |
| 33 | 2.12 | 2.37 | 2.41 | 2.35 | 2.17 | 2.05 | 1.25 | 1.14 | 1.00 | 1.10 | 1.20 | 1.47 |
| 34 | 1.80 | 1.28 | 1.86 | 1.91 | 1.46 | 0.94 | 0.98 | 0.93 | 1.09 | 1.02 | 0.84 | 0.87 |
| 35 | 2.19 | 1.46 | 1.68 | 1.58 | 2.04 | 2.00 | 0.74 | 0.57 | 0.71 | 0.57 | 0.59 | 0.75 |

TABLE F-4
WITHIN-GROUP ITEM PARAMETER ESTIMATES FOR PARAGRAPH COMPREHENSION

| | THRESHOLDS | | | | | | DISPERSIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | MALE | FEMALE | WHITE | POOR | BLACK | HISP | MALE | FEMALE | WHITE | POOR | BLACK | HISP |
| 1 | -0.52 | -0.30 | -0.34 | -0.45 | -0.26 | -0.58 | 0.88 | 0.91 | 0.88 | 0.92 | 0.84 | 0.93 |
| 2 | -1.42 | -1.23 | -1.36 | -1.25 | -1.48 | -1.21 | 0.98 | 0.88 | 0.87 | 0.79 | 1.06 | 1.00 |
| 3 | -0.53 | -0.58 | -0.78 | -0.48 | -0.59 | -0.38 | 0.62 | 0.60 | 0.64 | 0.57 | 0.67 | 0.56 |
| 4 | -0.25 | -0.22 | -0.06 | -0.17 | -0.40 | -0.32 | 1.15 | 1.08 | 1.07 | 1.17 | 1.10 | 1.13 |
| 5 | 0.23 | 0.08 | -0.02 | 0.08 | 0.27 | 0.27 | 0.71 | 0.87 | 0.74 | 0.85 | 0.81 | 0.75 |
| 6 | -0.16 | -0.10 | -0.51 | -0.36 | 0.38 | -0.03 | 1.41 | 1.44 | 1.27 | 1.30 | 1.62 | 1.50 |
| 7 | -0.34 | 0.18 | -0.39 | -0.05 | 0.23 | -0.11 | 0.97 | 1.03 | 1.03 | 0.86 | 1.07 | 1.03 |
| 8 | -0.69 | -0.81 | -0.50 | -0.73 | -0.79 | -0.98 | 1.42 | 1.19 | 1.47 | 1.34 | 1.04 | 1.38 |
| 9 | -0.11 | -0.34 | -0.11 | -0.12 | -0.41 | -0.27 | 0.80 | 0.84 | 0.81 | 0.84 | 0.87 | 0.75 |
| 10 | 1.38 | 1.60 | 1.46 | 1.49 | 1.42 | 1.60 | 0.90 | 0.86 | 0.76 | 0.93 | 0.87 | 0.96 |
| 11 | -0.27 | -0.84 | -0.08 | -0.71 | -0.74 | -0.69 | 1.23 | 1.39 | 1.35 | 1.30 | 1.30 | 1.29 |
| 12 | 0.73 | 1.06 | 0.68 | 0.90 | 0.99 | 1.02 | 1.23 | 0.98 | 1.13 | 1.07 | 1.10 | 1.12 |
| 13 | -0.23 | -0.50 | -0.33 | -0.59 | -0.22 | -0.33 | 0.75 | 0.70 | 0.78 | 0.71 | 0.72 | 0.69 |
| 14 | 0.70 | 0.76 | 0.61 | 0.80 | 0.69 | 0.83 | 0.89 | 0.85 | 0.96 | 0.89 | 0.79 | 0.85 |
| 15 | 1.49 | 1.24 | 1.72 | 1.66 | 0.91 | 1.17 | 1.81 | 2.32 | 2.03 | 2.50 | 1.90 | 1.82 |

TABLE F-5
WITHIN-GROUP ITEM PARAMETER ESTIMATES FOR AUTO & SHOP INFORMATION

| | THRESHOLDS | | | | | | DISPERSIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | MALE | FEMALE | WHITE | POOR | BLACK | HISP | MALE | FEMALE | WHITE | POOR | BLACK | HISP |
| 1 | -0.91 | -1.53 | -0.95 | -1.04 | -1.70 | -1.20 | 1.32 | 1.25 | 1.41 | 1.02 | 1.39 | 1.34 |
| 2 | -2.67 | -1.95 | -2.59 | -2.74 | -2.05 | -1.87 | 1.29 | 1.44 | 1.47 | 1.64 | 1.17 | 1.17 |
| 3 | -1.93 | 0.24 | -0.72 | -1.31 | -0.46 | -0.88 | 1.18 | 1.28 | 1.15 | 1.41 | 1.16 | 1.19 |
| 4 | -1.72 | -2.00 | -1.89 | -1.84 | -1.69 | -2.01 | 1.03 | 1.05 | 1.10 | 0.98 | 0.93 | 1.14 |
| 5 | -0.83 | -0.31 | -0.77 | -0.89 | -0.20 | -0.41 | 0.87 | 1.00 | 0.86 | 0.78 | 1.04 | 1.04 |
| 6 | -0.37 | 0.35 | -0.01 | -0.00 | -0.11 | 0.09 | 0.67 | 0.68 | 0.69 | 0.74 | 0.63 | 0.64 |
| 7 | 0.30 | 0.70 | 0.48 | 0.56 | 0.58 | 0.38 | 1.19 | 0.97 | 0.91 | 0.98 | 1.27 | 1.17 |
| 8 | -0.20 | -1.24 | -0.61 | -0.69 | -1.01 | -0.56 | 1.47 | 1.09 | 1.48 | 1.35 | 1.22 | 1.06 |
| 9 | -0.48 | -0.94 | -0.77 | -0.76 | -0.64 | -0.67 | 1.48 | 1.17 | 1.28 | 1.41 | 1.34 | 1.27 |
| 10 | 0.19 | -0.15 | -0.12 | 0.14 | -0.01 | 0.07 | 0.89 | 1.16 | 0.93 | 0.97 | 1.31 | 0.89 |
| 11 | 0.15 | -0.47 | -0.13 | -0.01 | -0.40 | -0.09 | 1.23 | 1.08 | 1.04 | 1.01 | 1.30 | 1.26 |
| 12 | 0.00 | -0.58 | -0.42 | -0.18 | -0.04 | -0.50 | 1.52 | 1.07 | 1.37 | 1.23 | 1.09 | 1.48 |
| 13 | -0.46 | -0.09 | -0.15 | -0.18 | -0.31 | -0.46 | 1.17 | 1.04 | 1.02 | 0.93 | 1.23 | 1.23 |
| 14 | -0.17 | -0.49 | -0.29 | -0.25 | -0.43 | -0.35 | 0.90 | 0.98 | 0.94 | 1.01 | 1.00 | 0.82 |
| 15 | 0.36 | 0.42 | 0.48 | 0.43 | 0.38 | 0.27 | 0.90 | 0.88 | 0.91 | 0.95 | 0.88 | 0.83 |
| 16 | 0.57 | 0.22 | 0.34 | 0.29 | 0.50 | 0.45 | 0.91 | 0.96 | 0.89 | 0.91 | 1.11 | 0.83 |
| 17 | 0.64 | 0.77 | 0.82 | 0.88 | 0.80 | 0.32 | 0.66 | 0.76 | 0.70 | 0.80 | 0.69 | 0.65 |
| 18 | 0.75 | 1.75 | 1.11 | 1.17 | 1.31 | 1.41 | 0.68 | 1.12 | 0.73 | 0.90 | 0.92 | 1.05 |
| 19 | -0.16 | -1.02 | -1.13 | -0.82 | -0.21 | -0.19 | 1.33 | 0.94 | 1.35 | 1.02 | 1.01 | 1.16 |
| 20 | 0.48 | 0.03 | 0.33 | 0.27 | 0.18 | 0.25 | 0.95 | 1.06 | 1.04 | 1.20 | 0.92 | 0.85 |
| 21 | 1.31 | 0.69 | 1.08 | 1.48 | 0.60 | 0.84 | 1.35 | 1.12 | 1.28 | 1.41 | 1.04 | 1.22 |
| 22 | 0.62 | 1.13 | 1.17 | 0.96 | 0.67 | 0.70 | 0.77 | 0.80 | 0.81 | 0.89 | 0.69 | 0.74 |
| 23 | 1.30 | 1.47 | 1.51 | 1.12 | 1.68 | 1.23 | 0.68 | 0.79 | 0.60 | 0.53 | 0.96 | 0.85 |
| 24 | 1.65 | 1.53 | 1.80 | 1.79 | 1.30 | 1.45 | 0.78 | 0.85 | 1.01 | 0.81 | 0.60 | 0.83 |
| 25 | 1.59 | 1.44 | 1.44 | 1.62 | 1.26 | 1.74 | 1.00 | 1.34 | 1.08 | 1.12 | 1.08 | 1.42 |

TABLE F-6
WITHIN-GROUP ITEM PARAMETER ESTIMATES FOR MATHEMATICS KNOWLEDGE

| | THRESHOLDS | | | | | | DISPERSIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | MALE | FEMALE | WHITE | POOR | BLACK | HISP | MALE | FEMALE | WHITE | POOR | BLACK | HISP |
| 1 | -2.39 | -2.07 | -2.72 | -2.42 | -1.89 | -1.90 | 0.88 | 0.80 | 0.86 | 0.87 | 0.79 | 0.86 |
| 2 | -1.93 | -2.06 | -2.33 | -1.87 | -2.13 | -1.65 | 1.27 | 1.48 | 1.64 | 1.14 | 1.45 | 1.27 |
| 3 | -0.67 | -0.55 | -0.99 | -0.81 | -0.20 | -0.45 | 1.00 | 0.79 | 1.18 | 0.82 | 0.73 | 0.84 |
| 4 | -0.90 | -1.53 | -1.04 | -1.10 | -1.58 | -1.12 | 0.90 | 1.09 | 0.88 | 1.04 | 1.03 | 1.03 |
| 5 | -2.37 | -1.23 | -3.36 | -1.69 | -1.04 | -1.10 | 1.71 | 1.41 | 2.44 | 1.51 | 1.14 | 1.16 |
| 6 | -0.35 | -0.48 | -0.22 | -0.37 | -0.53 | -0.53 | 0.79 | 0.83 | 0.81 | 0.89 | 0.87 | 0.68 |
| 7 | -0.82 | -1.17 | -0.88 | -1.00 | -1.15 | -0.94 | 0.93 | 1.12 | 0.95 | 0.99 | 1.09 | 1.08 |
| 8 | -0.54 | -0.35 | -0.58 | -0.46 | -0.36 | -0.39 | 1.18 | 1.13 | 1.18 | 1.07 | 1.13 | 1.22 |
| 9 | -0.70 | -1.09 | -0.79 | -0.85 | -1.01 | -0.93 | 1.04 | 1.06 | 1.03 | 1.00 | 1.00 | 1.16 |
| 10 | -0.49 | -0.77 | -0.72 | -0.80 | -0.66 | -0.34 | 1.16 | 1.19 | 1.30 | 1.26 | 1.05 | 1.09 |
| 11 | 0.23 | -0.40 | 0.20 | 0.13 | -0.31 | -0.37 | 1.04 | 1.26 | 1.20 | 1.37 | 0.98 | 1.04 |
| 12 | 0.33 | 0.35 | 0.30 | 0.24 | 0.58 | 0.23 | 0.82 | 0.90 | 0.84 | 0.86 | 1.01 | 0.74 |
| 13 | 0.09 | -0.27 | 0.10 | -0.14 | -0.07 | -0.23 | 0.70 | 0.74 | 0.65 | 0.69 | 0.83 | 0.71 |
| 14 | 0.41 | 0.80 | 0.60 | 0.27 | 1.10 | 0.45 | 0.76 | 0.74 | 0.78 | 0.88 | 0.71 | 0.64 |
| 15 | 1.52 | 1.82 | 1.72 | 2.26 | 1.18 | 1.53 | 2.26 | 2.45 | 1.72 | 2.74 | 2.29 | 2.68 |
| 16 | 0.01 | 0.01 | 0.22 | -0.16 | 0.15 | -0.18 | 1.07 | 1.04 | 1.01 | 0.99 | 1.16 | 1.07 |
| 17 | 0.65 | 0.43 | 0.59 | 0.47 | 0.49 | 0.60 | 0.70 | 0.75 | 0.76 | 0.68 | 0.75 | 0.69 |
| 18 | 0.26 | 0.23 | 0.53 | 0.36 | -0.02 | 0.11 | 0.73 | 0.84 | 0.72 | 0.83 | 0.75 | 0.83 |
| 19 | 0.08 | 0.12 | 0.13 | 0.08 | -0.01 | 0.20 | 1.05 | 1.83 | 1.20 | 1.30 | 1.52 | 1.74 |
| 20 | 0.82 | 0.59 | 0.93 | 0.36 | 0.78 | 0.76 | 1.47 | 1.30 | 1.34 | 1.13 | 1.66 | 1.42 |
| 21 | 0.88 | 1.01 | 1.33 | 1.04 | 0.79 | 0.62 | 1.28 | 1.08 | 1.10 | 1.27 | 1.29 | 1.06 |
| 22 | 1.14 | 1.11 | 1.64 | 1.57 | 0.52 | 0.77 | 0.84 | 0.81 | 0.71 | 0.85 | 0.86 | 0.89 |
| 23 | 1.34 | 1.81 | 1.57 | 1.53 | 1.56 | 1.63 | 0.76 | 0.61 | 0.59 | 0.64 | 0.68 | 0.82 |
| 24 | 1.65 | 2.06 | 1.80 | 1.73 | 1.95 | 1.93 | 1.11 | 0.87 | 1.05 | 1.05 | 0.83 | 1.02 |
| 25 | 1.74 | 1.63 | 1.98 | 1.63 | 1.84 | 1.30 | 0.92 | 0.64 | 0.71 | 0.65 | 0.84 | 0.91 |

TABLE F-7
WITHIN-GROUP ITEM PARAMETER ESTIMATES FOR MECHANICAL COMPREHENSION

| | THRESHOLDS | | | | | | DISPERSIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | MALE | FEMALE | WHITE | POOR | BLACK | HISP | MALE | FEMALE | WHITE | POOR | BLACK | HISP |
| 1 | -2.92 | -3.23 | -2.65 | -2.86 | -3.55 | -3.25 | 1.14 | 1.68 | 1.21 | 1.18 | 1.79 | 1.45 |
| 2 | -1.41 | -1.32 | -1.48 | -1.42 | -1.18 | -1.40 | 0.96 | 0.99 | 0.88 | 0.89 | 0.90 | 1.22 |
| 3 | -1.29 | -1.47 | -1.65 | -1.36 | -1.22 | -1.27 | 1.13 | 0.93 | 1.11 | 0.97 | 1.00 | 1.03 |
| 4 | 0.72 | 0.91 | 1.00 | 0.64 | 0.69 | 0.93 | 1.00 | 1.08 | 1.09 | 1.13 | 0.98 | 0.95 |
| 5 | -0.34 | -0.23 | -0.24 | -0.33 | -0.25 | -0.32 | 0.75 | 0.70 | 0.89 | 0.77 | 0.55 | 0.70 |
| 6 | -0.47 | -0.44 | -0.17 | -0.53 | -0.62 | -0.50 | 1.47 | 1.55 | 1.39 | 1.77 | 1.37 | 1.50 |
| 7 | -0.06 | 0.09 | -0.02 | 0.07 | 0.06 | -0.07 | 0.93 | 1.22 | 0.99 | 1.04 | 1.32 | 0.97 |
| 8 | -0.18 | -0.38 | -0.53 | -0.39 | -0.11 | -0.11 | 0.65 | 0.57 | 0.63 | 0.67 | 0.49 | 0.65 |
| 9 | -0.82 | -0.54 | -0.84 | -0.55 | -0.58 | -0.74 | 1.41 | 0.90 | 1.58 | 1.06 | 0.92 | 1.07 |
| 10 | -0.22 | -0.15 | -0.36 | -0.25 | 0.03 | -0.16 | 0.60 | 0.54 | 0.57 | 0.55 | 0.50 | 0.66 |
| 11 | -0.63 | -0.80 | -0.55 | -0.64 | -1.08 | -0.58 | 1.55 | 1.63 | 1.66 | 1.69 | 1.40 | 1.61 |
| 12 | -0.28 | 0.76 | 0.12 | 0.02 | 0.91 | -0.10 | 0.76 | 1.32 | 0.73 | 0.98 | 1.59 | 0.86 |
| 13 | -0.51 | -0.52 | -0.69 | -0.43 | -0.43 | -0.50 | 1.01 | 0.95 | 1.13 | 0.98 | 0.96 | 0.85 |
| 14 | 0.04 | -0.02 | -0.03 | -0.02 | 0.15 | -0.05 | 0.79 | 0.69 | 0.69 | 0.70 | 0.77 | 0.80 |
| 15 | -0.00 | 0.15 | 0.01 | 0.13 | 0.17 | -0.01 | 0.91 | 0.72 | 0.94 | 0.86 | 0.69 | 0.77 |
| 16 | -0.17 | -0.17 | -0.33 | -0.19 | -0.18 | 0.03 | 1.21 | 1.07 | 1.47 | 1.04 | 0.99 | 1.05 |
| 17 | 0.48 | 0.75 | 0.55 | 0.28 | 0.64 | 0.98 | 1.36 | 1.48 | 1.32 | 1.21 | 1.66 | 1.48 |
| 18 | 0.20 | 0.87 | 0.57 | 0.61 | 0.66 | 0.28 | 0.86 | 1.22 | 0.91 | 0.93 | 1.30 | 1.03 |
| 19 | 1.21 | 0.99 | 1.10 | 1.54 | 0.79 | 0.97 | 1.93 | 2.51 | 1.23 | 2.72 | 3.02 | 1.91 |
| 20 | 0.17 | 0.58 | 0.49 | 0.52 | 0.28 | 0.21 | 0.94 | 1.22 | 1.09 | 1.22 | 1.15 | 0.86 |
| 21 | 0.91 | 1.21 | 0.95 | 1.04 | 1.13 | 1.11 | 0.94 | 1.15 | 0.86 | 1.04 | 1.05 | 1.24 |
| 22 | 1.42 | 0.71 | 1.25 | 1.07 | 0.91 | 1.03 | 1.38 | 1.04 | 1.17 | 1.05 | 1.27 | 1.36 |
| 23 | 1.60 | 0.68 | 1.29 | 1.13 | 0.97 | 1.17 | 0.79 | 0.60 | 0.76 | 0.73 | 0.68 | 0.60 |
| 24 | 1.15 | 0.71 | 1.05 | 0.88 | 0.74 | 1.05 | 0.82 | 0.89 | 1.01 | 0.80 | 0.77 | 0.85 |
| 25 | 1.41 | 0.87 | 1.13 | 1.06 | 1.07 | 1.30 | 1.05 | 0.69 | 0.88 | 0.87 | 0.78 | 0.95 |

TABLE F-8
WITHIN-GROUP ITEM PARAMETER ESTIMATES FOR ELECTRONIC INFORMATION

| | THRESHOLDS | | | | | | DISPERSIONS | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ITEM | MALE | FEMALE | WHITE | POOR | BLACK | HISP | MALE | FEMALE | WHITE | POOR | BLACK | HISP |
| 1 | -1.71 | -1.71 | -1.66 | -1.63 | -1.77 | -1.78 | 1.01 | 0.90 | 0.95 | 0.92 | 0.84 | 1.11 |
| 2 | -1.10 | -0.99 | -1.12 | -1.01 | -1.14 | -0.91 | 0.84 | 1.03 | 0.82 | 0.89 | 1.05 | 0.98 |
| 3 | -1.11 | -0.66 | -1.12 | -1.22 | -0.58 | -0.62 | 0.93 | 0.82 | 0.83 | 0.95 | 0.85 | 0.85 |
| 4 | -1.36 | -0.81 | -1.14 | -1.10 | -1.17 | -0.93 | 0.83 | 0.70 | 0.75 | 0.76 | 0.76 | 0.79 |
| 5 | -1.44 | -1.37 | -1.37 | -1.25 | -1.42 | -1.58 | 1.16 | 0.84 | 0.98 | 0.96 | 0.81 | 1.27 |
| 6 | -0.81 | -0.76 | -0.81 | -0.72 | -0.75 | -0.85 | 0.67 | 0.72 | 0.63 | 0.73 | 0.77 | 0.66 |
| 7 | -1.23 | -0.68 | -0.92 | -0.64 | -1.12 | -1.16 | 1.08 | 1.32 | 1.06 | 1.12 | 1.32 | 1.29 |
| 8 | -0.52 | -0.83 | -0.70 | -0.72 | -0.70 | -0.58 | 0.93 | 0.83 | 0.89 | 0.97 | 0.82 | 0.82 |
| 9 | -1.24 | -0.89 | -1.21 | -1.07 | -0.92 | -1.07 | 1.13 | 1.28 | 1.13 | 1.20 | 1.16 | 1.34 |
| 10 | -0.22 | -0.43 | -0.23 | -0.29 | -0.45 | -0.33 | 0.80 | 0.72 | 0.90 | 0.81 | 0.63 | 0.69 |
| 11 | 0.84 | 0.37 | 0.66 | 0.53 | 0.58 | 0.64 | 1.80 | 1.44 | 1.46 | 1.43 | 2.08 | 1.51 |
| 12 | 0.36 | 0.24 | 0.23 | 0.48 | 0.34 | 0.14 | 0.77 | 0.69 | 0.81 | 0.74 | 0.76 | 0.60 |
| 13 | 0.43 | 1.44 | 0.40 | 0.48 | 1.38 | 1.49 | 0.85 | 1.06 | 0.93 | 0.90 | 1.01 | 0.97 |
| 14 | 1.15 | 0.04 | 0.79 | 0.51 | 0.18 | 0.90 | 1.44 | 1.64 | 1.77 | 1.72 | 1.26 | 1.41 |
| 15 | 0.12 | 0.55 | 0.34 | 0.40 | 0.33 | 0.27 | 0.74 | 0.91 | 0.79 | 0.82 | 0.85 | 0.83 |
| 16 | 0.84 | 0.82 | 0.92 | 0.91 | 0.99 | 0.50 | 1.24 | 1.32 | 1.36 | 1.38 | 1.37 | 1.00 |
| 17 | 4.20 | 3.79 | 3.81 | 3.60 | 4.52 | 4.07 | 2.49 | 2.77 | 1.94 | 2.32 | 3.56 | 2.72 |
| 18 | 0.62 | 0.42 | 0.79 | 0.56 | 0.33 | 0.40 | 0.97 | 0.72 | 0.97 | 0.92 | 0.84 | 0.67 |
| 19 | 0.88 | 0.34 | 0.76 | 0.88 | 0.37 | 0.42 | 0.97 | 0.93 | 1.07 | 0.89 | 0.87 | 0.96 |
| 20 | 1.31 | 1.13 | 1.58 | 1.31 | 1.02 | 0.99 | 0.75 | 1.04 | 0.97 | 0.79 | 0.71 | 1.09 |