DOCUMENT RESUME

ED 434 940	TM 030 193

AUTHOR	Krass, Iosif A.; Thomasson, Gary L.
TITLE	Defining Deficient Items by IRT Analysis of Calibration Data.
PUB DATE	1999-09-17
NOTE	20p.; Colored figures may not reproduce well.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; Algorithms; *Computer Assisted Testing; Nonparametric Statistics; Test Construction; *Test Items
IDENTIFIERS	Armed Services Vocational Aptitude Battery; *Calibration; Three Parameter Model

ABSTRACT

New items are being calibrated for the next generation of the computerized adaptive (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB) (Forms 5 and 6). The requirements that the items be "good" three-parameter logistic (3-PL) model items and typically "like" items in the previous CAT-ASVAB tests have resulted in a set of algorithmic rules that are general enough to be of interest to researchers and others working to improve CAT tests. For CAT-ASVAB on-line calibration, a seeded-item scheme is being used. Answers to the seeded items (new) are collected along with answers to nonseeded items to allow the estimation of examinee ability and the evaluation of the quality of each new item relative to the ability scale. Six types of deficient items are discussed. The first five types lead to nonparametric item characteristic curves that cannot be fitted well with a 3-PL model. The sixth category, in which the item has insufficient information, is the most difficult to correct. A set of factors has been developed to help deal with these problems and increase the value of item information. (Contains 6 tables and 6 figures.) (SLD)

Defining Deficient Items by IRT Analysis of Calibration Data

Iosif A. Krass
Gary L. Thomasson

Defining Deficient Items by IRT Analysis of Calibration Data

Iosif A. Krass and Gary L. Thomasson
Personnel Testing Division, Defense Manpower Data Center
Seaside, California

Introduction

The computer-adaptive testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB) has been in the field since 1994, with CAT-ASVAB Forms 1 and 2 as the first generation and CAT-ASVAB Forms 3 and 4 as the second (see Segall, Moreno, Bloxom, & Hetter, 1997). For CAT-ASVAB analysis, we have been using a 3-PL logistic model (Lord, 1977) to estimate the latent ability of an examinee and, consequently, to compute the examinee score; with this model the existing forms have provided good precision in estimating examinee ability.

Now we are calibrating new items in order to create the next generation of CAT-ASVAB, Forms 5 and 6. The requirements that the items be "good" 3-PL items and typically "like" items in the previous CAT-ASVAB tests have resulted in a set of algorithmic rules which, from our point of view, are rather general and thus should present interest for other researchers and item editors who are working to develop or improve their CAT tests.

For CAT ASVAB on-line calibration we are using a "seeded-item" scheme; that is, a new (seeded) item is given to every examinee in his or her second-through-forth position (chosen randomly) of each test in the battery. (For a detailed description of the seeded-item scheme, see Segall et. al., 1997.) Answers to the seeded items are collected along with answers to the non-seeded items, i.e., with answers on the CAT test, which allows us to estimate examinee ability for that particular test. It also allows us to judge the quality of the each new item relative to the ability scale.

After a long process of simulations and estimations, we came to the conclusion that a stable estimation of the Item Characteristic Curve (ICC) of a seeded item requires at least 1,400 answers on the studied item (Krass, 1998). Once the necessary number of responses per seeded item has been collected, the calibration process provides the 3-PL parameter estimation of the item (Krass, 1998).

Classification of Item Deficiencies and Example of Acceptable Item

The parametric model for the CAT-ASVAB test items is a 3-PL model, where the Item Characteristic Curve (ICC), of item i can be presented as,

$$P_i(\theta) = c_i + \frac{1-c_i}{1+\exp(l_i(\theta))}, \qquad (1)$$

where $P_i(6)$ is the probability for an examinee with the latent ability 6 to answer item i correctly. Here $l_i(6) = -D \cdot a_i \cdot (6 - b_i)$, and a_i, b_i, c_i; $i = 1, \cdots, I$ are the item discriminating, difficulty, and guessing indexes, correspondingly, and $D = 1.7$ is a scaling constant (Lord, 1980). We assume that examinee ability, $6 \in [-3.0, +3.0]$. Thus, after the process of calibration each seeded item should have its 3-PL representation (1) with estimated parameters.

Not all seeded items can be used in a future CAT or paper-and-pencil test because, for different reasons, the items are deficient. We have classified these deficiencies into six different categories:

A1 (classical) – biserial of correct answer is not significantly positive

A2 – a distractor has significant p-value and distracts examinees with higher test scores

A3 – a distractor has significant p-value and distracts examinees with scores not significantly lower than examinees with high test scores

A4 – examinees with high test scores lave less than .5 chance to answer item correctly

A5 – item has significantly non-monotonic experimental ICC; p-values for the highest and lowest are too close

A6 – item provides insufficient information, not discriminating among examinees

In Figure1 we present a seeded item from the Word Knowledge (WK)[1] test that is acceptable, WK2C8077. This figure also shows an example of the editorial data page that we provide for every item after calibration.

The editorial data page for an item consists of three parts: distractor analysis, ICC analysis, and a table of basic statistical parameters for the item.

1. The upper part provides visual analysis of the item options: correct answer and distractors. For every option there is a graph of option density. Different options density is approximated through the solution of the Pearson differential equation which conserves the first four momentum of experimental Optional Characteristic Curve (see Pollard, 1977).

2. The central part of editorial page provides visual ICC analysis. Here the solid line curve presents the result of a non-parametric approximation for the experimental item frequencies (experimental ICC), which is considered as a smoothing curve for an experimental ICC. This approximation is done by Dr. M. Levine's ForScore algorithm (Levine, 1984; Levine, Drasgow, Williams, McCusker, & Thomasson, 1992). The dotted line is the best 3-PL approximation of the non-parametric curve. Also presented on this same chart is the graph of frequencies for the experimental ICC, with 95% confidence intervals shown by vertical solid lines. Number of examinees in the different ranges of ability is printed in the low part of the ICC analysis chart.

3. The last part of the page provides classical statistical values in table format (P-values and biserials for every option), as well as some IRT values (means of examinee ability who chose a given option).

The 3-PL Deficiency (Categories A2 - A5)

If an item has a deficiency which belongs to categories A2 through A5 we identify it as a 3-PL deficient. All deficiencies of the items considered in this section are based on the fact that adaptive items in the CAT1 – CAT4 do not have items which satisfy the properties defined by A2 – A5 categories. From our practice in on-line calibration we have found that for an item with A2 – A5 categories of deficiency it is very hard to fit a

[1] The acronyms for each of the subtests are as follows: GS - General Science; AR - Arithmetic Reasoning; WK - Word Knowledge; PC - Paragraph Comprehension; AI - Auto Information; SI - Shop Information; MK - Mathematics Knowledge; MC - Mechanical Comprehension; EI - Electronic Information; AO - Assembling Objects.

reasonable 3-PL ICC curve to an experimental or smoothed – non-parametric ICC curve. Non-parametric ICCs for those items usually have more than one mode so fitting the 3-PL ICC to the non-parametric curves in those cases cannot be done satisfactorily (non parametric ICC behave like "spaghetti" type curve). For this reason if an item has one of A2 – A5 deficiency we call the item "not 3-PL item" and correspondent deficiency we called "3-PL deficiency". If the item do not have 3-PL deficiency, its non-parametric ICC behaves reasonably well and can be easily approximated by 3-PL curve (see the case of acceptable item WK2C8077 on the Figure 1).

As a rule, the 3-PL deficient items are either out of the "aptitude" the examinee brings to the testing situation, most often in the technical subtests (see Tables 1, 2, and 3 in this paper), or errors were made during item development that resulted in a confusing, incorrect item.. This phenomenon can be also due to present some obscure dimension in the item formulation.

Although the item which have category deficiency A1 formally do not belongs to categories characterized item as a not 3-PL item (it is too "classical" for 3-PL characteristic of an item) we put the category A1 in this section by reason described below.

A1 Deficiency

The A1 deficiency we call a "classical" category of deficiency: the biserial of the correct answer is not significantly positive. Although we have not formally used this category of deficiency to describe an item as unacceptable for CAT use (because CAT is strictly an IRT instrument), after the calibration of 2000 items we have not found an item which simply has an A1 deficiency. If an item belongs to the A1 category, it always has at least one other deficiency which belongs to the A2 -A5 categories. Therefore, we include items with this type of deficiency in this section of 3-PL deficient items for completeness of analysis. For the same reason we do not present an example of an item with the A1 deficiency.

A2 Deficiency

In items that are classified as A2 deficient there is a distractor j such that the mean of examinees μ_j who chose j instead of correct answer is greater than the mean μ_K of examinees who chose the correct answer ($\mu_j > \mu_K$). This chosen distractor has to have a significantly positive P-value (compared with the P-value of correct answer) to be a reason for A2 deficiency of the item. We denote by π_j, $j = 1,\ldots,J$, the P-value of distractor j, and by μ_j, $j = 1,\ldots,J$ the mean of ability of examinees, who chose this distractor, as the answer to the item, where J is the total number of item distractors. Then the item will have an A2 deficiency if two hypothesis – inequalities:

$$\frac{\pi_j}{\pi_K} > 0; \quad \text{and} \quad \mu_j - \mu_K > 0, \qquad (2)$$

are held with 95% of confidence; here π_K is the P-value of the correct answer option. This type of deficiency we will call a "stronger" case of existing a "too hard distractor". An example of this case of a considerably "stronger" distractor (B) is presented in Figure 2 for a WK item, WK4E8005.

A3 Deficiency

Based on the analysis of items in existing CAT- ASVAB versions, we have determined that the mean of ability of examinees who chose a particular distractor instead of the correct answer is typically significantly less than the mean of the examinees who answered the item correctly. Therefore, we developed a "weaker" case of presence of a "too hard distractor". By this we mean a distractor exists which has a significant P-value (compared with the P-value of correct answer), and the mean of the ability examinees who chose this distractor is not significantly less than the mean of the examinees who chose the correct response. In other words, the first inequality in (2) is holding with 95% of confidence, and the difference $\mu_K - \mu_j$ is not significantly positive. Figure 3 presents an item illustrating the A3 deficiency (MKB18078). Although definitions of A2 and A3 deficiencies are very close, we choose to divide them having in mind that, if an item have only A3 deficiency it can be reconsider as a candidate for set of acceptable items. Up to now we did not find an example of this expected phenomena, but we only at the beginning of calibration process using seeded item design.

A4 Deficiency

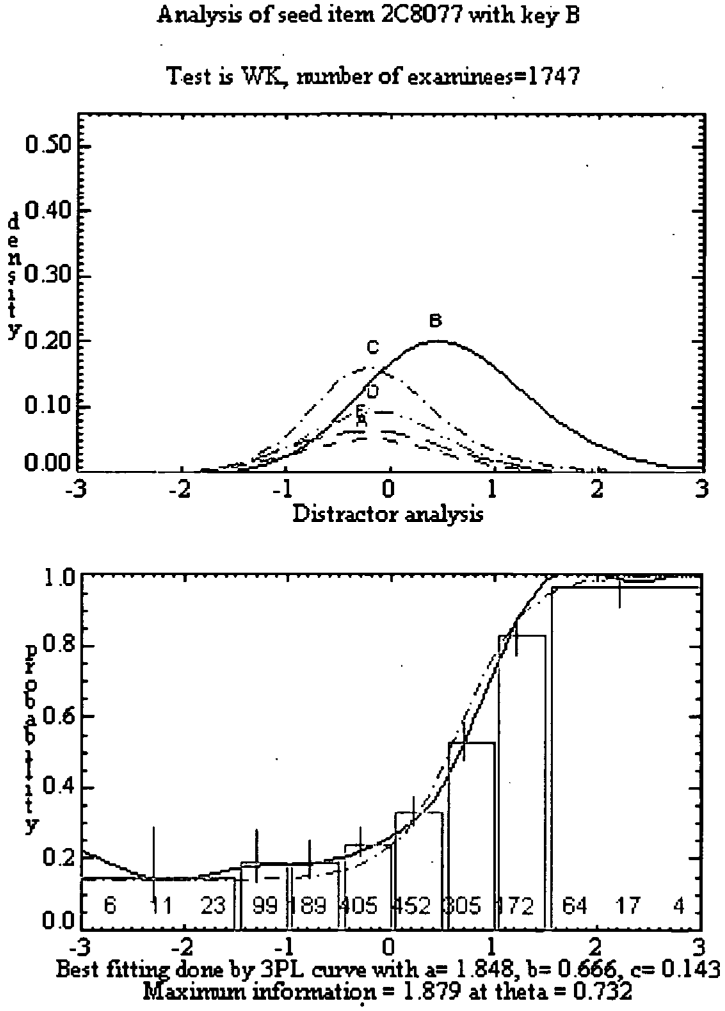
Another case of item deficiency is connected with the excessive hardness of the item for the recent population of examinees (category A4). To check this type of deficiency we define the range of ability $6 \in (1.5, +3.0]$ as a range "more able" examinees for the given test. For this range we compute value P^m - percent correct answers for the seeded item. If the seeded item is a "good" 3-PL item the value of P^m should be close to 1. By analyzing item pools of existing CAT-ASVAB tests we have found that, for items in those pools, the inequality $P^m \geq 0.5$ is held. The case of the opposite inequality $P^m < 0.5$ means that even higher scoring ("more able") examinees have less than a 0.5 chance to answer the item correctly. Therefore, we think that if $P^m < 0.5$, the item is excessively hard and should not be used for this population.

The editorial page presented in Figure4 illustrates an example of a category A4 item (ARB28181) for the AR test. Typical for this type of 3-PL deficient items, the non-parametric ICC for this item has more than one mode, and the steepest increase of the ICC occurs in the area where we do not have a sufficient number of examinees to make a reliable estimation. Item ARB28181 also has A2 deficiency; the mean of ability for examinees, who chose distractor B ($\mu_B = 0.545$) is higher than mean of ability of examinees, who answer the item correctly ($\mu_K = 0.213$), and the P-value of this distractor significantly positive.

A5 Deficiency

The next case of item deficiency (A5) is connected with a strongly non-monotonic response frequency and leads to an absence of a reasonable 3-PL model that should approximate this response frequency. To check the monotonicity of the experimental frequency we define the range of ability $6 \in (-3.0, -1.5)$ as the "lower" examinee ability area, compared with the interval of ability $6 \in (1.5, +3.0)$, defined earlier as the "higher" examinee ability area. For the area of "lower" ability intervals we compute the percent correct answers for the seeded item P^l. The value of P^l should be close to the guessing parameter c of the seeded item for an acceptable 3-PL item, and the value of P^m should be close to 1 for an acceptable seeded item. Thus, if P^m is not significantly higher than P^l, we will state that the experimental ICC is not monotonic and the item should be checked and possibly rewritten by the editors. "Significance" here is a 95% test of significance of the hypothesis $P^m > P^l$, checked by a one-sided Student distribution.

The item in Figure 5 provides an example of an item that is judged A5 deficient. In the case of the WK4B8028 item, $P^l > P^m$ and the 3-PL approximation is not acceptable. This item also has an A3 deficiency: the mean ability for examinee who chose distractor C ($\mu_C = 0.114$) is not significantly higher than the mean of ability for examinees, who answer the item correctly ($\mu_K = 0.163$), and the P-value of this distractor is significantly



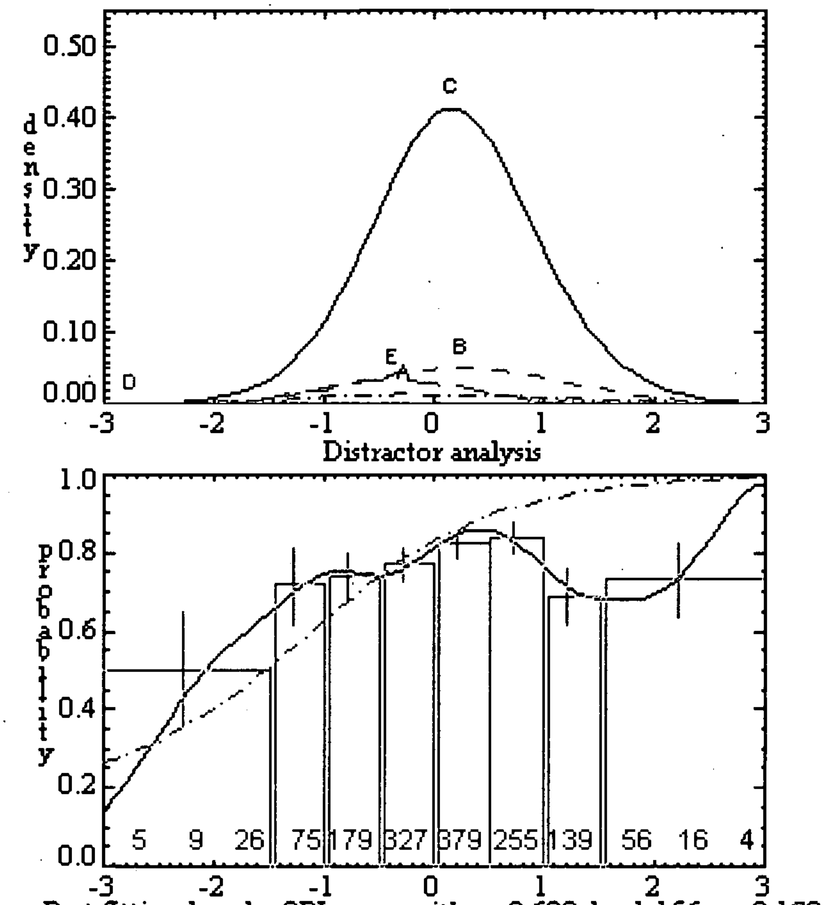
CLASSICAL TABLE

key	A	***B***	C	D	E
Pval	0.089	0.397	0.246	0.152	0.116
Biserial	-0.233	0.539	-0.221	-0.235	-0.231
Means	-0.193	0.588	-0.080	-0.146	-0.167

Figure 1. Example of an acceptable item for CAT-ASVAB (WK2C8077).

Analysis of seed item 4E8005 with key C

Test is WK, number of examinees=1470



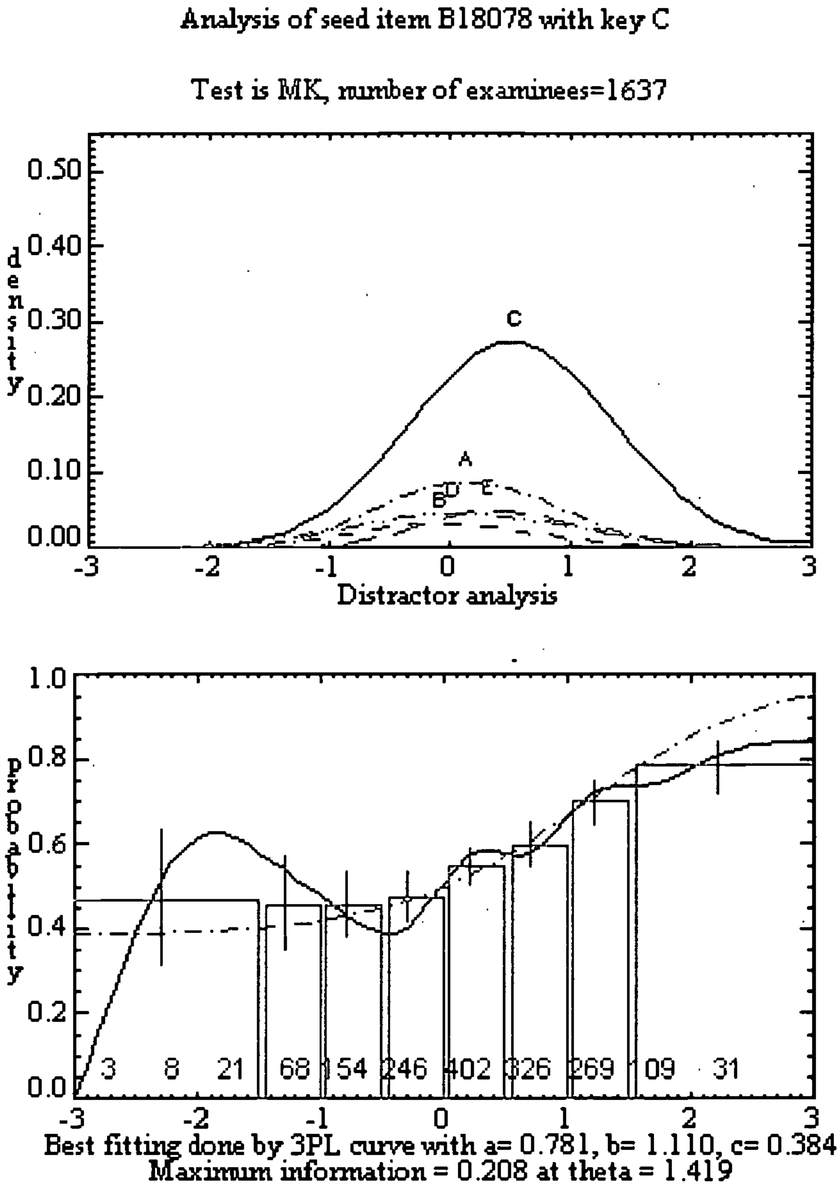
Best fitting done by 3PL curve with a= 0.683, b=-1.156, c= 0.179
Maximum information = 0.240 at thata =-0.945

CLASSICAL TABLE

key	A	B	C	D	E
Pval	0.040	0.121	0.773	0.003	0.063
Biserial	0.028	0.075	0.083	-0.553	-0.304
Means	0.203	0.257	0.180	-1.398	-0.348

Distractor B distract too smart people

Figure 2. Example of an unacceptable item because of A2 deficiency (WK4E8005).

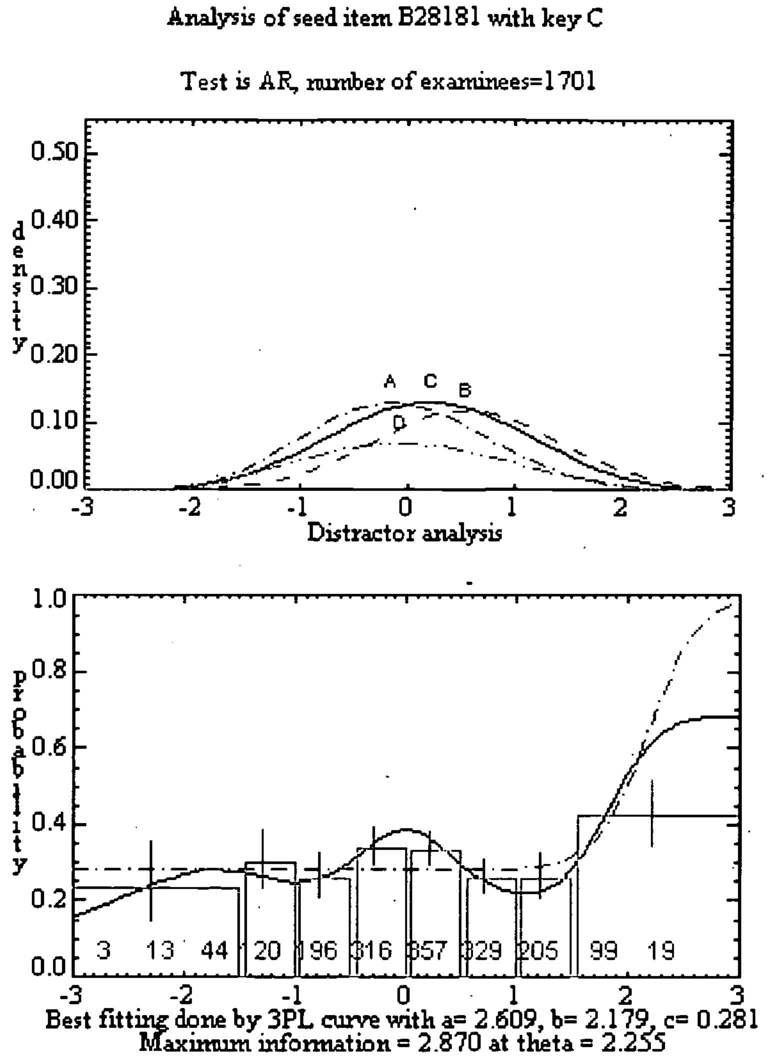


CLASSICAL TABLE

key	A	B	C	D	E
Pval	0.162	0.065	0.577	0.109	0.086
Biserial	-0.138	-0.229	0.226	-0.196	0.071
Means	0.183	-0.036	0.518	0.065	0.497

Distractor E distracts too smart people

Figure3. Example of an unacceptable item because of A3 deficiency (MKB18078).


Analysis of seed item B28181 with key C

Test is AR, number of examinees=1701

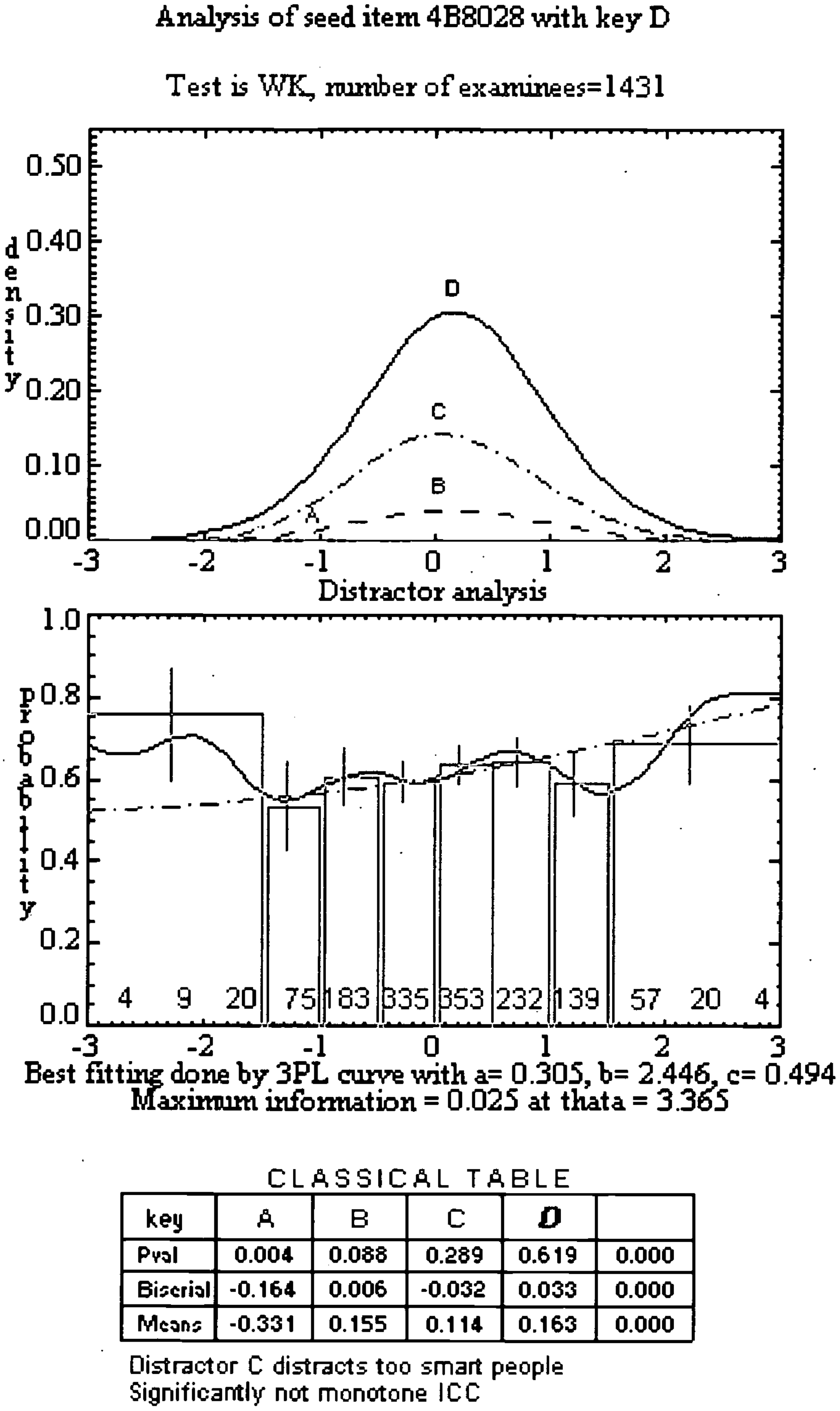
Best fitting done by 3PL curve with a= 2.609, b= 2.179, c= 0.281
Maximum information = 2.870 at theta = 2.255

CLASSICAL TABLE

key	A	B	C	D	
Pval	0.274	0.259	0.303	0.165	0.000
Biserial	-0.232	0.316	0.028	-0.140	0.000
Means	-0.075	0.545	0.213	-0.009	0.000

Distractor B distract too smart people
People has lesser than 50% chance to answer right

Figure 4. Example of an unacceptable item because of A4 and A2 deficiency (ARB28181).


Analysis of seed item 4B8028 with key D

Test is WK, number of examinees=1431

Best fitting done by 3PL curve with a= 0.305, b= 2.446, c= 0.494
Maximum information = 0.025 at thata = 3.365

CLASSICAL TABLE

key	A	B	C	D	
Pval	0.004	0.088	0.289	0.619	0.000
Biserial	-0.164	0.006	-0.032	0.033	0.000
Means	-0.331	0.155	0.114	0.163	0.000

Distractor C distracts too smart people
Significantly not monotone ICC

Figure 5. Example of an unacceptable item because of A3 and A5 deficiency (WK4B8028).

positive. This particular item presents a very easy vocabulary word ("brighten"), but the word has become so familiar that it no longer has one true definition, and that prevents accurate measurement.

Let us note that "cutoff" values $6 = -1.5; 6 = +1.5$ for "less able" or "more able" examinee will be converted to correspondent percentile "cutoffs" when more seeded item will be processed.

Unacceptable Items Without 3-PL Deficiency

If an item has no deficiencies in categories A1-A5, fitting the 3-PL ICC curve is reliable and yields the 3-PL parameters (a, b, c) of the item. Using these parameters we can compute item information $I(\theta) = \dfrac{D^2 a^2 (1-c)}{(c + e^{Da(\theta-b)})(1 + e^{-Da(a-\theta)})^2}$, at the ability level. Item information defines the precision of the CAT definition of examinee ability, because the CAT- ASVAB defines the ability approximately by maximizing the likelihood of examinee answers, and in this case Birnbaum (1968) shows that item information is reciprocal to the standard error of 6 estimation. Thus, if the item information is too small, that item is not a good instrument in ability estimation. A small value of item information is responsible for the last category of an item deficiency (A6).

A6 Deficiency

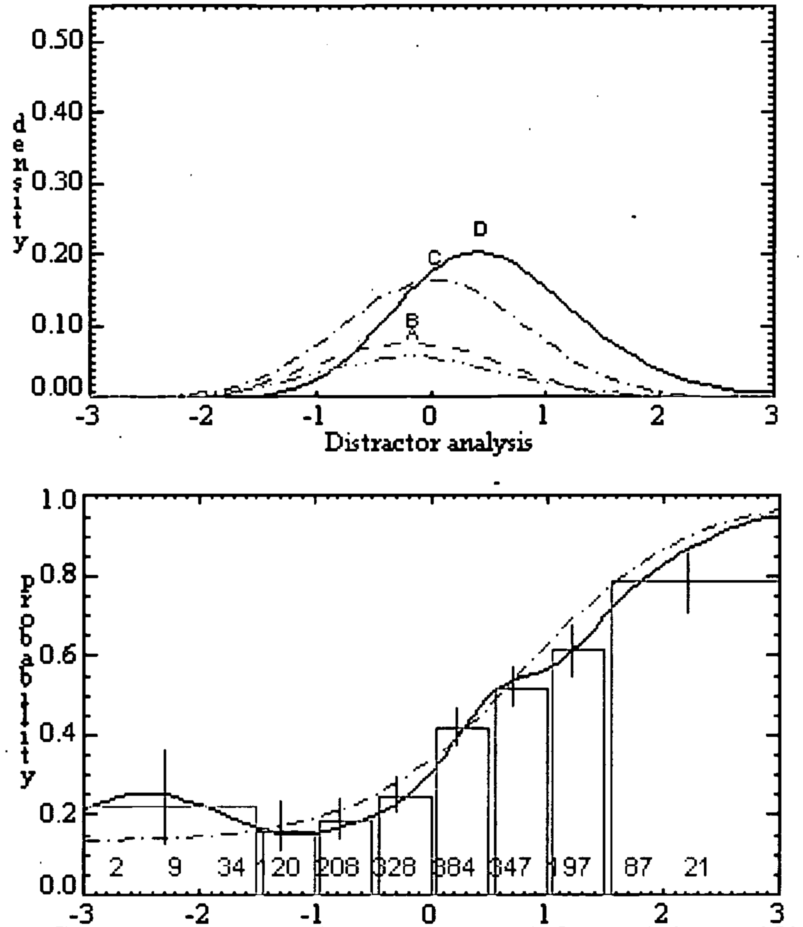
The last case of item deficiency (A6) is related to the efficiency of ability estimation and the method of selection of the next item in the test used in the CAT-ASVAB.. This method, excluding small randomization to restrict item exposure (Hetter & Sympson, 1997), is based completely on an information table. If an item has insufficient information, which means that the discriminating parameter a is too small, the item has no chance to be selected in a CAT test exam. To get the minimum requirements on item information, we went through the existing CAT item pools and estimate the maximum information that can be provided by the item, getting an "information profile" of the test.

Maximum information $I_{\max}$ for the particular item can be reached on maximizing ability $6_{\max}$ which is different for different items. (If the guessing parameter c is not too big, then $6_{\max}$ is close to the b parameter of the item). Thus, for a given test, we have determined the minimum level of $I_{\max}$ value (maximum information) in each ability interval for items used in CAT1 - CAT4, and these levels define the minimum boundary for $I_{\max}$ value all new seeded items in a particular ability interval. (In other words, there is not just one minimum boundary for a test, or a battery of tests; it depends on the ability interval where maximum information is reached.)

Both parameters $I_{\max}$ and $6_{\max}$ for each seeded item are computed and provided on the item analysis page. When maximum information provided by the seeded item is less than the lowest boundary defined by the test "information profile", this item is deficient in category A6. An example of this is presented in Figure 6 for item ARB28178. As we can see, the item ICC behaves more or less normally, but its "ascending limb" (ascending part of ICC curve) is not steep enough to meet our standards in that ability interval. This item would not have been chosen for use in any of the current CAT-ASVAB forms.

Analysis of seed item B28178 with key D

Test is AR, number of examinees=1737



Best fitting done by 3PL curve with a= 0.842, b= 0.801, c= 0.131
Maximum information = 0.398 at theta = 0.937

CLASSICAL TABLE

key	A	B	C	*D*	
Pval	0.118	0.151	0.332	0.400	0.000
Biserial	-0.220	-0.270	-0.167	0.433	0.000
Means	-0.135	-0.180	0.029	0.558	0.000

Information too small comparing with Operational CATs

Figure 6. Example of an item (ARB28178) with insufficient information (A6).

Correlation of Different Types of Item Deficiencies

Over the past year we have calibrated 200 items per ASVAB test (altogether 2000 items) and have identified deficient items in the different deficiency categories for different ASVAB tests (Tables 1, 2). We also have found that very often if an item is unacceptable by one of the 3-PL deficiencies, i. e. the item is unacceptable for a CAT-ASVAB form, it typically has more than one case of deficiency (as in the case of items WK4B8028 and ARB28181). This property is mirrored in the correlation table (Table 3).

Table 1. Absolute numbers of distribution of deficient items by ASVAB tests

Test	A1	A2	A3	A4	A5	A6	Mean
GS	30	36	15	21	26	17	24
AR	08	11	04	08	10	20	10
WK	06	08	02	02	07	23	08
PC	06	10	10	09	05	09	08
AI	43	53	21	30	32	22	33
SI	47	52	26	52	42	37	42
MK	09	09	07	13	09	08	09
MC	26	28	15	20	23	20	22
EI	29	34	19	36	24	03	24
AO	03	02	03	02	03	01	02
Mean	20	24	12	19	18	16	18

Table 2. Relative numbers of distribution of deficient items by ASVAB tests.

Test	A1	A2	A3	A4	A5	A6	Mean%
GS	15	18	07	10	13	09	12
AR	04	05	02	04	05	10	05
WK	03	04	01	01	03	11	04
PC	03	05	05	05	02	05	04
AI	21	27	10	15	16	11	17
SI	23	26	13	26	21	18	21
MK	05	05	03	07	05	04	05
MC	13	14	07	10	11	10	11
EI	15	17	10	18	12	01	12
AO	01	01	01	01	01	00	01
Mean%	10	12	06	10	09	08	09

Table 1 provides, for each test, the absolute number of deficient items in every category for 200 calibrated test items. Table 2 shows the same distribution in relative numbers (percentage or estimated probabilities).

Table 3. Matrix showing correlation among different deficiency categories.

Test	A1	A2	A3	A4	A5	A6
A1	1.00	0.99	0.95	0.92	0.99	0.57
A2	0.99	1.00	0.95	0.90	0.97	0.56
A3	0.95	0.95	1.00	0.96	0.94	0.44
A4	0.92	0.90	0.96	1.00	0.93	0.48
A5	0.99	0.97	0.94	0.93	1.00	0.63
A6	0.57	0.56	0.44	0.48	0.63	1.00

From Table 1 and Table 2, we see that items in AI, SI, GS, MC, and EI (the technical subtests) are more prone to categories A1 - A5 item deficiencies than the other tests. This is probably because we are measuring aptitude for which examinees are provided no general education; what they learn they learn on their own, and many examinees do not learn the material that is tested. The AFQT tests (AR, WK. PC, and MK) are less prone to these deficiencies because the tested material is part of every high school curriculum, and Test AO is less prone to those types of item deficiencies because it does not require any special knowledge.

Table 3 shows that categories A1-A5 are rather correlative among themselves, i.e., typically, if an item is unacceptable in one of the A1-A5 categories, it is unacceptable by another type of deficiency category. However, the A6 deficiency category – insufficient information – is not as strongly correlated with the other item deficiency categories. We apply this category only for the acceptable 3-PL items or to the items which do not have deficiencies in A2-A5 (and automatically in A1) categories. On average, we have identified about 170 (of 200) acceptable 3-PL seeded items per test. Moreover, we have found that all types of item deficiencies for tests AI, SI, and GS are highly correlative, which may be the result of the special nature of those tests.

Factors with Influence on Item Information Deficiency

Of the different categories of deficiencies, A6 (insufficient information) is the most puzzling for the item editors. If an item is deficient due to category A1-A5, an experienced editor typically understands what is wrong with the item and can often correct it and put it back into the item pool for on-line recalibration. But if the item is deficient due to insufficient information, its ICC curves often look normal, its distractors look normal, its set of classical data looks normal, and it is not obvious how to make the item more discriminative (e.g., to make its ICC curve more steep).

It appears to us that the most influential factors on the magnitude of item information are the item distractors (alternative choices). We have come to the conclusion that there are seven factors which most influence maximum information of an item or its discriminative parameter a. These factors are listed below.

The first factor is:

$$F_1 = \sum_{j=1}^{J} (\mu_K - \mu_j) \cdot P_j,$$

where μ_j, $j = 1,\ldots,J$ is mean of ability of examinees who chose distractor j, is μ_K - mean of ability of examinees who chose the correct answer, P_j is P-value of distractor j, and J is number of distractors (currently three or four). Factor F_1 we are calling "total leverage." A distractor has a big impact on this factor if its P-value is quite high or its mean is far from the mean of the correct answer.

The second factor is:

$$F_2 = (\mu_K - \mu_{bst}),$$

where μ_{bst} is the mean of the "best" distractor (distractor with maximal P-value). This factor can be called "best mean" – the farther the mean of the best distractor is from the mean of the correct answer, the higher the value for the factor.

The third factor is "total leverage with Standard Deviation":

$$F_3 = \sum_{j=1}^{J} (\mu_K - \mu_j) \cdot s_j \cdot P_j,$$

where s_j is estimation of SD for the frequency distribution of ability of those examinees who chose distractor j instead of correct answer. Value s_j influences the shape of distractor density shown on the editorial data page for the item. This factor is very close to the total leverage factor (F1); in fact, factors F1 and F3 are highly correlated.

The fourth factor is the "best leverage" factor:

$$F_4 = (\mu_K - \mu_{bst}) \cdot P_{bst}.$$

This factor is similar to the best mean (F_2) factor and correlates with it for some tests.

The fifth factor is "total relative leverage with SD":

$$F_5 = \sum_{j=1}^{J} \frac{\mu_K - \mu_j}{\theta_{\max} - \mu_j} \cdot s_j \cdot P_j,$$

where $6_{\max}$ is the upper bound of latent ability range. $6_{\max} = 3.0$ for CAT-ASVAB. Usually this factor is also highly correlative with "total leverage" factor F_1.

Distractor biserials (or its negative values) generate the next two factors (F6 and F7) because for an acceptable item, the biserials of distractors are negative, and we would like to have positive valued factors:

$$F_6 = -\sum_{j=1}^{J} BS_j,$$

$$F_7 = -BS_{bst},$$

where BS_j, BS_{bst} are biserials of distractor j or the best distractor correspondingly. Traditionally, editors prefer to work with distractor biserials, than with distractor IRT values, because there has not been much research to provide any connection between the IRT values of the distractors and the quality of the item. For this reason we include these biserial - based factors in the set of our factors.

Finally we have a compound factor to get maximal influence on the Maximum Information of an item:

$$F_8 = \sum_{i=1}^{7} w_i \cdot F_i,$$

where $w_i \geq 0$ weight of original factor F_i in the compound factor, and $\sum w_i = 1$. We chose weights to maximize heuristically the correlation between factor F_8 and the maximum information of the correspondent item. This is done by iterative applications of a SAS Canonical Correlation program.

Table 4 provides the correlations among the different factors and the maximum information of the item for different tests. As we can see, for most of the tests the highly influential factors are "total leverage," "best mean," or "total leverage with SD" (factors F_1, F_2, F_3 correspondingly).

Table 4. Correlation matrix among different factors and item maximum information

	F1	F2	F3	F4	F5	F6	F7	F8
GS	0.353	0.451	0.323	0.364	0.291	0.250	0.344	0.550
AR	0.336	0.313	0.312	0.288	0.273	0.095	0.123	0.491
WK	0.278	0.457	0.256	0.230	0.238	0.314	0.315	0.636
PC	0.337	0.509	0.297	0.233	0.252	0.352	0.374	0.642
AI	0.642	0.451	0.592	0.565	0.524	0.047	0.110	0.649
SI	0.567	0.672	0.509	0.614	0.480	0.344	0.417	0.754
MK	0.645	0.322	0.594	0.538	0.540	-0.062	0.044	0.654
MC	0.375	0.272	0.334	0.278	0.302	0.176	0.127	0.422
EI	0.539	0.273	0.524	0.542	0.510	-0.149	-0.001	0.569
AO	0.332	0.603	0.103	0.300	0.028	0.424	0.577	0.735

This is even more obvious in Table 5 which gives the correlation of F_8 with maximum information at least 0.42 and at most 0.75. In Table 5, maximizing weights are given as percentages. As we can see, factor F_1 is the most influential in most tests. Note that the zeroes in Table 5 of optimal weights can be rather deceiving. For example, factor F_3 for test GS has zero weight in the compound factor;

this is a result of high correlation between factor F_1 and F_3("total leverage" and "best leverage") for this test. In Table 6 we present the inter-correlation among different factors for the GS test. (We have not included correlation tables for the other tests in order to save space). As we can see, the correlation between factors F_1 and F_3 is 0.996 which means that these two factors estimated about the same property of the item and are highly exchangeable. Particularly, if we put the weight of factor F_1 to zero and the weight of factor F_3 to 57%, leaving the other weights without change for the GS test, we will get a correlation between compound factor F_8 and maximum information 0.538 instead of 0.55, which is the same influence on item information from a practical point of view.

Table 5. Optimal weights for compound factors

	F1	F2	F3	F4	F5	F6	F7	F8
GS	57.0%	23.4%	0.0%	9.5%	0.0%	10.2%	0.0%	0.550
AR	53.7%	19.7%	0.0%	0.0%	17.7%	8.9%	0.0%	0.491
WK	62.3%	28.4%	0.0%	0.0%	0.0%	9.2%	0.0%	0.636
PC	66.1%	21.3%	0.0%	0.0%	0.0%	12.6%	0.0%	0.642
AI	90.1%	0.0%	0.0%	2.0%	0.0%	7.8%	0.0%	0.649
SI	22.2%	21.8%	0.0%	43.9%	0.0%	12.1%	0.0%	0.754
MK	89.8%	10.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.654
MC	83.3%	4.4%	0.0%	0.0%	0.0%	12.3%	0.0%	0.422
EI	40.7%	10.0%	0.0%	49.3%	0.0%	0.0%	0.0%	0.569
AO	80.0%	14.4%	0.0%	0.0%	0.0%	5.6%	0.0%	0.735

Table 6. Correlation of original factors for test GS

	F1	F2	F3	F4	F5	F6	F7
F1	1.000	0.165	0.996	0.846	0.955	-0.256	-0.066
F2	0.165	1.000	0.145	0.309	0.105	0.534	0.937
F3	0.996	0.145	1.000	0.846	0.960	-0.288	-0.086
F4	0.846	0.309	0.846	1.000	0.849	-0.174	0.209
F5	0.955	0.105	0.960	0.849	1.000	-0.324	-0.109
F6	-0.256	0.534	-0.288	-0.174	-0.324	1.000	0.540
F7	-0.066	0.937	-0.086	0.209	-0.109	0.540	1.000

Conclusion

After our experience with on-line calibration, using a seeded-item design in the CAT-ASVAB, we have come to the conclusion that item unacceptability for future CAT- ASVAB tests can be classified into six categories of deficiencies. The first five categories lead to non-parametric Item Characteristic Curves that can not be fitted well with a 3-PL model, so we placed them in a larger 3-PL deficiency category. The sixth category – item has insufficient information - is the most difficult for editors to correct. For this reason we have developed a set of factors, which is easier to measure and which in most cases influences the maximum information provided by the item.

Table 5 of Optimal weights shows that for all tests, except SI and EI, the largest positive impact on the value of item information is the "total leverage" factor. As we see from the formula for this factor, it can be increased by adding a distractor which has a high P-value, or a distractor which will be attractive for examinees who have less knowledge of the subject matter. Those examinees will be in the "lower" ability level for the test, and the mean of ability for those examinees will be far away from the mean of ability for the examinees who knows the test material and answer the item correctly. This will increase the difference between the means, which defines the product in the "leverage" factor, and therefore, due to the presented correlation, will increase the value of item information.

From Table 6 (and analogous tables for other tests not presented here) we learn that for nearly all tests factors F_1 and F_4 are strongly correlated, which means that increased item information can be reached by changing only the distractor with the largest P-value - making this distractor more attractive for the "less able" examinees. In Figure 1, where we have an example of an acceptable and rather informative item, although the means for distractors B and D are not too far from the mean of the correct answer option, the P-values of those distractors are rather large (about 20%) which provide a rather high value for the "total leverage" factor.

It is interesting to note (again from the Table 6 and its analogous tables), that in most tests the "best biserial" factor F_7 is very highly correlated with the "best mean" factor F_2. This supports the idea that "classical" judging of item quality based on distractors biserials can be made with the same success as judging an item based on its IRT value.

Acknowledgment

The authors would like to express acknowledgment to G. Glick whose remarks and huge editorial experience inspired and much influence this research.

References

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Hetter, R. D., & Sympson, J. B. (1997). Item exposure control in CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing (pp. 141-145). Washington, DC: American Psychological Association.

Krass, I. A. (1998, April). Application of direct optimization for On-Line calibration in computer adaptive testing. Paper presented at the 1998th Annual meeting of National Council on Measurement in Education, San Diego, CA.

Krass, I. A. (1988, June). Recent advances in On Line calibration as applied to ASVAB. Paper presented at 1998th Annual meeting of Psychometric Society. Urbana, IL.

Levine, M. (1989). Formula scoring. Basic Theory and applications. (Technical Report 89-1) Champaign, IL: Model Based Measurement Laboratory, University of Illinois.

Levine, M., Drasgow, F., Williams, B., McCusker, C., & Thomasson, G. (1992). Measuring the difference between two models. Applied Psychological Measurement, Vol. 16, #3, 261-278.

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

Pollard, J. H. (1977). Numerical and Statistical Techniques. Cambridge University Press, Cambridge, London, New York.

Segall, D. O., Moreno, K. E., Bloxom, B. M., & Hetter, R. D. (1997). Psychometric procedures for administering CAT-ASVAB. In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing (pp. 131-140). Washington, DC: American Psychological Association.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

ERIC®

TM030193

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Defining Deficient Items by IRT Analysis of Calibration Data

Author(s): Iosif A. Krass and Gary L. Thomasson

Corporate Source: Defense Manpower Data Center
DoD Center-Monterey Bay

Publication Date: 09/17/99

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

□

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

□

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document es indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees end its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service egencies to satisfy information needs of educators in response to discrete inquiries.

Sign here,→ please

Signature: JKrass

Printed Name/Position/Title: Iosif A. Krass STATISTICIAN

Organization/Address: Defense Manpower Data Center
DoD Center-Monterey Bay
400 Gigling Road, Se
Seaside, CA 93955-6771

Telephone: 831/583-2400

FAX: 831/583-2339

E-Mail Address: KRASSIA@OSD.PENTAGON.MIL

Date: 09/17/99

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com