Research Report

A Graphical Approach to Item Analysis

Samuel A. Livingston

Neil J. Dorans

March 2004

RR-04-10

A Graphical Approach to Item Analysis

Samuel A. Livingston and Neil J. Dorans

ETS, Princeton, NJ

March 2004

Research Reports provide preliminary and limited dissemination of ETS research prior to publication. They are available without charge from:

Research Publications Office
Mail Stop 7-R
ETS
Princeton, NJ 08541

Abstract

This paper describes an approach to item analysis that is based on the estimation of a set of response curves for each item. The response curves show, at a glance, the difficulty and the discriminating power of the item and the popularity of each distractor, at any level of the criterion variable (e.g., total score). The curves are estimated by Gaussian kernel smoothing, a weighted moving average process with a parameter that can be varied at the user's discretion. The response curve for the correct answer can be accompanied by curves indicating a confidence region. The response curves also form the basis for estimating item statistics for any group of examinees for which the distribution of the criterion variable is known.

Key words: Item analysis, smoothing

Some things in educational testing have changed greatly over the past half-century. Others have not. Computing capabilities have changed tremendously. The purpose of item analysis has changed very little. Item analysis is still, as it was 50 years ago, the production of statistics to enable test developers to evaluate (a) the difficulty of each test item and (b) the relationship between performance on the item and performance on some more general measure—typically, the test containing the item. Fifty years ago, the statistics that could be used for this purpose were severely constrained by the limited computing capabilities available. Today, these constraints no longer apply. But in a large testing organization as in other environments, procedures tend to take on a life of their own as people become accustomed to them. As long as those procedures serve their purpose adequately, people see no compelling reason to look for something better. Thus it was that ETS found itself using, in the 1990s, item analysis procedures developed during the 1950s. As each new generation of computers arrived, the old statistical routines were adapted for the new machines.

As the 20th century entered its last decade, ETS began a major redesign of its statistical analysis procedures. The statisticians involved in the item analysis portion of this effort (a group that included the authors of this article) were asked, "What item analysis statistics should the system compute, and how should the results be presented to the test developers?" The purpose of this paper is to describe, explain, and illustrate the answers to those questions.

Why a Graphical Approach?

The statisticians working on the statistical analysis redesign project agreed that the most useful statistical information an item analysis could provide would be a series of conditional probability estimates. For each possible value of the criterion variable (e.g., the total score), these estimates would indicate the examinee's probability of answering the item correctly—and of choosing each distractor. These estimates could be plotted on a graph showing a response curve for the correct option and a response curve for each distractor. Such a graph would allow the test developer to see, at a glance, the most important statistical characteristics of the item: the difficulty of the item, the way its difficulty varied with the examinee's score on the criterion variable, the popularity of the individual distractors, and the way their popularity varied with the examinee's score on the criterion variable.

Why Not IRT?

Item response theory (IRT) was not a practical possibility when ETS first developed the item analysis procedures used from the 1950s to the 1990s. The computing technology necessary for large-scale operational use of IRT was either nonexistent or prohibitively expensive. But today, IRT and its applications are frequently used in the assembly and scoring of tests. (Wainer et al., 2000, and Thissen & Wainer, 2001, illustrate this point vividly.) Why did we and our colleagues choose not to use IRT?

The IRT models that are typically used in practice assume that an examinee's probability of choosing a particular response depends only on a single ability factor common to all the items on the test. This strong assumption produces local independence—statistical independence among items for examinees having any specified value of the ability factor. In addition, the mathematical form imposed on the response curve by most IRT models (e.g., a logistic ogive) is highly restrictive. IRT, with its use of a strong mathematical model, implies an obligation to test for model fit. But what if the goodness-of-fit test showed that the data for several items did not fit the model? Users of the system would be left without estimates of the response curves for those items. The developers of the ETS system chose a more flexible approach—one that allows the estimated response curve to take the shape implied by the data. Nonmonotonic curves, such as those observed with distractors, can be easily fit by this approach.

If Not IRT, Then What?

The approach we have chosen is a modified version of a technique used by Ramsay (1991). It is based on a weighted-moving-average smoothing procedure. This procedure is applied separately in estimating the response curve for each answer option: the correct answer, each incorrect answer, and the option of omitting the item. Each of the response curves is estimated separately. Each point on the response curve indicates the estimated probability that an examinee with a particular score on the criterion variable will choose that option.

The use of a weighted-moving-average smoothing procedure is based on the assumption that if all possible examinees were grouped into score levels on the criterion variable, the proportion choosing a given answer to a given item would change gradually—not abruptly—as the scores on the criterion increased. If this assumption is correct, each examinee's response to an item contains information that is useful for estimating performance on the item by examinees at the same score level *and also by examinees at nearby score levels*. The closer the score level, the more relevant the information, and this varying relevance is incorporated into the weights applied to the data.

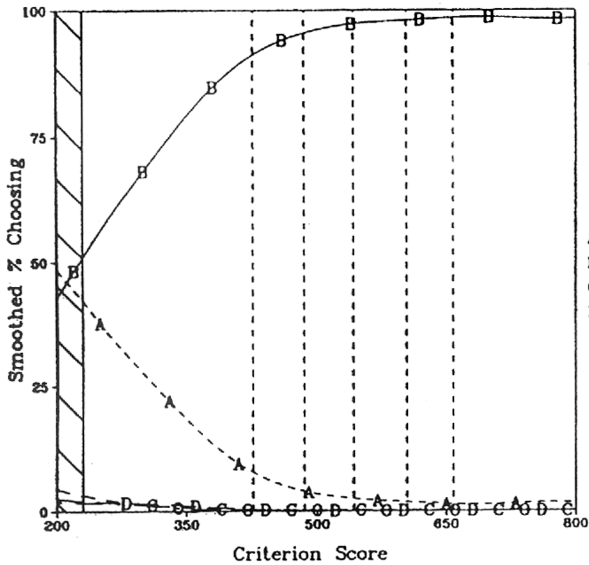
What Do the Graphs Look Like?

Figures 1 to 7 are examples of graphs based on actual data from the pretesting of items in a large-scale testing program. The items are four-option or five-option multiple-choice items. For a five-option item, a score of 1 is assigned to a correct responses, -1/4 is assigned to an incorrect response, and 0 to a nonresponse. An examinee responding at random (e.g., without reading the question) has a .20 probability of answering correctly. For a four-option item, an incorrect response is assigned a score of -1/3, and an examinee responding at random has a .25 probability of answering correctly.

The horizontal axis of the graph represents the score scale of the criterion variable. In these examples, the criterion variable is the examinee's scaled score on the test, and the numbers on the horizontal axis range from 200 to 800. The vertical axis represents the probability scale, from .00 to 1.00. The graph for each item includes a curve for the correct answer, a curve for each distractor, and a curve for omitting the item. The height of the curve at any point indicates the examinee's probability of choosing that answer option, given the examinee's score on the criterion variable.

Users of the system have the option to include in the graph a series of dashed vertical lines, indicating selected percentiles of the distribution of the criterion variable. These lines allow the test developer to relate the information in the graph to the abilities of the group of examinees. A high correct-answer probability in the middle of the score scale may mean one thing to the test developers if that point is near the 50th percentile of the score distribution, but it may mean something quite different if that point is near the 10th percentile. These lines also help the test developer see where the data were sparse and where the data were plentiful. The choice of percentiles is up to the user; in the following examples, the vertical lines represent the 20th, 40th, 60th, 80th, and 90th percentiles. (The tables to the right of the plot contain various statistics that describe the examinees' responses to the item. These statistics are described in a later section of this report.)

Figure 1 shows the graph for an easy item. Even the weakest examinees have a 50% probability of answering correctly, and this probability rises rapidly as the criterion score increases. One single distractor (A) seems to account for most of the item's effectiveness. The small diagonally hatched area at the far left of the graph indicates a region in which the response curves are not well estimated.



Multiple Choice, 4 choice

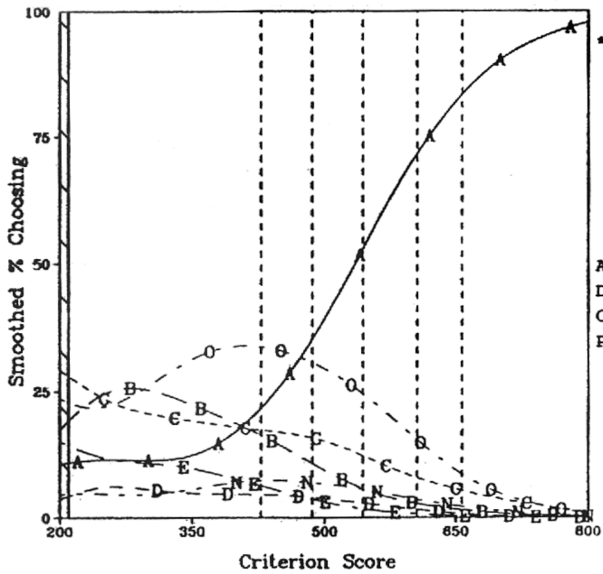
	N	%Tot	Criterion Mean	SD	Top 10%
A	673	6.1	405.7	104.4	1.2
*B	10277	93.2	523.9	100.7	98.8
C	32	0.3	395.6	114.1	0.0
D	29	0.3	367.2	80.4	0.0
Omt	21	0.2	378.1	85.2	0.0
NR	0				
Rch	11032	100.0	515.6	105.4	

	Observed	Ref.
Average Item Score	0.91	0.90
Delta	7.10	7.38
Correlation with Crit.	0.53	0.55
Percent Reached	100.00	

***Figure 1.* An easy item.**

Figure 2 shows an item that discriminates throughout the entire score range. The weakest examinees have a low probability of answering correctly (lower than they would if they responded at random). The strongest examinees have a probability of nearly 100% of answering correctly. The curve for the correct answer (A) rises steeply from the 20th percentile to the 90th percentile of the score distribution. A substantial number of examinees omit the item, and no single distractor seems particularly attractive to the examinees who answer incorrectly.

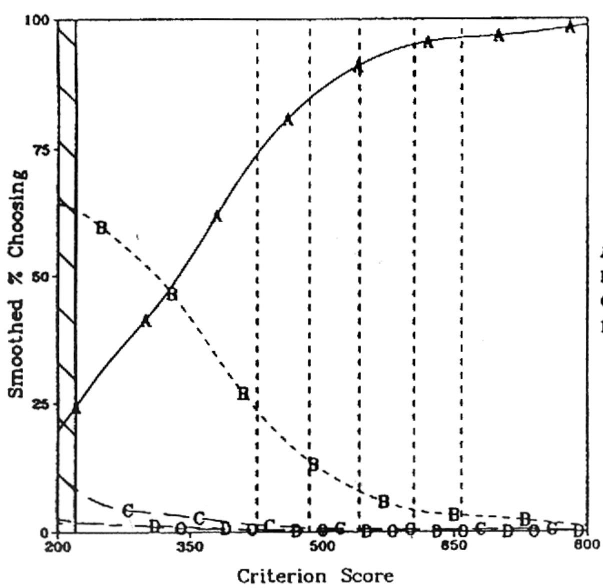
Figure 3 shows the graph for an item that discriminates well in the lower portion of the score range—below the 60th percentile of the score distribution. The examinee's probability of choosing the correct answer (A) rises from slightly better than chance (20%), for the weakest examinees, to nearly 90% at the 60th percentile of the score distribution. Again, a single distractor (B) accounts for most of the item's effectiveness.



	N	%Tot	Criterion Mean	SD	Top 10%
*A	4529	44.4	577.5	92.0	93.4
B	929	9.1	435.1	79.1	0.5
C	1252	12.3	470.5	90.3	2.3
D	299	2.9	460.4	80.1	0.1
E	352	3.5	418.6	78.4	0.1
Omt	2288	22.5	477.3	79.8	2.7
NR	541	5.3	482.9	77.9	0.9
Rch	9649	94.7	516.7	105.1	

	Observed	Ref.
Average Item Score	0.40	0.35
Delta	12.80	13.20
Correlation with Crit.	0.61	0.63
Percent Reached	94.69	

Figure 2. **An item that discriminates throughout the range score.**



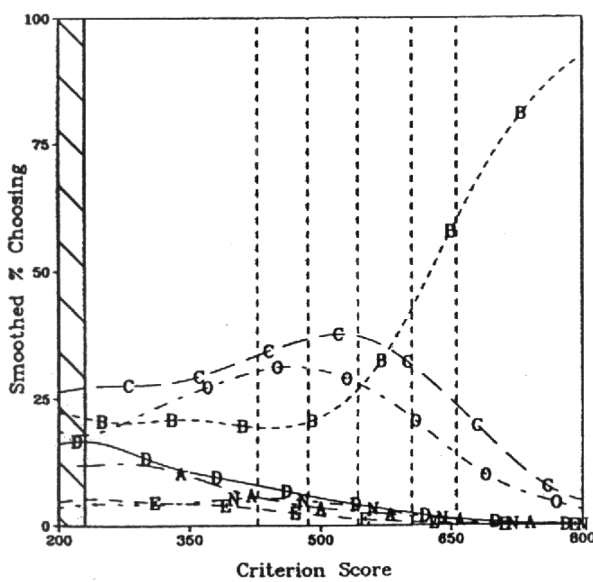
Multiple Choice, 4 choice

	N	%Tot	Criterion Mean	SD	Top 10%
*A	9134	82.8	535.1	97.0	97.1
B	1694	15.4	421.7	91.0	2.4
C	130	1.2	426.2	116.4	0.5
D	36	0.3	398.6	86.2	0.0
Omt	38	0.3	419.7	86.6	0.0
NR	0				
Rch	11032	100.0	515.6	105.4	

	Observed	Ref.
Average Item Score	0.77	0.75
Delta	9.18	9.49
Correlation with Crit.	0.59	0.61
Percent Reached	100.00	

Figure 3. **An item that discriminates in the lower part of the score range.**

Figure 4 shows the graph for an item that discriminates well in the upper portion of the score range—above the 60th percentile of the score distribution. Below the 40th percentile, the examinee's probability of choosing the correct answer (B) is about that of a person who responds at random. But above the 60th percentile, the probability rises rapidly, exceeding 90% for the strongest examinees. Again, one distractor (C) is substantially more popular than the others, particularly with the middle-ability examinees. The third most popular choice, after the correct option B and distractor C, is to omit the item.



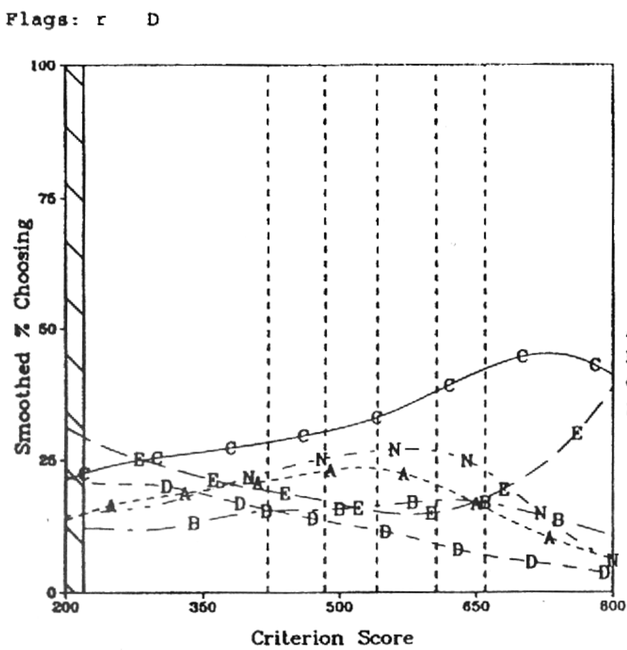
Multiple Choice, 5 choice

	N	%Tot	Criterion Mean	Criterion SD	Top 10%
A	390	3.8	439.3	93.0	0.5
*B	3189	31.3	571.1	113.4	81.2
C	3092	30.3	504.7	87.5	11.3
D	521	5.1	445.3	88.6	0.6
E	193	1.9	439.4	81.0	0.1
Omt	2436	23.9	494.0	83.2	6.1
NR	369	3.6	471.8	75.0	0.2
Rch	9821	96.4	516.6	104.7	

	Observed	Ref.
Average Item Score	0.22	0.19
Delta	14.33	14.54
Correlation with Crit.	0.38	0.39
Percent Reached	96.38	

***Figure 4.* An item that discriminates in the upper portion of the score range.**

Figure 5 shows the graph for an item that is too difficult for this population of examinees. The examinee response data do not clearly indicate the correct answer (E); someone attempting to infer the correct option from the graph might well choose distractor (C). Examinees at the 90th percentile performed at the chance level on this item. Even the strongest examinees chose distractor (C) more frequently than the correct answer. Note, however, that the slope of the curve for the correct option does begin to increase sharply at the upper end of the criterion score scale, suggesting that the item might discriminate effectively in a more able population. Because this item was the last item in a timed section, the curve for the examinees who did not respond to the item is labeled "N," for "not reached." Note that many examinees did not respond to this item, either because it was too difficult or because they ran out of time.



Multiple Choice, 5 choice

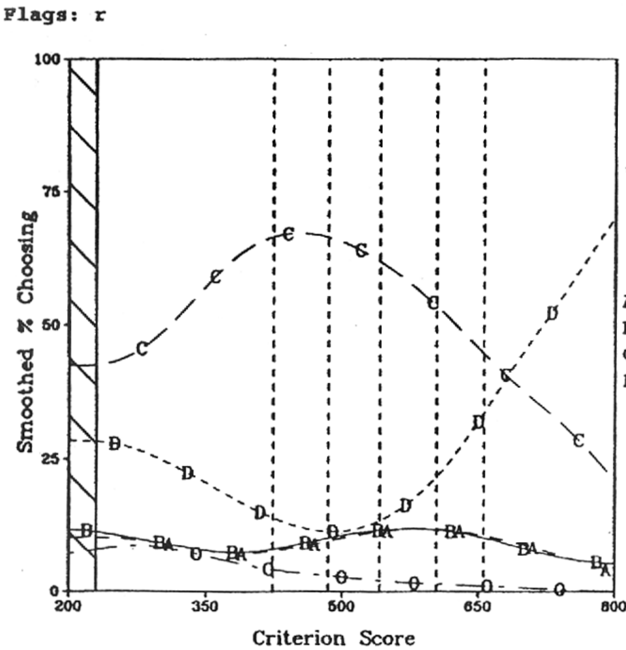
	N	%Tot	Criterion Mean	SD	Top 10%
A	1676	15.4	504.5	93.9	8.6
B	1298	11.9	519.2	102.6	11.6
C	2754	25.3	533.6	109.3	39.0
D	1043	9.6	474.9	102.2	4.3
*E	1542	14.2	511.1	126.2	22.1
Omt	0				
NR	2571	23.6	516.4	91.0	14.3
Rch	8313	76.4	514.0	109.5	

	Observed	Ref.
Average Item Score	-0.02	-0.02
Delta	16.51	16.66
Correlation with Crit.	-0.02	-0.02
Percent Reached	76.38	

Figure 5. **An item that is too difficult for the population of examinees.**

Figure 6 shows the graph for an item that illustrates the limitations of item response theory models for estimating item response curves. The response curve for the correct answer (D) is clearly not monotonic. The probability of answering this item *decreases* from the lowest score levels to the middle of the score distribution; then it rises sharply. Distractor C appears to be particularly attractive to examinees in the lower-middle portion of the score range, causing their probability of choosing the correct answer to be substantially less than it would be if they responded at random. However, even the strongest examinees have only about a 70% probability of answering this item correctly.

Figure 7 shows the graph for an item that is clearly not functioning as a measure of the skills measured by the test as a whole. The probability of answering this item correctly is at or below chance for examinees at all score levels. It is impossible to identify the correct answer (D) from the examinee response data. The most popular answer option throughout almost the entire score range is to omit the item. This item did not survive the pretest screening.

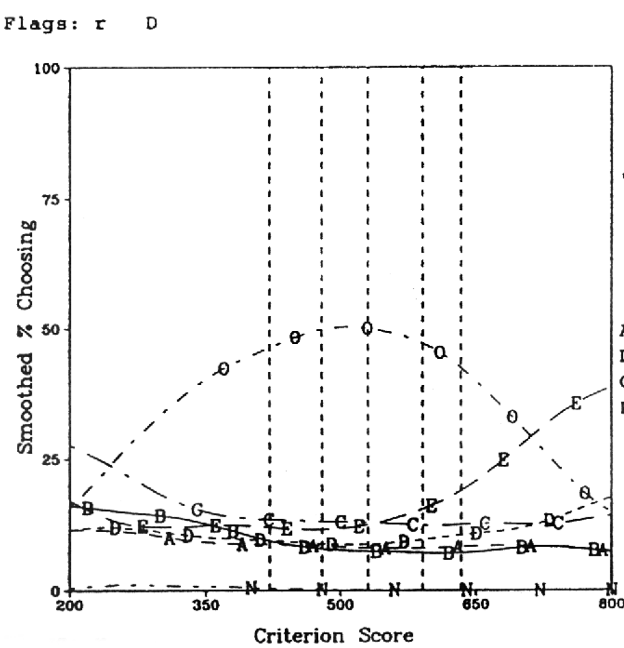


Multiple Choice, 4 choice

	N	%Tot	Criterion Mean	SD	Top 10%
A	951	9.6	522.1	98.1	7.2
B	970	9.8	519.6	98.1	6.9
C	5680	57.5	503.3	92.5	31.9
*D	1993	20.2	555.2	134.2	53.6
Omt	291	2.9	442.7	94.9	0.4
NR	0				
Rch	9885	100.0	515.4	106.1	

	Observed	Ref.
Average Item Score	-0.05	-0.07
Delta	16.23	16.37
Correlation with Crit.	0.19	0.20
Percent Reached	100.00	

Figure 6. **An item that is hardest for middle-ability examinees.**



Multiple Choice, 5 choice

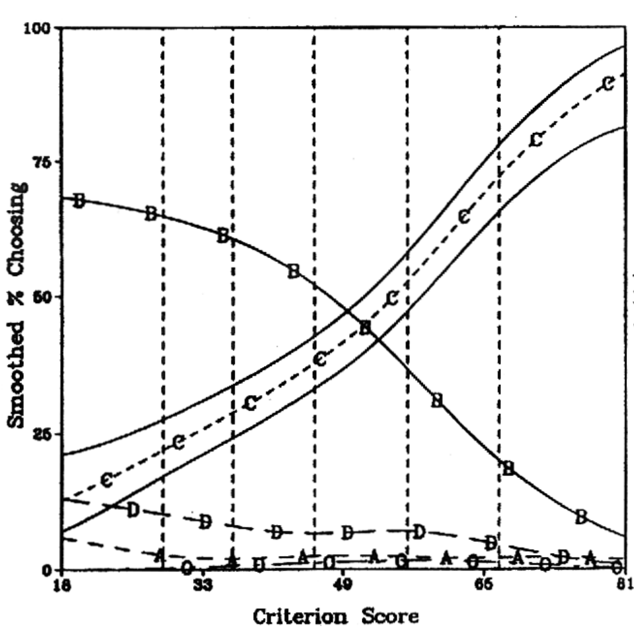
	N	%Tot	Criterion Mean	SD	Top 10%
A	995	8.5	499.9	103.9	8.6
B	1009	8.7	485.4	109.0	7.8
C	1574	13.5	495.8	107.8	13.5
*D	1137	9.8	510.0	109.7	12.0
E	1694	14.5	532.4	114.7	27.5
Omt	5219	44.8	504.5	87.0	30.6
NR	31	0.3	407.4	75.1	0.0
Rch	11628	99.7	505.9	100.9	

	Observed	Ref.
Average Item Score	-0.02	-0.02
Delta	16.51	16.51
Correlation with Crit.	0.00	0.00
Percent Reached	99.73	

Figure 7. **An item that does not work for this population.**

How Do the Graphs Indicate Sampling Variability?

The response curves that appear on the item analysis graph are estimates for a population of examinees like those included in the analysis. Because the examinees in the analysis are only a sample, the estimated response curves are affected by sampling variability. In some parts of the score range depicted in the graph, these effects can be quite large. It is important for the test developers to have a sense of how much confidence they can place in the information communicated by the response curves, particularly the curve for the correct answer. The item analysis procedure can communicate this type of information in two different ways at the discretion of the user. One way is to include in the graph a confidence band—a pair of curves above and below the estimated response curve for the correct answer, as in Figure 8. For any given score on the criterion variable, these curves indicate the upper and lower limits of an approximate 90% confidence interval for the probability of a correct response in the examinee population.[1] (The 90% confidence level is a default value; the user has the option to choose a confidence level other than 90%.)



Multiple Choice, 4 choice

	N	%Tot	Criterion Mean	SD	Top 10%
A	11	2.5	43.6	15.2	2.1
B	205	46.5	41.2	10.8	6.4
*C	189	42.9	54.7	13.3	91.5
D	32	7.3	41.7	12.4	0.0
Omt	4	0.9	52.0	7.6	0.0
NR	0				
Rch	441	100.0	47.2	13.8	

Average Item Score	0.43
Delta	13.71
Correlation with Crit.	0.57
Percent Reached	100.00

Figure 8. **A graph that uses confidence bands to indicate sampling variation.**

The item in Figure 8 comes from a different testing program than the items in Figures 1 to 7. The item analysis for this test used raw scores on the test as a criterion. The test consisted of four-option multiple-choice items, scored 1 for a correct response and 0 otherwise. The highest-scoring examinee answered 81 items correctly. While the samples of examinees (the number of examinees reaching the item) in Figures 1 to 7 ranged from 9,785 to 11,659, the sample in Figure 8 included only 441 examinees, all of whom reached the item. (The number of examinees reaching the item is listed as "Rch" in the statistics printed at the right of each graph.)

With samples as large as those in Figures 1 to 7, the confidence bands are so narrow that users of the system prefer not to include them in the graph. The system offers users an alternative method of indicating sampling variability. The user can request graphs with diagonal hatching that indicates the regions in which the response curve for the correct option is likely to be inaccurately estimated. (Figures 1 to 6 provide examples.) The user must specify a maximum acceptable value for the size of the confidence interval. The diagonal hatching indicates those portions of the graph where the confidence interval for the correct option is larger than this specified maximum.

Why the Numerical Information?

Although the graphs make it possible to evaluate the item at a glance, numerical item statistics still can be useful. When a test developer has to review a large number of items in a short time, it is useful to have the computer produce a list of items that may require special attention. Decision rules based on item statistics are the basis for this list. At the option of the user, items can be listed for being too difficult, for having too low a correlation with the item analysis criterion, for being too frequently omitted, or for having a particular incorrect option selected by too many of the examinees with very high scores on the criterion.

It is also useful to have a statistic that summarizes the difficulty of each item, so that the difficulty of a group of items can be summarized in the form of a distribution. For the same reason, it is useful to have a statistic that summarizes the discriminating power of each item. Finally, presenting the item statistics along with the graph has helped to ease the transition from a primarily numerical approach to a primarily graphic approach.

To the right of each plot are two tables. The upper table has a row for each answer option (including "omit" and "NR" for "not reached"). The entries in the row are statistics that refer to the examinees who chose that answer option. "N" is the number of examinees who chose the option. "% Tot" is the percentage of all the examinees who chose the option; it is "N" divided by the total number of examinees. "Criterion Mean" is the average score of these examinees on the criterion (which, in Figures 1 to 7, is their scaled score on the test). "Criterion SD" is the standard deviation of their scores on the criterion. The column labeled "Top 10%" shows the popularity of each answer option among the 10% of the examinees having the highest scores on the criterion variable. An option that was intended to be incorrect but was chosen by a substantial percentage of these high-scoring examinees will get a second look from the test developers. It could be simply a common misconception or a plausible wrong answer, but it could be an answer that is correct under some circumstances or under a plausible alternative interpretation of the question.

The information in the upper table reflects an approach opposite to that of the graph. The graph treats the examinee's score on the criterion as an input measure and the examinee's response to the item as an output. The statistics in the table effectively do the opposite; they treat the examinee's response to the item as input and the examinee's score on the criterion as output. In these statistics, the criterion scores of examinees who knew the correct answer are lumped together with those of examinees who chose the correct answer for the wrong reason and those of examinees who chose it by guessing at random. Inclusion of the mean score on the criterion variable for examinees choosing each option was a concession to historical practice. Some staff members were accustomed to using these statistics and would have objected to any item analysis system that did not include them. The standard deviations were added in the hope of discouraging overinterpretation of the option means. A large standard deviation for an option indicates that the examinees choosing that option varied substantially in ability, as indicated by the criterion variable.

The lower table contains two sets of summary statistics for the item. The statistics in the column labeled "Observed" refer to the group of all examinees included in the analysis. The statistics in the column labeled "Reference" are estimated for some other group of examinees—a reference group. The reference group can be any group of examinees—an actual group or a hypothetical group—whose distribution of scores on the criterion is known.[2]

The first row of the lower table contains a difficulty statistic: the average item score. For an item on which the only possible scores are 1 (for a correct answer) and 0 (for any other response), this statistic is simply the percentage choosing the correct answer. On a test, such as the SAT® I or SAT II, that is scored with a penalty for incorrect guessing, examinees who choose an incorrect response will have a negative score for the item. Therefore, the average item score will be somewhat lower than the percentage of examinees answering correctly—substantially lower, if the item is difficult.

The second row of the lower table contains the "delta" difficulty statistic. This statistic is simply a nonlinear transformation of the average item score, defined so that a high delta value indicates a difficult item. To transform the average item score to a delta value, let y represent the average item score, and compute $p = (y - y_{\min}) / (y_{\max} - y_{\min})$, where $y_{\min}$ and $y_{\max}$ represent the lowest and highest possible y-values.[3] Let z represent the pth percentile of the normal (0,1) distribution (i.e., the z-score that corresponds to the pth percentile rank in a normal distribution). Then the delta value for the item is 13–4z.

The third row of the table shows the correlation of the item with the criterion. If the only possible scores on the item are 1 and 0, this correlation is a biserial correlation. If more than two different scores on the item are possible, the correlation is a polyserial correlation, which is a generalization of the biserial correlation. The fourth row of the table, labeled "Percent Reached," shows the percentage of the examinees who reached the item (i.e., who answered either that item or at least one item appearing later in the same section of the test).

How Are the Response Curves Estimated?

The response curve for each answer option is estimated separately, as a series of data points—a separate point on the curve for each possible score on the criterion variable. (If the criterion variable is a continuous variable, it must be made discrete by partitioning the range of possible scores into discrete score levels. However, the user may specify as many as 5,000 score levels, making the variable effectively continuous.) The height of the curve at each point represents the estimated probability that an examinee with a particular score on the criterion variable will choose the answer option.

The first step in estimating the probabilities is to classify the examinees according to their scores on the item analysis criterion. Next, the procedure counts the number of examinees at each score and computes the proportion who chose each answer option. If the observed proportions for an answer option were simply plotted on the graph, the resulting curve would be a jagged, irregular line. The graph would be difficult to read, and *the irregularities in the curve would not generalize to another group of examinees*.[4] To produce a useful estimate of the response curve, it is necessary to apply a smoothing process. This process removes the irregularities while preserving the general shape of the curve.[5]

"Moving-average" smoothing replaces each observed data point with an estimate that averages the data from that point and nearby points. A simple moving-average process, applied at a given score on the criterion variable, would use the data for all examinees whose scores on the criterion variable are within a specified distance—and only those examinees. However, it does not seem reasonable to assume that as examinees' scores on the criterion get farther away from the value at which the response probability is to be estimated, the examinees' responses to the item change abruptly from being fully relevant to being completely irrelevant. It seems more reasonable to assume that their responses become gradually less and less relevant. This reasoning leads to the use of "*weighted*-moving-average" smoothing. The "smoothed" probability estimate at each criterion score level is a weighted average of the observed proportions at that score level and at other score levels—the closer the score level, the heavier the weight given to the data.

The particular variety of weighted-moving-average smoothing used in ETS item analysis is called "Gaussian kernel smoothing."[6] The weight given to each examinee's response is proportional to a normal (Gaussian) density function

$$\exp\left[-\frac{1}{2h}\left(\frac{x_i - x_k}{s_x}\right)^2\right]$$

where x_i is the examinee's score on the criterion, x_k is the score at which the proportion is to be estimated, h is a smoothing parameter, and s_x is the standard deviation of the scores on the criterion. Putting this formula into words, the weight given to the response of an examinee with criterion score x_i for determining the response probability at criterion score x_k is proportional to the height, at criterion score x_i, of a normal (Gaussian) density function centered at x_k with standard deviation $s_x\sqrt{h}$. The Gaussian density function is simply a convenient mathematical function that has the desired shape.

The value of the smoothing parameter is currently a function of the number of examinees included in the analysis. ETS staff are planning to modify the formula so that the default value will be a function of both the number of examinees included in the analysis and the number of score levels on the criterion variable.

How Is the Average Item Score Estimated for a Reference Group?

The average item score for a reference group—a group of examinees other than the group on which the analysis was done—is estimated by a post-stratification procedure. The stratifying variable is the item analysis criterion—the variable depicted in the horizontal axis of the graphs. Let i index the possible scores on the criterion variable and j index the possible scores on the item. Let $\hat{p}_{ij}$ represent the estimated probability that an examinee with a score of x_i on the criterion variable will earn a score of y_j on the item. The average item score, for examinees in the reference group with criterion score x_i , is estimated to be

$$\hat{\bar{y}}_i = \sum_j \hat{p}_{ij} y_j \,.$$

Then if n_i represents the number of examinees in the reference group with scores of x_i on the criterion variable, the average item score for the entire reference group is estimated to be

$$\frac{\sum_i n_i \hat{\bar{y}}_i}{\sum_i n_i} \,.$$

The delta statistic for the item in the reference group is estimated by applying the delta transformation (described above) to the estimated average item score (on the 0-to-1 scale) in the reference group.

How Is the Polyserial Correlation Estimated?

The polyserial correlation was originally defined as the correlation of two continuous variables in a bivariate normal distribution, when one of the variables can be measured only in terms of categories. However, the estimation procedure used in the ETS item analysis system does not assume a bivariate normal distribution. It is based on a more general model that includes the bivariate normal distribution as a special case. The estimation procedure was developed at ETS by Lewis (Lewis, Thayer, & Livingston, 2003), who called the estimated correlation "r-polyreg" (an abbreviation for "r-polyserial estimated by regression"). The procedure assumes that the item score Y is determined by the examinee's position on an underlying latent continuous variable η, which represents the examinee's ability to perform the task required by that item. The distribution of η for candidates with a given criterion score x is assumed to be normal with mean = βx and variance = 1, where β is an item parameter estimated from the data. The model can be written

$$P(Y \leq y_j \mid x) = P(\eta \leq \alpha_j \mid x) = \Phi(a_j - \beta x) ,$$

where y_j is the jth possible score on the item, α_j is the value of η corresponding to y_j, and Φ is the unit normal cumulative distribution function. The item analysis procedure estimates the value of β for each item by maximum likelihood. It uses this estimate of β to compute the polyserial correlation, by the formula

$$r_{polyreg} = \frac{\beta \sigma_x}{\sqrt{\left(\beta^2 \sigma_x^2 + 1\right)}} ,$$

where σ_x is the standard deviation of scores on the criterion variable in the group of examinees for which the polyserial correlation is to be estimated. That group of examinees could be the group of all examinees included in the analysis, or it could be any other group for which the standard deviation of scores on the criterion variable is known.

Summary

ETS's graphical approach to item analysis is based on the estimation of a set of response curves for each item. The response curves show, at a glance, the difficulty and the discriminating power of the item and the popularity of each distractor at any level of the criterion variable (e.g., total score). The curves are estimated by Gaussian kernel smoothing, a weighted-moving-average process. The response curve for the correct answer can be accompanied by curves indicating a confidence region. The response curves also form the basis for estimating item statistics for any group of examinees for which the distribution of the criterion variable is known.

References

Dorans, N. J. (2002). Recentering the SAT score distributions: How and why. *Journal of Educational Measurement, 39*(1)*,* 59-84.

Lewis, C., & Livingston S. A. (2003). *Confidence bands for a response probability function estimated by weighted moving average smoothing.* Manuscript submitted for publication.

Lewis, C., Thayer, D., & Livingston, S. A. (2003). *A regression-based polyserial correlation coefficient.* Unpublished manuscript.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56,* 611-630.

Thissen, D., & Wainer, H. (2001). *Test scoring.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., et al. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Notes

[1] The procedure for determining the confidence bands is described in Lewis and Livingston (2003).

[2] For example, the reference population for the SAT® I: Reasoning Test is the group of examinees whose scores were used in the 1990 analysis conducted for the purpose of recentering the score scale to a mean of 500 and a standard deviation of 110 (Dorans, 2002).

[3] Computing *p* in this way, for an item scored (1, 0, -1/[*k*-1]), has the effect of transforming the item scores to (1, 1/k, 0) and computing the average item score.

[4] An exception is the irregularities that are produced when the item analysis criterion is a score on a multiple-choice test on which noninteger item scores are possible but the total score is rounded to the nearest integer. These irregularities tend to replicate across groups of examinees and tests with the same number of questions, but they do not provide information that is useful to the test developers.

[5] There is an exception to the statement that the irregularities would not generalize to another group of examinees. The exception is the irregularities that are produced by the rounding of noninteger item scores. These irregularities tend to replicate across groups of examinees and across tests with the same number of questions, but they do not provide information that is useful to the test developers.

[6] This procedure is a simplified version of a similar procedure used by Ramsay (1991).