DOCUMENT RESUME

ED 230 569 TM 830 250

AUTHOR Dorans, Neil J.
TITLE Effects on Score Distributions of Deleting an Unkeyable Item from a Test.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-83-5
PUB DATE Feb 83
NOTE 65p.; Some tables contain small print.
AVAILABLE FROM Educational Testing Service, Research Publications Rm 116, Princeton, N.J. 08541.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS College Entrance Examinations; *Equated Scores; *Item Analysis; Mathematical Models; Psychometrics; *Scaling; Scores; Secondary Education; *Statistical Analysis; Test Construction; *Test Items
IDENTIFIERS Educational Testing Service; Flawed Items; *Item Deletion; Scholastic Aptitude Test; *Score Distribution

ABSTRACT

A formal analysis is presented of the effects of item deletion on equating/scaling functions and reported score distributions. The phrase "item deletion" refers to the process of changing the original key of a flawed item to either all options correct, including omits, or to no options correct, i.e., not scoring the flawed item. There are two aspects to the present analysis. The first aspect is analytical, focusing on the development of a formal model for the item deletion effect by decomposing it into its constituent elements. The second component of the analysis is empirical, involving the use of actual Scholastic Aptitude Test data to illustrate and supplement the analytical results. The analytical decomposition demonstrates how the effects of item properties, test properties, individual examinee responses and rounding rules combine to produce the item deletion effect on the equating/scaling function and candidate scores. In addition, the analytical component of the report examines the effects of not scoring vs. scoring all options correct and the effects of re-equating vs. not re-equating, as well as the interaction between the decision to re-equate or to not re-equate and the scoring option chosen for the flawed item. (Author/PN)

RESEARCH REPORT



# EFFECTS ON SCORE DISTRIBUTIONS OF DELETING AN UNKEYABLE ITEM FROM A TEST

Neil J. Dorans



February 1983



Educational Testing Service
Princeton, New Jersey

# Effects on Score Distributions of Deleting an Unkeyable Item From a Test

Neil J. Dorans

College Board Statistical Analysis

Educational Testing Service

This report benefited from the continued encouragement and careful reviews offered by Gary L. Marco and Nancy S. Petersen. Edwin O. Blew's adeptness with data analysis was invaluable, as was the secretarial support of Georgiana Thurston.

## ABSTRACT

The purpose of this report is to present a formal analysis of the effects of item deletion on equating/scaling functions and reported score distributions. The phrase "item deletion" shall be used to refer to the process of changing the original key of a flawed item to either all options correct, including omits, or to no options correct, i.e., not scoring the flawed item. There are two aspects to the present analysis. The first aspect is analytical, focusing on the development of a formal model for the item deletion effect by decomposing it into its constituent elements. The second component of the analysis is empirical, involving the use of actual data to illustrate and supplement the analytical results. The analytical decomposition demonstrates how the effects of item properties, test properties, individual examinee responses and rounding rules combine to produce the item deletion effect on the equating/scaling function and candidate scores. In addition to demonstrating how the deleted item's psychometric properties can affect the equating function, the analytical component of the report examines the effects of not scoring vs. scoring all options correct and the effects of re-equating vs. not re-equating, as well as the interaction between the decision to re-equate or to not re-equate and the scoring option chosen for the flawed item. The empirical portion of the report uses data from the May 1982 administration of the SAT, which contained the circles item, to illustrate the effects of item deletion on reported score distributions and equating functions. The empirical data verify what the analytical decomposition predicts.

EFFECTS ON SCORE DISTRIBUTIONS OF DELETING AN UNKEYABLE ITEM FROM A TEST

Within the past few years, the pyramid problem on a 1980 form of the PSAT/NMSQT and the adjacent circles item on a May 1982 form of the SAT have generated a great amount of press about items with indefensible keys. Wainer's (1981) large sample analysis of the PSAT pyramid problem, cleverly entitled, "Pyramid Power: Searching for an Error in Test Scoring with 830,000 Helpers", demonstrated that even a statistical analysis based on almost 830,000 examinees would not have revealed that the pyramid problem was miskeyed. It appears safe to anticipate a similar conclusion would be reached if a large sample analysis were performed on the adjacent circles problem. In addition to the highly visible external tempest created by these items, there has been a less visible yet vibrant discussion about what to do about defective items.

One of the available policy options is to delete the item from the test. This option was employed with the SAT adjacent circles item. Item deletion is operationalized by not scoring the item and effectively reducing the test length by a single item. Petersen (Note 1) summarized the effects of not scoring the adjacent circles item on equating and reported scores.

Very recently, two problem items appeared on a Biology Achievement test that was administered in June 1982. Consideration was given to either giving everyone credit or not scoring the two items under conditions of re-equating vs. not re-equating. When re-equating is employed, there is no psychometric difference between scaled scores based on giving everyone credit on the problem items and scaled scores based on not scoring the problem items. There is, however, a very

noticeable difference between giving everyone credit and not scoring, however, when re-equating is not employed. Petersen (Note 2, Note 3) summarized the effects of re-equating vs. not re-equating under various scoring options for the two problem items on the Biology Achievement Test.

The purpose of this report is to present a formal analysis of the effects of item deletion on equating/scaling functions and reported score distributions. The phrase "item deletion" shall be used to refer to the process of changing the original key of a flawed item to either all options correct, including omits, or to no options correct, i.e., not scoring the flawed item. Although neither not scoring the item nor scoring the item all options correct involve deletion of the item in a physical sense from the test booklet, the flawed item is, in both cases, deleted psychometrically from the test scores that candidates receive. A psychometrically deleted item has no psychometric impact on scores that individuals receive on the test. In other words, regardless of whether the item is difficult or easy, discriminating or not, it has the same impact on all candidates scores: If the item is not scored, all candidates receive no points for that item; If the item is scored all options correct, everyone receives one raw score point regardless of how they responded to the flawed item. As the title implies, this report is limited to item deletion in the sense just indicated. The effects of multiple keying of an item are not studied.

There are two aspects to the present analysis. The first aspect is analytical, focusing on the development of a formal model for the item

deletion effect by decomposing it into its constituent elements. Portions of this development are mathematically complex. For the benefit of the general reader, the salient features of this development are summarized in the next few paragraphs. The second component of the analysis is empirical, involving the use of actual data to illustrate and supplement the analytical results. This empirical component, which is less mathematically demanding to read than the analytical component, is summarized in the last two paragraphs of this introduction.

The analytical decomposition demonstrates how the effects of item properties, test properties, individual examinee responses and rounding rules combine to produce the item deletion effect on the equating/scaling function and candidate scores. Prior to decomposing the item deletion effects, the fundamentals of item response theory (IRT) true-scoring equating are described with the focus placed on the compositional nature of the IRT true-score equating process. In short, the equating process is composed of various new form and old form components. Item deletion affects the equating/scaling process through its effects on the new form components. The psychometric characteristics of items and rounding rules for formula (or raw) and scaled scores contribute to changes in the equating/scaling function. An item's difficulty determines where the change in equating function occurs along the raw (or formula) score scale. An item's discriminating power determines the abruptness and direction of the change. The item's suseptibility to guessing moderates the effects induced by the item's difficulty and discrimination. Rounding rules can exaggerate small effects in a rather unpredictable way.

In addition to demonstrating how the deleted item's psychometric properties can affect the equating function, the analytical component of the report examines the effects of not scoring vs. scoring all options correct and the effects of re-equating vs. not re-equating, as well as the interaction between the decision to re-equate or to not re-equate and the scoring option chosen for the flawed item. Not scoring the flawed item and scoring it all options correct affect the equating function in opposite ways. While the item's psychometric properties determine where the effect occurs, not scoring the flawed item results in a shorter test that is harder than the original test, while scoring all options correct results in a test that is as long as but easier than the original test. The flawed item's difficulty and the scoring decision determine how much the new conversion approximates the original for a given formula score. For example, while deleting a very difficult item may have no substantial effect on the equating function when the item is not scored, scoring that same difficult item all options correct can have a very noticeable impact on the equating function.

The analytical decomposition also examines the issue of re-equating vs. not re-equating. Not re-equating allows the flawed item's psychometric properties to have a substantial impact on scaled scores and also allows the decision to score all options correct vs. not score to impact on final reported scores. In contrast, re-equating makes the scoring decision irrelevant and mitigates the impact of the flawed item's psychometric properties on reported scores. Hence, re-equating after deletion of a flawed item is clearly better from a psychometric point of view.

The empirical portion of the report uses data from the May 1982 administration of the SAT, which contained the circles item, to illustrate the effects of item deletion on reported score distributions and equating functions. Six items from that test, in addition to the circles item, were selected for item deletion.

The effects of item deletion were studied in the following manner. Each of the six items was deleted from the 60-item total test containing the circles item and the 59-item test that excluded the circles item. As a consequence, six separate 59-item tests and six separate 58-item tests were simulated. Each of these 12 tests will be compared to the full 60-item test. One type of comparison focuses on equating/scaling functions and differences induced by deletion of items with certain properties. Effects on both rounded and unrounded converted scores will be assessed. Difference plots are used to examine these effects. In addition, effects of item deletion on examinee formula scores are assessed. This step necessitated rescoring for a representative sample of 45,579 examinees, the same set of examinees used to assess the effect of deleting the circles item. Differences among rounded and unrounded scaled scores are summarized.

The empirical data verify what the analytical decomposition predicts. For example, item difficulty determines where the change in equating functions occur and by how much reported scores produced by not re-equating under each scoring option (not score vs. score all options correct) differs from those produced by re-equating under either scoring option. The illustrative data demonstrates that re-equating mitigates

the impact of the deleted item's psychometric properties on reported scores. In fact, re-equating reduces the impact of the flawed item's characteristics on reported scores to an effect that is smaller than that associated with the rounding of scaled scores to two significant digits. In contrast, not re-equating enables the item's properties and the scoring decision to have very noticeable impacts on reported score distributions. In short, the illustrative data vividly demonstrates the importance of re-equating, and given re-equating the relative unimportance of flawed items psychometric characteristics. In the process, the relative importance of rounding rules is also illustrated.

## Analytical Decomposition

The effects of deleting an unkeyable item can be accounted for by three components:

- Changes in the equating function that maps rounded formula scores on the new form onto formula scores on the old form.
- Changes in the formula scores of individual examinees.
- The rounding of scaled scores to their two-significant-digit reported-score form.

The effects of these three components will be addressed in sequence.

### Changes in the Equating Function

This component of the item deletion effect is affected by the psychometric properties of the item. The particular effect depends on which equating method is employed, e.g., IRT true-score equating, linear equating, or equipercentile equating. Here, we focus on IRT true-score equating, the method employed for the SAT and the PSAT/NMSQT.

IRT true score equating. Lord (1980, pg. 198) demonstrates that observed scores on two tests cannot satisfy certain equating requirements unless either (1) both scores are perfectly reliable or (2) the two tests are strictly parallel, in which case equating is unnecessary. Since perfect reliability is virtually unattainable, observed-score equating is either unnecessary or impossible. Consequently, Lord advocates true-score equating.

Lord (1980, p. 199) cites three important requirements for equating two unidimensional tests that measure the same ability:

1. Equity: For every ability level, the conditional frequency distribution of equated scores from test X for a given ability level should equal the conditional frequency distribution of equated scores from test Y at that same ability level.

2. Invariance across groups: The equating function should be the same regardless of the population from which it was determined.

3. Symmetry: The equating relationship should be the same regardless of whether X is equated to Y or Y is equated to X.

IRT true-score equating meets these three requirements because true scores on tests measuring the same ability are perfectly related, i.e., there is an exact unique functional relationship between the true scores on the two tests. The equity condition is met because IRT true-score equating depends solely on IRT item parameters which theoretically are invariant across populations of examinees. Finally, symmetry follows from the identity relationship.

To appreciate the mechanics of IRT equating, we need to introduce some mathematical concepts and notation. We begin with the item response function, $P_g(\theta)$. The item response function is a mathematical expression for describing the probability of success on an item as a function of a single characteristic of the individual answering the item, his or her ability, and multiple characteristic of the item. The IRT model used for the SAT and the PSAT/NMSQT is the three-parameter logistic,

$$(1) \qquad P_g(\theta) = c_g + (1-c_g)\,[1+e^{-1.7a_g(\theta-b_g)}]^{-1},$$

where:

$P_g(\theta)$ - the probability that an examinee with ability $\theta$ answers item g correctly;

$a_g$ - item discrimination parameter for item g;

$b_g$ - item difficulty parameter for item g;

$c_g$ - lower asymptote of the item response curve, the probability that an examinee with extremely low ability answers item g correctly.

In (1), $\theta$ is the ability parameter, a characteristic of the examinee, and $a_g$, $b_g$ and $c_g$ are the item parameters that determine the shape of the item response function.

For a test composed of n items, summing the item response functions over the n items yields the test characteristic function

$$R = \sum_{g=1}^{n} P_g(\theta) . \tag{2}$$

The test characteristic function identifies the expected number-right score for each level of $\theta$. This expected number right score is the number-right true score on that test.

If test X and test Y are measures of the same ability $\theta$, then their number-right true scores are related to $\theta$ by their test characteristic functions

$$R_x = \sum_{i=1}^{n} P_i(\theta) \; ; \; R_y = \sum_{j=1}^{m} P_j(\theta) . \tag{3}$$

Note that $R_x$ and $R_y$ are functionally related to each other through their relationships with $\theta$. Substituting values of $\theta$ into $R_x$ and $R_y$ in (3) yields pairs of X and Y true scores. These pairs of true scores define $R_x$ as a function of $R_y$ and vice versa, and constitute an equating of true scores.

Let $t_x$ and $t_y$ refer to the test characteristic function transformations that convert $\theta$ to $R_x$ and $R_y$, respectively, i.e.,

$$R_x = t_x(\theta) \; ; \; R_y = t_y(\theta) . \tag{4}$$

Then, we can express $\theta$ as a function of $R_x$ and $R_y$ via

(5) $$t_x^{-1}(R_x) = \theta = t_y^{-1}(R_y) .$$

Let us designate X as the old form and Y as the new form. To find the transformation that equates Y to X, we first find the $\theta$ corresponding to a particular number-right true score on Y via

(6) $$\theta = t_y^{-1}(R_y) .$$

Next, we find the number-right true score on X corresponding to that $\theta$ via

(7) $$R_x = t_x(\theta) = t_x(t_y^{-1}(R_y)) .$$

Substituting a value of $R_y$ into (7) yields its equivalent in $R_x$ metric.

Both the SAT and the PSAT/NMSQT are formula-scored tests. In IRT, true formula scores on X and Y are defined via

(8) $$FS_x(\theta_a) = \sum_{i=1}^{n_a} P_i(\theta_a) - \sum_{i=1}^{n_a} [(1-P(\theta_a))/(A_i-1)]$$

and

(9) $$FS_y(\theta_a) = \sum_{j=1}^{m_a} P_j(\theta_a) - \sum_{j=1}^{m_a} [(1-P_j(\theta_a))/(A_j-1)]$$

where $n_a$ and $m_a$ are the number of items on X and Y that were reached by examinee a, and $A_i$ and $A_j$ are the number of response alternatives on items i and j, respectively. When an examinee reaches all items, and all items have A options, (8) and (9) simplify to

(10) $$FS_x = (AR_x-n)/(A-1) \; ; \; FS_y = (AR_y-m)/(A-1) .$$

For simplicity of exposition, we will assume all examinees reach every item, all items have the same number of options A, and $f_x$ and $f_y$ represent the transformations in (10), i.e.,

$$FS_x = f_x(R_x) \; ; \; FS_y = f_y(R_y) \; . \tag{11}$$

Rearrangement of terms in (10) yields,

$$R_x = ((A-1)FS_x+n)/A \; ; \; R_y = ((A-1)FS_y+m)/A \; , \tag{12}$$

which can be expressed as

$$R_x = f_x^{-1}(FS_x) \; ; \; R_y = f_y^{-1}(FS_y) \; . \tag{13}$$

IRT true formula-score equating proceeds as follows: The true formula score on new form Y is converted to a number right true score on Y via

$$R_y = f_y^{-1}(FS_y) \; . \tag{14}$$

Then, (6) is used to convert $R_y$ to $\theta$, and (7) converts $\theta$ to $R_x$, yielding

$$R_x = t_x(\theta) = t_x(t_y^{-1}(R_y)) = t_x(t_y^{-1}(f_y^{-1}(FS_y))) \; . \tag{15}$$

Next $R_y$ is converted to $FS_x$ via (11), yielding

$$FS_y = f_x(R_x) = f_x(t_x(t_y^{-1}(f_y^{-1}(FS_y)))) \; . \tag{16}$$

Equation (16) expresses the equating of true formula scores on Y to true formula scores on X. In practice, the transformation $s_x$ from $FS_x$ to scaled score is applied to the equated scores in (16) to place the Y test on scale, i.e.,

$$SS_y = s_x(f_x(t_x(t_y^{-1}(f_y^{-1}(FS_y))))) = S_y(FS_y) \; . \tag{17}$$

In sum, the scaling function for formula scores on new test y is

$$s_y = s_x \circ f_x \circ t_x \circ t_y^{-1} \circ f_y^{-1} . \quad (18)$$

where o indicates composition of functions.

Table 1 contains a convenient summary of the various functions involved in IRT equating and the scores they operate upon. In this table, the four types of scores, ability estimate ($\theta$), number right score ($R_x$, $R_y$), formula score ($FS_x$, $FS_y$), and scaled score ($SS_x$, $SS_y$) for old form X and new form Y are defined at the bottom. Above these score designations is a list of the functions. Alongside each function is a description of the mapping accomplished by that function and the number of the first equation containing that function. For example, $t_x$ maps $\theta$ onto number-right true score on test X, while $s_y$ maps formula scores on test Y, $FS_y$, onto the reported score scale for that test, $SS_y$.

The effect of item deletion. Equation (18) shows that the scaling function for test Y is a composite of several functions. Two of these functions deal with the relationships between test Y items and ability $\theta$, two deal with the relationships between test X items and ability $\theta$, and the fifth is the scaling function for test X. When an item is deleted from test Y, three of these functions are unaffected, namely, the functions associated with the old test X. The functions relating test Y true scores to $\theta$ are affected by item deletion. Hence, we shall focus on the effects of these functions. In this section and the following section, we presume that deletion of the item is accomplished by a decision not to score the deleted item. After the development for

Table 1

Functions and Scores Employed in IRT Equating[1]

| Function | maps | Score | onto | Score | Equation # |
|---|---|---|---|---|---|
| $t_x$ | : | $\theta$ | → | $R_x$ | 4 |
| $t_y$ | : | $\theta$ | → | $R_y$ | 4 |
| $t_x^{-1}$ | : | $R_x$ | → | $\theta$ | 5 |
| $t_y^{-1}$ | : | $R_y$ | → | $\theta$ | 5 |
| $f_x$ | : | $R_x$ | → | $FS_x$ | 11 |
| $f_y$ | : | $R_y$ | → | $FS_y$ | 11 |
| $f_x^{-1}$ | : | $FS_x$ | → | $R_x$ | 13 |
| $f_y^{-1}$ | : | $FS_y$ | → | $R_y$ | 13 |
| $s_x$ | : | $FS_x$ | → | $SS_x$ | 18 |
| $s_y$ | : | $FS_y$ | → | $SS_y$ | 17 |

| | Old Form | New Form[1] |
|---|---|---|
| Ability | $\theta$ | $\theta$ |
| Number Right | $R_x$ | $R_y$ |
| Formula Score | $FS_x$ | $FS_y$ |
| Scaled Score | $SS_x$ | $SS_y$ |

[1] A parallel set of scores and related functions exist for shortened tests.

deleting and not scoring, the alternative of "deleting" and giving credit to all options, including omits, will be considered.

The first function to be considered is $f_y^{-1}$ in (14). Let $f_{y'}^{-1}$ represent the formula score to number-right true-score conversion for the reduced test Y' composed of the m-1 items that remain after deletion of item k. Likewise, $t_{y'}$ represents the relation between number right true score and $\theta$ on test Y'. Hence, the scaling function for Y' is

$$s_{y'} = s_x \circ f_x \circ t_x \circ t_{y'}^{-1} \circ f_{y'}^{-1} . \tag{19}$$

The function $t_{y'}$ is defined by

$$R_{y'} = \sum_{\substack{j=1 \\ j \neq k}}^{m} P_j(\theta) . \tag{20}$$

Note that $R_{y'}$ can be related to $R_y$ for the full test via

$$R_{y'} = R_y - P_k(\theta) . \tag{21}$$

Hence, $t_y$ and $t_{y'}$ differ by the item characteristic function for the deleted item k. Note that for all $\theta$, $R_y \geq R_{y'}$. Since the item characteristic function for any item is a function of three item parameters, the particular change from $t_y$ to $t_{y'}$ is a function of these parameters. Hence, the deleted item's psychometric properties, embodied in the item discrimination parameter ($a_k$), the item difficulty parameter ($b_k$), and the lower asymptote ($c_k$), affect the equating function through

their affect on $t_y$. Deletion of the item affects $t_y$ in another important way, namely, it constricts the range of the function. Whereas $t_y$ maps $\theta$ onto a scale bounded by 0 and m, $t_{y'}$ maps $\theta$ onto a scale bounded by 0 and m-1. This restriction in the range of the function occurs regardless of the deleted item's psychometric properties and may be the major contributing factor to differences between $t_y$ and $t_{y'}$.

The function $f_{y'}^{-1}$ is embodied in

$$R_y = ((A-1)FS_{y'} + (m-1))/A \quad (22)$$

Given the relation in (21), $FS_y$ can be related to $FS_{y'}$ via

$$\begin{aligned} (FS_{y'}(A-1) + (m-1))/A &= R_y - P_k(\theta) \\ FS_{y'}(A-1) &= AR_y - AP_k(\theta) - (m-1) \\ FS_{y'} &= (AR_y - m)/(A-1) + (1-AR_k(\theta))/(A-1) \\ FS_{y'} &= FS_y + (a-AP_k(\theta))/(A-1) \quad . \end{aligned} \quad (23)$$

In (23), it is clear that true formula score is affected by the properties of the deleted item. Note that $FS_{y'}$ is greater than $FS_y$ for values of $\theta$ for which $P_k(\theta)$ is less than 1/A. This is an interesting result because it states that the expected formula score for individuals of very low ability can increase when an item is deleted despite the fact that the test is shortened by one item. For very low level examinees, the maximum increase is $(1-Ac_k)/(A-1)$. Since the minimum $c_k$ value is zero, the maximum gain in expected formula score is 1/(A-1), which for a five-choice item is .25. At the other extreme, very high ability individuals exhibit decreases in expected formula-score of (1-A)/(A-1), which is -1, precisely the decrease in expected number-right score at that level of ability.

Since $P_k(\theta)$ is always greater than zero, it can be inferred from (21) that for all $\theta$, $R_y \geq R_{y'}$. This inequality reflects that it is always easier to obtain a given number right on Y on than Y' regardless of which item is deleted. How much higher the expected score on Y is than the expected score on Y' depends on the properties of the deleted item and the individual's ability level. Since $R_y \geq R_{y'}$ for the same $\theta$, it follows that a given number-right score from Y' will convert to a larger (or as large) $\theta$-value than will the same number-right score from Y. In short, the longer test Y appears easier than the shorter test Y' because of the inequality $R_y \geq R_{y'}$, i.e., at all values of number right score, the function $t_{y'}^{-1}$ will exceed or equal $t_y^{-1}$, i.e., $t_{y'}^{-1} \geq t_y^{-1}$.

The same effect tends to occur for formula-scores as well. In particular,

(24) $$FS_y \geq FS_{y'} \quad \text{for all } P_k(\theta) \geq 1/A.$$

Since $P_k(\theta)$ should exceed 1/A for most values of $\theta$, it follows that for almost all values of $\theta$, a given formula score on Y' will convert to a larger (or as large) $\theta$-value than the same formula score on Y, i.e., $(t_y^{-1} \circ f_y^{-1}) \leq (t_{y'}^{-1} \circ f_{y'}^{-1})$ for a given formula score. When $c_k \geq 1/A$, this inequality will hold for all values of $\theta$. In short, the longer test Y will appear easier than the shorter test Y' for most if not all values of $\theta$. How much easier and for what values of $\theta$ will depend on the psychometric properties of the deleted item.

Since, $t_y^{-1} \leq t_{y'}^{-1}$ for all number right scores, and $f_y^{-1} \leq f_{y'}^{-1}$, will be true for most formula scores, the scaling function for Y' will tend to be higher than that of Y for most formula scores, i.e., $s_{y'} \geq s_y$ for most formula scores.

Effects of deletion on scaled scores. The relationship between $s_{y'}$ and $s_y$ can be constrained by the effects of the psychometric properties of the deleted item on the relationship between $FS_y$ and $FS_{y'}$. The general relationship can be expressed as:

- $SS_{y'} \geq SS_y$ for all $P_k(\theta) \geq 1/A$, and
  $SS_{y'} < SS_y$ for all $P_k(\theta) < 1/A$ .

Note that the point of which $P_k(\theta)$ equals $1/A$ depends on $a_k$, $b_k$, and $c_k$. Examination of specific cases of this general relationship proves enlightening.

- If $c_k = 1/A$ and $a_k = 0$, then $SS_{y'} = SS_y$ for all formula scores.

This unrealistic item is characterized by a flat item characteristic curve of height $1/A$. Such a curve would be observed if all examinees responded randomly to the item.

- If $c_k > 1/A$, then $SS_{y'} > SS_y$ for formula scores.

Whenever the lower asymptote exceeds the chance level, the shortened test is harder than the longer test at all levels of $\theta$.

- If $c_k = 1/A$, and $a_k$ is extremely large, then
  $FS_{y'} = FS_y$ for all $\theta < b_k$, and
  $FS_{y'} = FS_y - 1$ for all $\theta \geq b_k$,
  Hence, $SS_{y'} = SS_y$ for all $\theta < b_k$,
  and $SS_{y'} \geq SS_y$ for all $\theta \geq b_k$.

This item has a lower asymptote at the chance level and exhibits a very steep climb from $P_k \doteq 1/A$ to $P_k \doteq 1$ at $\theta = b_k$. In short, it is a highly discriminating item on which examinees either know the answer or guess randomly. While deletion of the item has no effect for those below $\theta = b_k$, the shorter test is harder for those whose $\theta \geq b_k$. Hence, $SS_{y'} \geq SS_y$ for this latter group.

- If $c_k = 0$ and $a_k$ is extremely large, then
  $FS_{y'} = FS_y + 1/(A-1)$ for all $\theta < b_k$, and
  $FS_{y'} = FS_y - 1$ for all $\theta \geq b_k$.
  Hence, $SS_{y'} < SS_y$ for all $\theta < b_k$.
  $SS_{y'} = SS_y$ at $\theta = b_k$, and
  $SS_{y'} > SS_y$ at $\theta > b_k$.

This is a sharply discriminating item that clearly separates those who know it from those who do not. Deletion of this item from the test has an interesting effect. The shortened test is harder for those with $\theta$ above $b_k$, and easier for those with $\theta$ below $b_k$. This example illustrates a general result that follows from the general relationship stated earlier: Deletion of an item makes the shortened test easier for examinees who perform below chance level on that item. For all others, the test is either as hard or harder than before.

Changes in Individual Examinee Scores

When an item is deleted from a test, the formula score of an individual examinee may or may not change. Wainer (Note 4) referred to the score on a test that an individual will have when a particular item is deleted from the test as the item's influence function (IIF). In the

formula score metric, each and every item has three item influence functions, one for each possible score on the item (omit, correct, or incorrect). These three distinct functions are the same across all items. In the formula score metric, these three IIFs are:

(26)
$$IIF\ (FS_y,\ Y_k = 1) = -1.0$$
$$IIF\ (FS_y,\ Y_k = 0) = 0.0$$
$$IIF\ (FS_y,\ Y_k = -1/(A_k-1)) = 1/(A_k-1).$$

Note that the three functions are independent of the examinee's ability and the item's psychometric properties. They depend solely on the examinee's response to the deleted item and the number of response alternatives.

The impact of rounding rules. When expressed in the reported scaled score metric, however, the three item influence functions do depend on examinee ability and the item's psychometric properties. In addition, rounding conventions for both formula scores and scaled scores impact on these influence functions, a point overlooked by Wainer (Note 4) in his treatment of item deletion. For a correct response to the deleted item, the influence function is

(27)
$$IFF(SS_y,\ Y_k=1) = r_{ss}(s_y, r_{fs}(FS_y^{-1}))) - r_{ss}(s_y(r_{ss}(FS_y)))$$

where in (27), $r_{ss}$ and $r_{fs}$ refer to the ETS rounding rules for reported scaled scores and formula scores, respectively. For both the SAT and the PSAT/NMSQT, reported scaled scores are rounded to two significant digits, e.g., on the SAT, $r_{ss}(444.97) = 440$, while $r_{ss}(445.01) = 450$. In addition, formula scores are rounded to integers, e.g., $r_{fs}(15.75) = 16$, while $r_{fs}(13.25) = 13$. In (27), $r_{fs}\ (FS_y^{-1}) = r_{fs}\ (FS_y) - 1$.

For an omit, the influence function is

$$(28) \quad IIF(SS_y, Y_k=0) = r_{ss}(s_{y'}(r_{fs}(FS_y) - r_{ss}(s_y(r_{fs}(FS_y))),$$

because $FS_{y'} = FS_y$.

For an incorrect response, there are two possible item influence functions. If $r_{ss}(FS_{y'}) = r_{ss}(FS_y)$, then equation 28 is the item influence function for an incorrect response. If, however, $r_{ss}(FS_{y'}) = r_{ss}(FS_y) + 1$, then the item influence function is described by

$$(29) \quad IIF(SS_y, Y_k=-1/(A_k-1))=r_{ss}(s_{y'}(r_{fs}(FS_y)+1) - r_{ss}(s_y(r_{fs}(FS_y))).$$

In (27)-(29), note that the item influence function depends on an item score component, namely the rounding conventions for scaled scores and formula scores ($r_{ss}$ and $r_{fs}$) and the individual's response to the deleted item ($Y_k$), as well as an equating function component ($s_y$ and $s_{y'}$), which is affected by the item's psychometric properties and the examinee's particular ability level. In (27)-(29), note that the arbitrary rounding rules can have a noticeable impact.

Parallel Analyses for "Deleting" Item and Giving Credit to All Options

The preceding analyses presume that the deleted item is also not scored. As an alternative to not scoring the deleted item, one can consider giving credit to all candidates for all options including omit. There are pros and cons associated with not scoring vs. giving credit to all options. From a psychometric viewpoint, it makes no difference as long as the new test is re-equated. From a public relations viewpoint, however, it makes a difference. On the surface, giving credit appears

more palatable than not scoring a problem item because no raw formula scores go down, i.e., those who gave the keyed response keep the same formula score, while everyone else gets a higher formula score. In contrast, not scoring the deleted item reduces the formula score of those who gave the keyed response and increases the scores of some of those who answered the item incorrectly. Re-equating yields the same scaled scores for each candidate regradless of which scoring option is used with the deleted item. As a consequence of re-scoring and re-equating, the scaled scores for those candidates who originally "got the deleted item right" will either go down or stay the same. In contrast, those who originally "got the item wrong" will retain the same scaled scores or obtain higher ones.

Since re-equating makes the scoring option (not score vs. all options correct) irrelevant, the choice between not scoring and giving everyone credit should be based on public relations considerations, which could differ depending on whether the item has a defensible key. In my opinion, when there is no defensible key, as was the case with the circles item, it is easier to explain a lower scaled score to a candidate when not scoring than when scoring all options correct. When not scoring the deleted item you can tell the candidate: "your score went down because you had 'correctly' answered an item which has been dropped from the test because it had no correct key; consequently your new raw score is one point lower than it was. In contrast, when scoring all options correct, you might have to say: "while your raw score was unchanged, everyone was given credit on the item, as a consequence the test became easier than it was and your unchanged raw score led to lower scaled score." When there are several defensible keys, however, a

stronger argument for scoring all options correct exists, as was the case with the problem Biology items.

Scoring all options correct does affect the equating function differently than the decision to not score the item. A parallel analysis of equating functions and item influence functions can be conducted for the decision to score all options correct. Rather than repeat the analyses of the two preceding sections, I will summarize how they would differ from the analysis for not scoring the item.

The item influence functions (IIF) for scoring all options correct would be

(30)
$$IIF(FS_y, Y_k=1) = -1.0 + 1.0 = 0.0$$
$$IIF(FS_y, Y_k=0) = 0.0 + 1.0 = 1.0$$
$$IIF(FS_y, Y_k=-1/(A_k-1) = 1/(A_k-1) + 1.0,$$

i.e., one point higher than the IIFs in (26) for not scoring the item.

When the item was deleted and not scored, the new test Y' was harder than the original test Y. When the item is scored all options correct, however, the new test is easier than the original test. This impacts on the equating function analysis. The ultimate effect is that the equating function for the new test is <u>always</u> lower than the equating function for the original test, i.e., it is easier to obtain a particular formula score on the new test than it was on the original test. Hence, for any given formula score, the scaled score on the original test exceeds that of the new test. The item's psychometric properties determine by how much these conversions differ, as will be illustrated in the empirical section of this report.

Summary of Analytical Decomposition

The effects of item properties, test properties, individual examinee responses and rounding rules combine to produce the item influence functions in the reported score metric described earlier. In the section on the effects on the equating function, the equating/scaling function was decomposed into its old form and new form components as depicted in (18). Then, it was shown that only the new form components were affected by item deletion. Finally, the impact of various psychometric properties, such as difficulty and discrimination, on these new form transformations was discussed and summarized in the section entitled the effects of item parameters on equating functions.

Next, the effects of individual examinee's responses to the deleted item were examined, and the item influence function was introduced. In the formula score metric, there are three item influence functions, one for each possible item score, that are independent of ability and the same across all items with the same number of response alternatives. In the reported score metric, however, item influence functions were shown to depend on examinee ability and changes in the equating function as well. In addition, the impact of rounding rules was noted. In sum, we have decomposed the item deletion effect into its various components. As a consequence, for any given examinee, we can project their new reported score from their original formula score, their response to the deleted item, and the test characteristic function for the original test. Hence, we can project the individual effects for all examinees. These individual effects culminate into effects on reported score distributions. The ultimate effect on reported score distributions depends on the ability distribution in the population of interest and

the responses of members of that population to the deleted item. The particular nature of these effects will vary from setting to setting. These points are illustrated in the next section with data from the May 1982 administration of the SAT.

## Illustrations of Item Deletion Effects

Since the circles item appeared on the May 1982 administration of the SAT, data from this administration will be used to illustrate the effects of item deletion on reported score distributions and equating functions. These data provide answers to several interesting 'what if' questions. We can examine the effects associated with deleting items that have certain psychometric characteristics, enabling us to answer what would have occurred if the circles item had different psychometric characteristics than it had. In particular, six items were selected for deletion. Item MD was the most difficult item on the test ($b_{MD} = 2.87$); item LD was the least difficult ($b_{LD} = -2.48$). In addition to these two extremes on the difficulty continuum, a highly discriminating item ($a_{AC} = 1.73$) with a high lower asymptote ($c_{AC} = .27$), a highly discriminating item ($a_{Ac} = 1.48$) with a low lower asymptote ($c_{Ac} = .05$), a poorly discriminating item ($a_{aC} = .55$) with a high lower asymptote ($c_{aC} = .27$), and a poorly discriminating item ($a_{ac} = .52$) with a low lower asymptote ($c_{ac} = .03$) were selected for deletion. These four items are denoted by AC, Ac, aC, and ac, respectively. Table 2 contains the item parameters for the six items and the circles items.

## Item Deletion Simulation Procedures

The effects of item deletion were studied in the following manner. Each of the six items was deleted from the 60-item total test containing the circles item and the 59-item test that excluded the circles item. As a consequence, six separate 59-item tests and six separate 58-item tests were simulated. Each of these 12 tests will be compared to the full 60-item test.

Table 2

Item Parameter Estimates for Deleted Items

| Item | a(discrimination) | b(difficulty) | c(lower asymptote) |
|---|---|---|---|
| MD | .83 | 2.88 | .15 |
| LD | .53 | -2.48 | .08 |
| AC | 1.73 | .94 | .27 |
| Ac | 1.48 | 2.16 | .05 |
| aC | .55 | 1.61 | .27 |
| ac | .52 | .03 | .03 |
| Circles | 1.30 | 1.10 | .24 |

One type of comparison focuses on equating/scaling functions and differences induced by deletion of items with certain properties. Effects on both rounded and unrounded converted scores will be assessed. Difference plots are used to examine these effects.

In addition, effects of item deletion on examinee formula scores are assessed. This step necessitated rescoring for a representative sample of 45,579 examinees, the same set of examinees used to assess the effect of deleting the circles item. Differences among rounded and unrounded scaled scores are summarized.

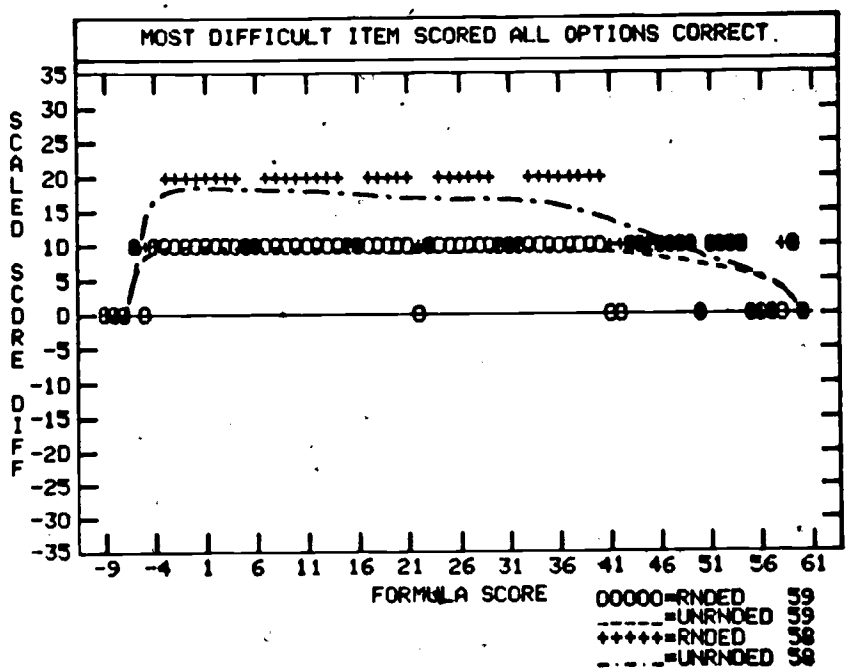
Results

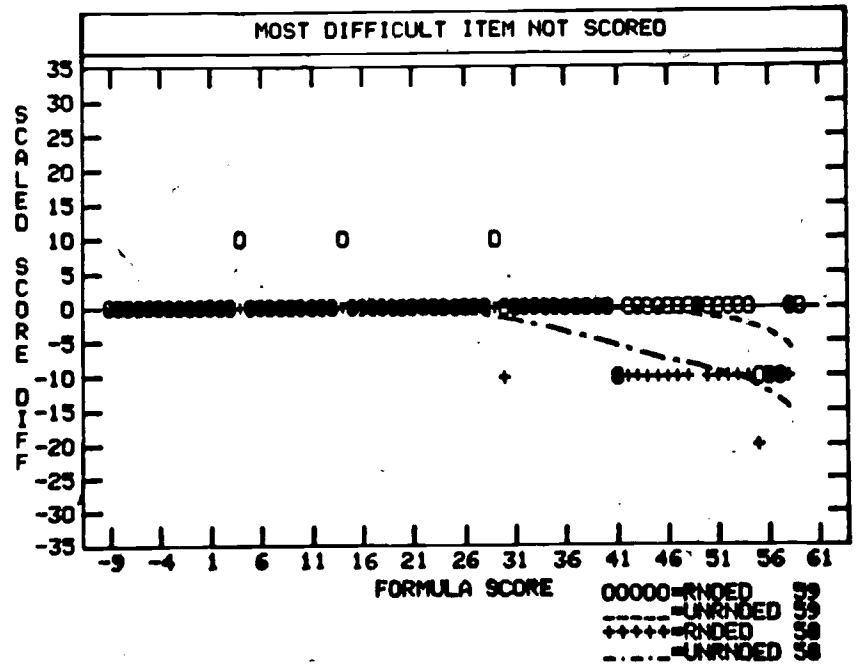
Equating/scaling functions. Figure 1 depicts the effects on the equating/scaling functions produced by deleting the most difficult item by itself (indicated by the 00000 for rounded scores and the ----- for unrounded scores), and with the circles item ( indicated by +++++ for the rounded scores and -.-.- for the unrounded scores). In this figure, and all subsequent figures, the rounded differences are discrete, taking on one of the five possible values: -20, -10, 0, +10, +20. These rounded differences are obtained by subtracting, for a given formula score, the re-equated rounded scaled score for the 59-item test (or the 58-item test) produced by deleting the most difficult item by itself (or by deletion of that item and the circles item as well) from the rounded scaled score for the original 60-item test. In contrast, the "unrounded" differences are differences of unrounded scaled score conversions that are rounded to the units place.

This figure and subsequent figures contain an upper and lower panel. Both panels contain four difference plots, two rounded and two unrounded, two for deleting a single item (in Figure 1, item MD) and two

Figure 1

Differences in Unrounded and Rounded Equating/Scaling Functions for the 59-item (and 58-item) Tests Produced by Deletion of the Most Difficult (MD) Item (and the Circles Item) (Original Equating - Re-equating)



for deleting that item plus the circles item. In the upper panel, the differences between the original conversions and those obtained by scoring all options correct and re-equating are plotted. The lower panel contains the differences between the original conversions and those obtained by not scoring and re-equating.

Examination of the bottom panel of Figure 1 reveals that, at the unrounded scaled score level, deletion of the most difficult item has a negligible effect on the equating function for not scoring the item at formula scores less than 50. In fact, the only unrounded difference which exceeds -5.0 in magnitude occurs at a formula score of 59. There is hardly any effect evident on rounded scaled score level either. The rounded differences of +10 that occur for the three formula scores below 50 are actually negligible unrounded differences: 294.9656 vs. 295.5527 at a formula score of 4; 384.6948 vs. 385.2297 at a formula score of 14; 514.9980 vs. 515.3547 at a formula score of 29. The -10 at a formula score of 41 is also the result of a negligible difference, 625.1816 vs. 624.9834. In short, deletion of the most difficult item has very little effect on the equating function for not scoring the item because the shortened test is almost as easy as the longer test.

When the circles item is also deleted to produce a 58-item test, a greater effect occurs. Note, in the bottom panel, that for formula scores of 41 and greater, the rounded conversions for the 58-item test exceed those of the full 60-item test at all but one formula score, 49, where rounding produces an equal scaled score of 700 because the 60-item conversion is 695.2595 while the 58-item conversion is 703.8240. The plot of the unrounded conversion for the 58-item test, denoted by -.-.-, indicates a noticeable downward slope that begins at about a formula

score of 30. At a formula score of 40, the difference in unrounded conversions begins to systematically exceed -5.0 in magnitude. By 53, this difference exceeds -10 in magnitude. Clearly the additional deletion of the circles item has a much greater effect on the equating function for not scoring than did deletion of just the most difficult item.

The upper panel of Figure 1 provides a sharp contrast to the lower panel of the same figure. Here the differences associated with scoring the item all options correct are depicted. At every formula score, the original conversions exceed the new conversions that result from re-equating. Obviously, deletion of the most difficult item and scoring all options correct has a substantial effect on the equating function. In short, the decision to score a difficult item all options correct makes the new test noticeably easier than the old test. Deletion of the circles item as well merely increases the differential in difficulty.

The contrast between the panels in Figure 1 can illuminate discussion about the necessity to re-equate tests after item deletion. Recall that it was stated earlier that re-equating makes it psychometrically irrelevant whether the deleted item is scored all options correct or deleted. Figure 1, as well as all subsequent figures, can be used to demonstrate that how one scores the item becomes very important when the decision is made not to re-equate the test after item deletion. The upper panel reflects the differences in scaled scores for a given formula score that would be obtained if the original conversion were used instead of the conversion produced by re-equating the new test. Except at very low and very high scaled scores, and a few points in between where rounding impacts on the results, use of the

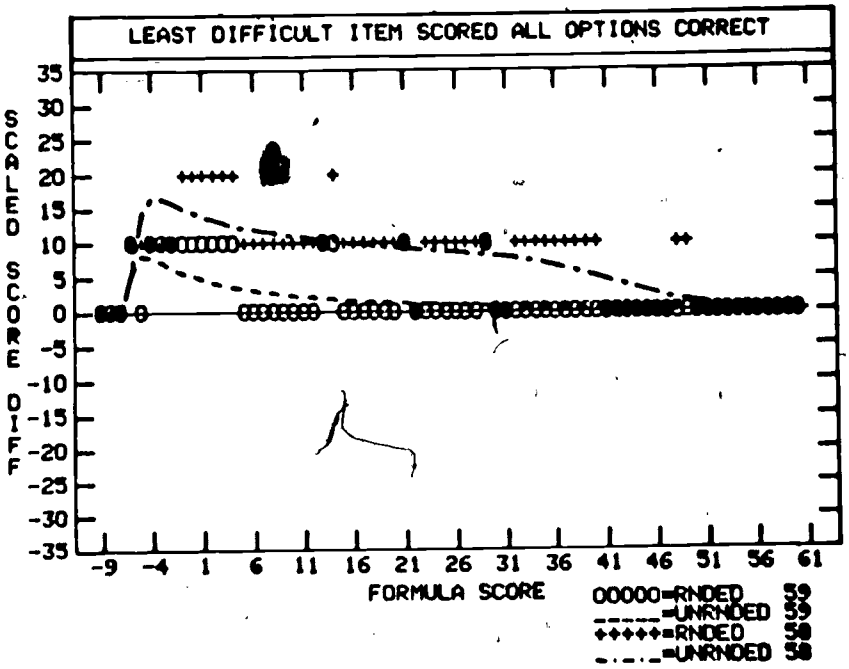
original conversion with the adjustment of formula scores resulting from scoring the flawed item all options correct would yield scaled scores that were 10 to 20 points higher than what the re-equating would suggest are appropriate. In short, there would be a systematic positive bias introduced into the scaled scores.

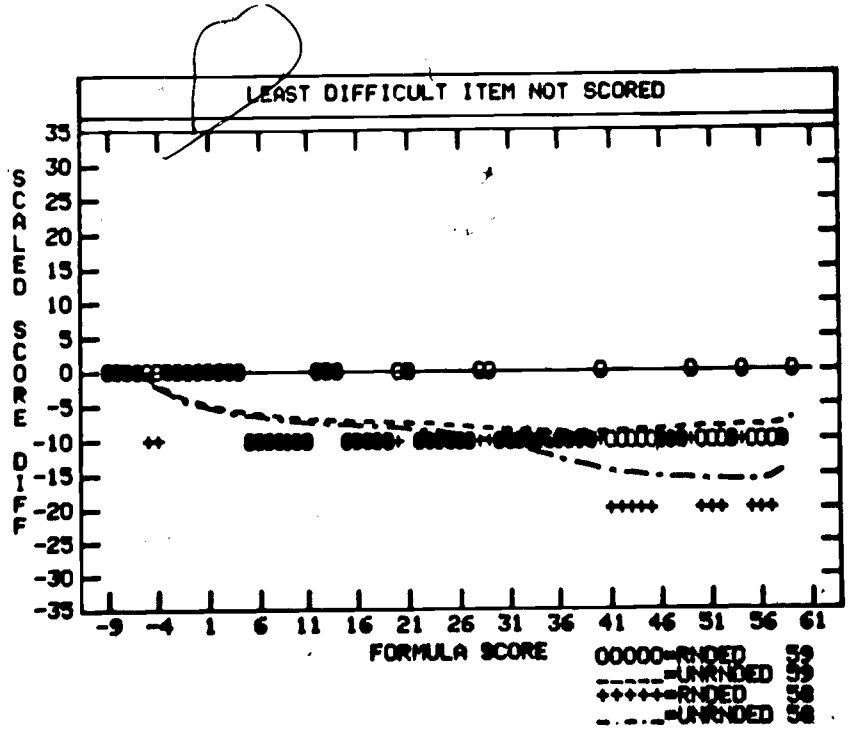
In contrast, the bottom panel in Figure 1 reveals that use of the original conversion with the adjustment of formula scores resulting from not scoring the flawed item would yield scaled scores that were equal to the scaled scores produced by re-equating except at high formula scores where the re-equated scores would exceed the original conversions (reflected by negative differences in the bottom panel). In short, the decision to not re-equate, if it were made, would make the scoring decision, not score vs. score all options correct, very important: Giving credit tends to make all scores higher than the re-equated scores, while not scoring tends to make all scores lower than the same re-equated scores. In addition, not re-equating allows the psychometric properties of the deleted items to impact on the nature of these differences, as will be seen in subsequent figures.

Figure 2 provides a striking contrast to Figure 1. Here, the effects of deleting the least difficult item are depicted. Note that all four difference plots in the upper panel reflect positive differences, while all four difference plots in the bottom panel indicate negative differences, which implies that all formula scores above 6 are easier to obtain on the 60-item test than on either the 58-item or 59-item tests under not scoring, and harder on the 60-item test than they are under scoring all options correct. In short, all four difference curves in the bottom panel are consistent with the fact

Figure 2

Differences in Unrounded and Rounded Equating/Scaling Functions for the 59-item (and 58-item) Tests Produced by Deletion of the Least Difficult (LD) Item (and the Circles Item) (Original Equating - Re-equating)



that deleting the easiest item provides a shorter more difficult test, while the upper panel indicates that scoring all options correct makes the test easier. By a formula score of 2.0, the unrounded conversions in the bottom panel for both the 58-item and 59-item tests exceed -5.0 in magnitude. Note that at approximately a formula score of 36, the unrounded difference plot in the bottom panel for the 59-item test has leveled off, while that of the 58-item test continues to slope downward. This latter effect, separation of the two unrounded difference plots at around 30, which also occurs in the upper panel, also occurred in Figure 1 and reflects deletion of the circles item.

Examination of Figures 1 and 2 provide insight into what might happen if one decided not to score a flawed item and not to re-equate, which is unsound from a psychometric vantage point. If the flawed item were hard, as is the case in Figure 1, the converted scores would agree up to formula scores in the high fifties (see lower panel). In fact, use of the original conversion would avoid the three roundoff problems at 4, 14, and 29 that were discussed earlier. Deletion of the second item, however, introduces consistent differences above 41 that would be ignored if the original conversion were used.

While use of the original 60-item conversion on the 59-item test resulting from deleting and not scoring the most difficult item might not affect scores much, deleting and not scoring the easy item is another story. Here, use of the original conversion with the 59-item and 58-item tests, whose equating function differences are depicted in the lower panel of Figure 2, would yield substantially lower converted scores than would re-equating.

Comparison of the upper panels of Figures 1 and 2 provides us with insight into what would happen if the decision to score all options correct were accompanied by a decision to not re-equate. As noted in the discussion of Figure 1, scoring the most difficult item all options correct has a profound impact on the equating function compared to deleting and not scoring that same item. In contrast, the upper panel in Figure 2 reveals that scoring the easiest item all options correct only affects low formula scores, those below .6, while the lower panel reveals that not scoring the easiest item affects almost all scores above 6. In short, the decision not to re-equate allows both the flawed item's psychometric properties and the scoring decision to impact significantly on reported scores. When there is no re-equating, scoring the most difficult item all options correct produces the largest scaled scores, followed by scoring the easiest item all options correct. Also under no re-equating, not scoring the easiest item produces the lowest scaled scores, followed by deleting and not scoring the hardest item. Hence, not re-equating allows the properties of the deleted problem item and the scoring decision to interact and impact significantly on reported scores. In contrast, re-equating makes the scoring decision irrelevant and, as will be seen, mitigates the effects of the deleted item's psychometric properties.

In the four remaining figures, attention will be paid to the lower panels only where discussion will focus on the effects of deletion on equating and scaling functions. Since re-equating is clearly desirable, re-equating vs. not re-equating will not be discussed explicitly with these figures. The reader, however, can compare the upper and lower panels of these subsequent figures to project the effects of re-equating

vs. not re-equating. The re-equating vs. not re-equating issue will be revisited explicitly when the effects on score distributions are addressed.

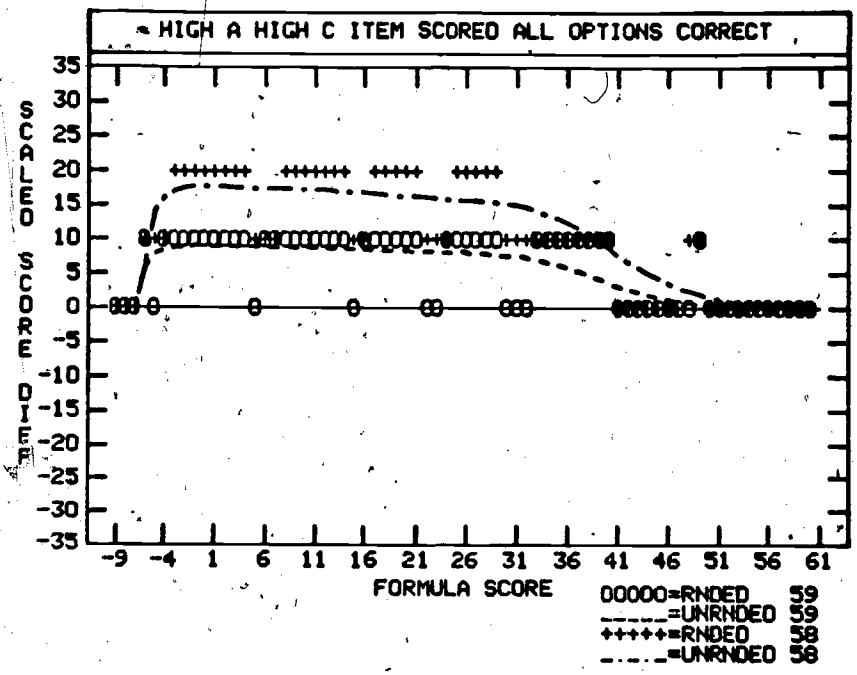
Figure 3 depicts the effects of deleting a highly discriminating item with a high lower asymptote, item AC. The first thing to note in the figure is that all four difference curves in the lower panel are below zero at all formula scores. This follows from the fact that $c_{AC} = .27$, i.e., the lower asymptote exceeds chance level performance. Hence, the shortened test is harder than the longer test at all formula score levels, a point noted earlier in the analytical analysis. Another aspect worth noting is that for formula scores below 20, the unrounded differences are close to zero, a fact that is characteristic of highly discriminating items. It should be noted that the circles item was highly discriminating also, $a_c = 1.30$. Observe that the difference curves begin to descend noticeably and level off quickly also, a characteristic of highly discriminating items.

Figure 4 depicts the effects of deleting another highly discriminating item, Ac. Item Ac, however, has a low lower asymptote, $c_{Ac} = .05$, and a high difficulty, $b_{Ac} = 2.16$ to accompany its high $a_{Ac} = 1.48$. As a consequence, the shortened tests are easier than the 60-item test for sizeable portions of the formula score range: up to a formula score of 28 for the 58-item test, and up to a formula score of 45 for the 59-item test. As in Figure 3, sharp declines and leveling off occur in the difference curves. The nine +10 differences observed for the rounded 59-item conversion are clearly rounding artifacts.

The sharp declines and abrupt leveling off observed in the bottom panels of the last two figures are not replicated in the next two

Figure 3

Differences in Unrounded and Rounded Equating/Scaling Functions for the 59-item (and 58-item) Tests Produced by Deletion of the High A - High C (AC) Item (and the Circles Item) (Original Equating - Re-equating)



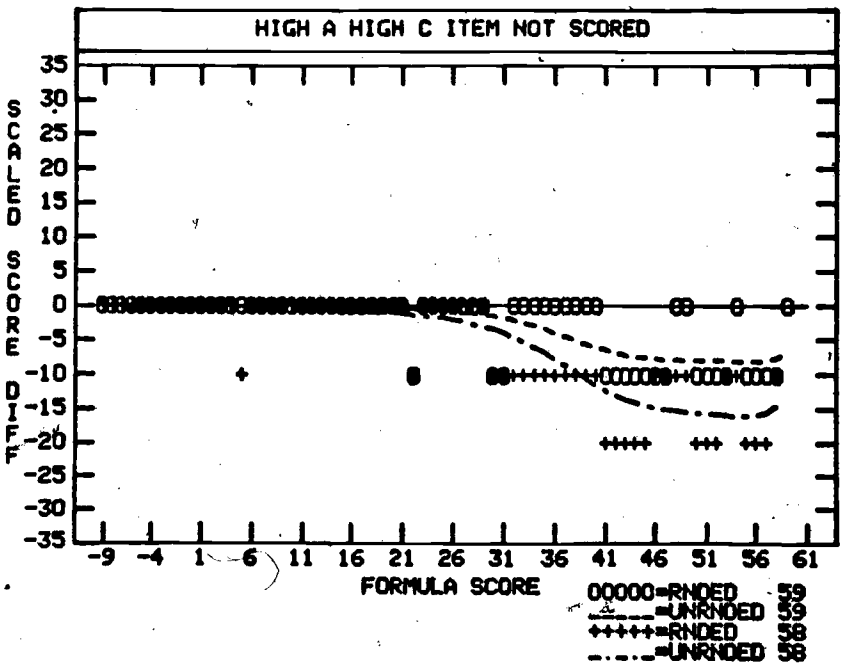
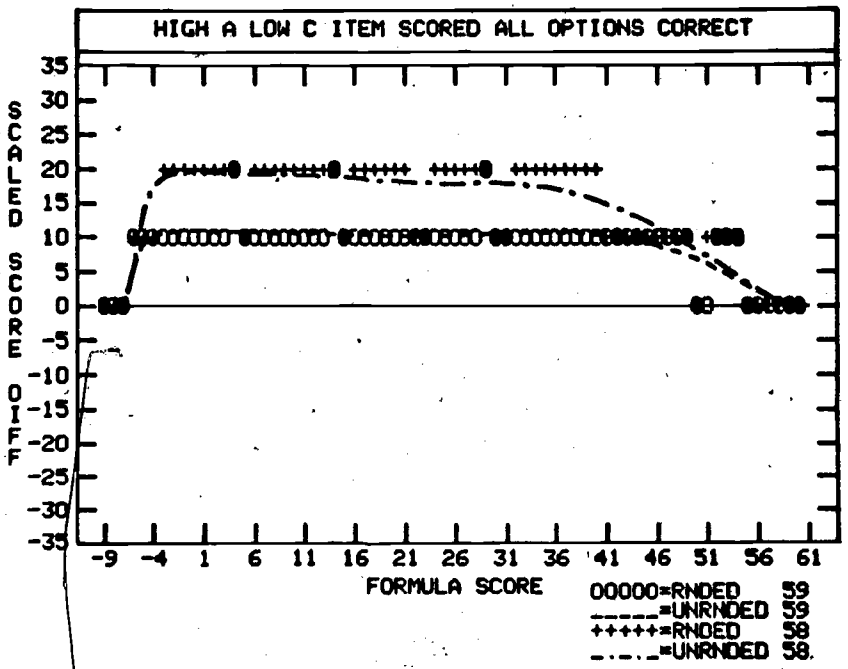
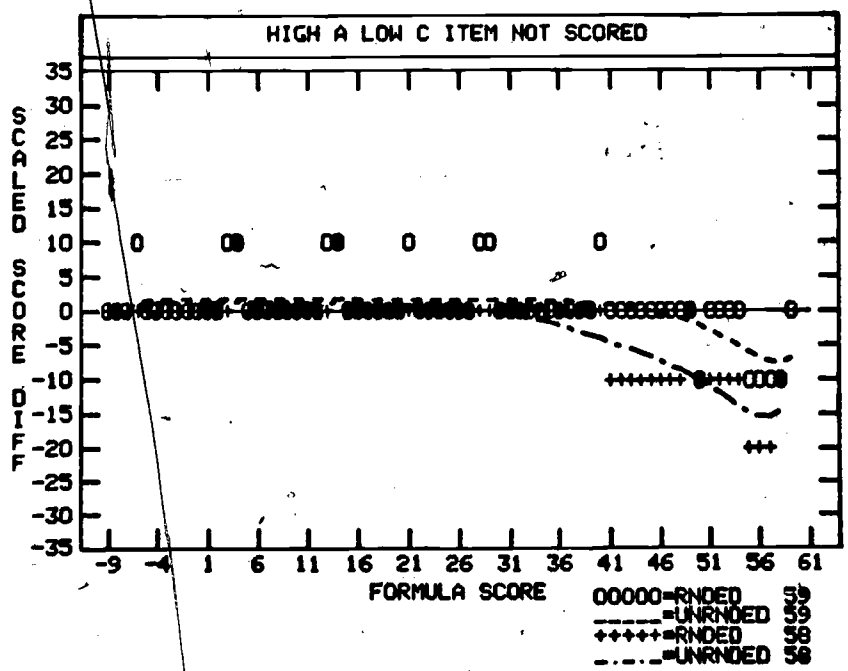



Figure 4

Differences in Unrounded and Rounded Equating/Scaling Functions for the 59-item (and 58-item) Tests Produced by Deletion of the High A - Low C (Ac) Item (and the Circles Item) (Original Equating - Re-equating)



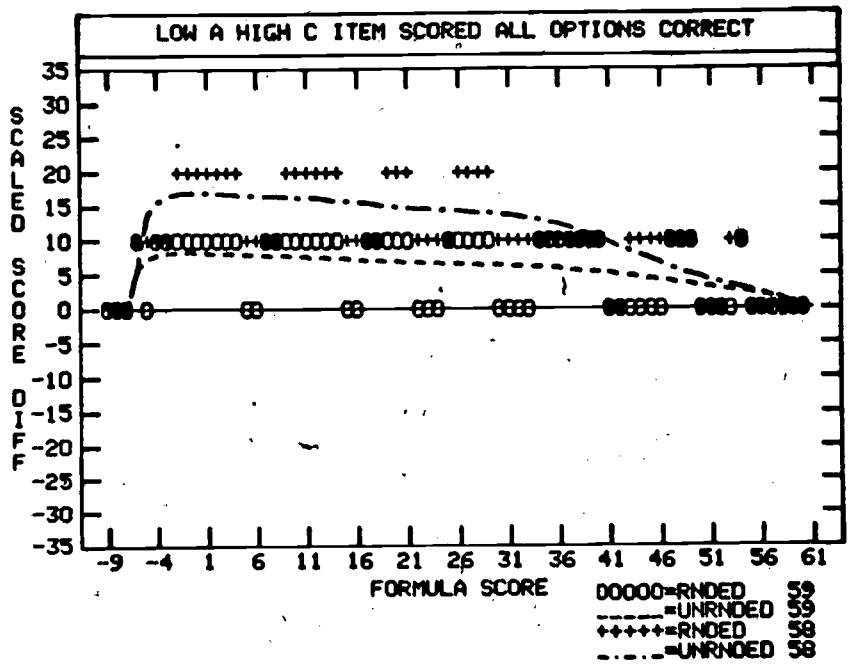
figures, which depict the effects on equating functions of deleting items with low discriminating power. Figure 5 depicts the effects of dropping the aC item ($a_{aC}$ = .55, $b_{aC}$ = 1.61, $c_{aC}$ = .27), while deletion of the ac item ($a_{ac}$ = .52, $b_{ac}$ = .03, $c_{ac}$ = .03) is depicted in Figure 6. In both these figures, the declines in the unrounded difference curves are gradual. In the latter figure, the decline starts sooner because it is an easier item, which also accounts for the larger number of -20 rounded differences evident in this figure. Note that in contrast to Figure 5, unrounded differences in Figure 6 can be positive since the lower asymptote is nearly zero.

Figures 1-6 depict the effects of deleting various items from the full 60-item test on the equating functions. Several effects were noted. The magnitude of the c-parameter constricts the range of the effect. Sufficiently high c-parameters preclude the occurence of negative differences. The location of the b-parameter determines where the effect occurs along the formula score range, while the a-parameter affects the sharpness and duration of the effect. In addition to the effects of these item parameters, we observed the effects of the rounding rules. In fact, in Figures 1 and 4, the rounding effects tended to be the dominant effects.

Finally, the discussion of Figures 1 and 2 made it clear that from a psychometric viewpoint not re-equating is less desirable than re-equating since not re-equating allows the deleted item scoring decision and the deleted item's psychometric properties to have significant impact on the converted scores. In contrast, re-equating makes the scoring decision (all options correct vs. not score) irrelevant and attempts to mitigate the impact of the item's

Figure 5

Differences in Unrounded and Rounded Equating/Scaling Functions for the 59-item (and 58-item) Tests Produced by Deletion of the Low A - High C (aC) Item (and the Circles Item) (Original Equating - Re-equating)



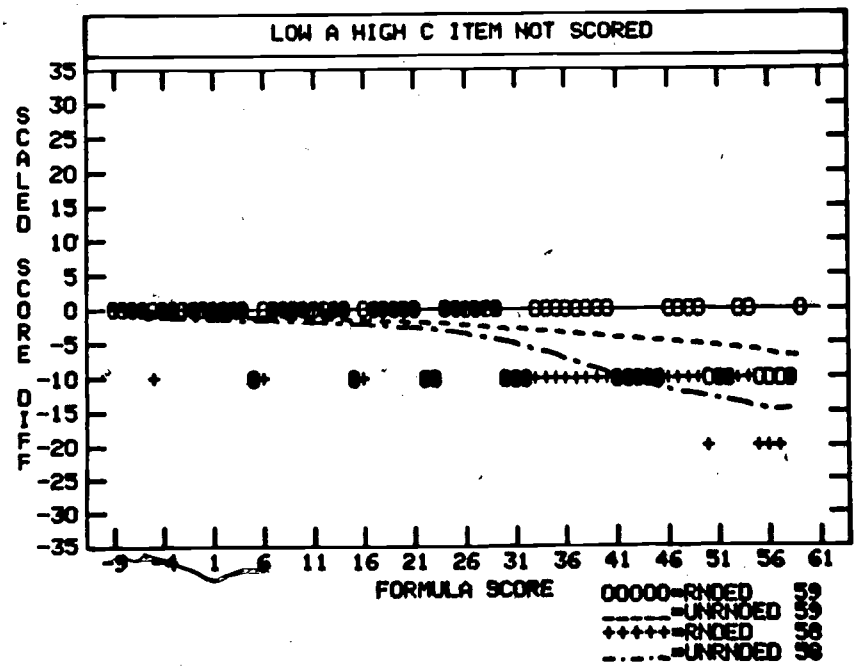
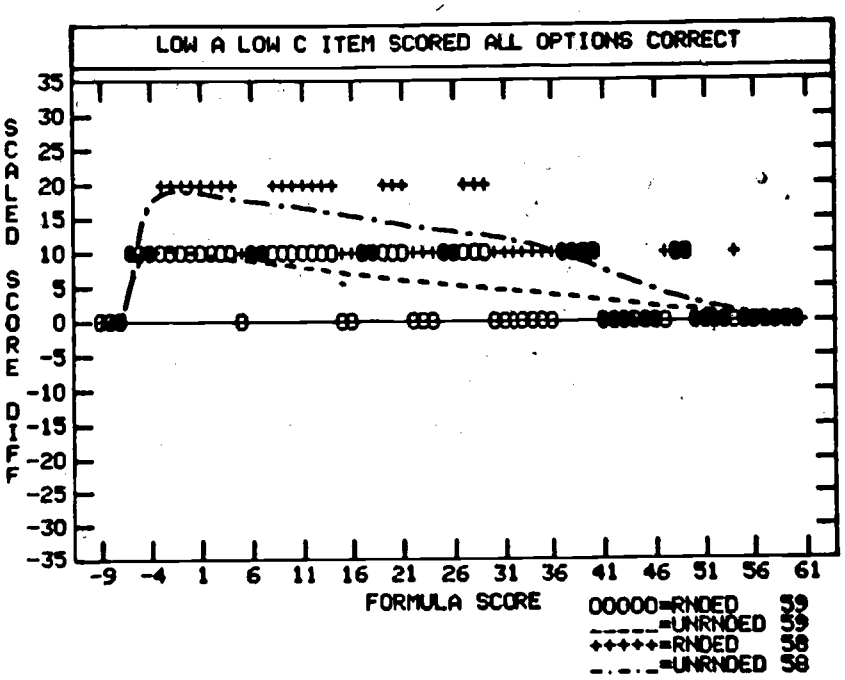
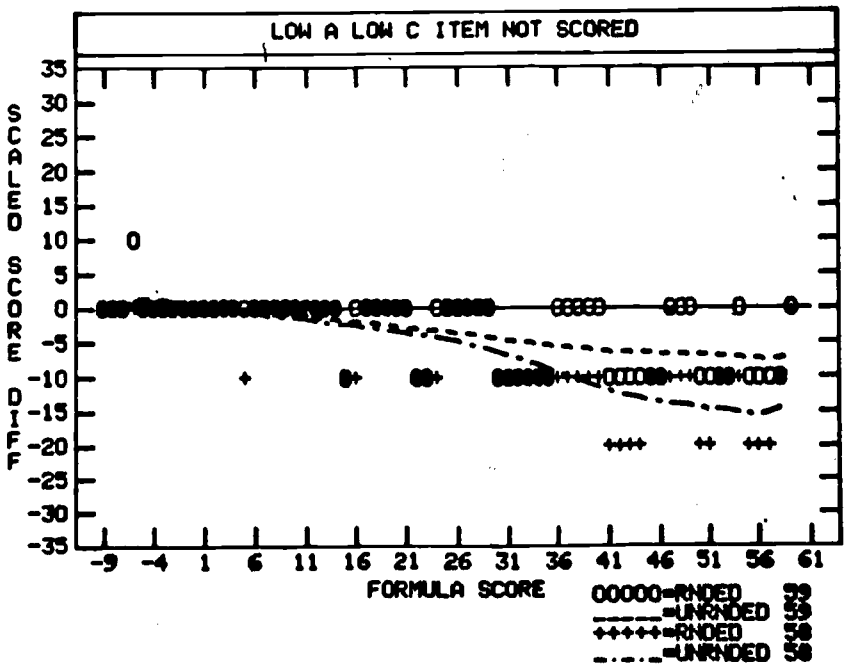



Figure 6

Differences in Unrounded and Rounded Equating/Scaling Functions for the 59-item (and 58-item) Tests Produced by Deletion of the Low A - Low C (ac) Item (and the Circles Item) (Original Equating - Re-equating)



psychometric properties. The results presented in the next two sections clarify these points.

Equated score distributions after re-equating. In addition to altering the equating function, deletion of an item can affect the formula score of an individual. There are different item influence functions for each possible response to the item: correct, incorrect or omit. In addition, the scoring decison has an impact. Here, we limit the discussion to not scoring the deleted item. Those who answered the deleted item correctly lose one formula score point, those who omitted the deleted item keep the same formula score, and those who answered incorrectly either retain the same formula score or gain one formula score point. The effect on the equating function, illustrated in Figures 1-6, and the effects on formula scores combine to impact on scaled scores. For a group of individuals, these effects translate into a distribution of difference scores for that group. Tables 3-8 summarize these distributions of differences resulting from deleting the seven items under study, and correspond to Figures 1-6, respectively. Each table contains nine columns, the first of which is scaled score difference, ranging from 30 to -20. The next four columns contain the absolute and relative frequencies for unrounded and rounded differences between individual's scaled scores on the 59-item test resulting from item deletion and their original score on the 60-item test. A positive difference indicates the new score exceeds the original score. The last four columns present the same data for the 58-item test produced by deleting the circles item from the 59-item test. At the bottom of each table are means, standard deviations, sample sizes and modes.

Table 3 contains the result for the most difficult item; item MD. Figure 1 is the corresponding figure depicting equating function differences. Note that over 71% of the rounded differences equal zero for the 59-item test. Of the remaining 29.9%, slightly more than half are -10. In the unrounded difference distributions for the 59-item test, over 73% of the distribution is in the range -2 to 4. In addition, there are two other peeks at 8 to 10 and -10 to -5. Note also that use of the rounding rule tends to spread out differences, e.g. despite the fact that the maximum unrounded difference is 13, 19 people receive +20 over their original scaled score. This spreading is reflected in larger standard deviation for rounded differences. Note that rounding also pushes the mean farther from its original value.

Deleting the circles item from the 59-item test has the expected effect of spreading scores out even more. The trimodal nature of the unrounded distribution is less apparent than it was before deleting the circles item. In addition, differences of -20 are produced for rounded scores, where only 52.7% of the difference scores are zero. The summary statistics at the bottom of the table are interesting. As expected, deletion of the second item increases the spread of both rounded and unrounded difference scores. More importantly, the mean difference for rounded scores is pushed even further from the original mean, while the mean of the unrounded scores remains relatively close to the original mean. The impact of rounding that is evident here will be evident in subsequent tables and will tend to be a relatively major factor.

Table 4 portrays the distributional effects produced by deletion of the easiest item (LD). The trimodality evident with deletion of the most difficult item is missing in the distribution of unrounded

Table 3

Distributions of Differences Between Re-equated(RE) and Original(OS) Rounded and Unrounded Scaled Scores Associated with Deletion of Item MD (RE - OS)

| Scaled Score Difference | Item MD 59-item unrounded | | Item MD 59-item rounded | | Item MD 58-item unrounded | | Item MD 58-item rounded | |
|---|---|---|---|---|---|---|---|---|
| 30.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 29.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 28.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 27.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 26.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 25.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 24.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 23.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 22.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 21.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 20.000 | 0 | 0.0 | 19 | 0.0 | 0 | 0.0 | 362 | 0.8 |
| 19.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 18.000 | 0 | 0.0 | | | 25 | 0.1 | | |
| 17.000 | 0 | 0.0 | | | 82 | 0.2 | | |
| 16.000 | 0 | 0.0 | | | 104 | 0.2 | | |
| 15.000 | 0 | 0.0 | | | 241 | 0.5 | | |
| 14.000 | 0 | 0.0 | | | 231 | 0.5 | | |
| 13.000 | 4 | 0.0 | | | 459 | 1.0 | | |
| 12.000 | 19 | 0.0 | | | 443 | 1.0 | | |
| 11.000 | 6 | 0.0 | | | 595 | 1.3 | | |
| 10.000 | 308 | 0.7 | 6247 | 13.7 | 1163 | 2.5 | 9854 | 21.5 |
| 9.000 | 2527 | 5.5 | | | 5570 | 12.2 | | |
| 8.000 | 3673 | 8.0 | | | 317 | 0.7 | | |
| 7.000 | 9 | 0.0 | | | 265 | 0.6 | | |
| 6.000 | 15 | 0.0 | | | 713 | 1.6 | | |
| 5.000 | 0 | 0.0 | | | 597 | 1.3 | | |
| 4.000 | 21 | 0.0 | | | 719 | 1.6 | | |
| 3.000 | 80 | 0.2 | | | 1466 | 3.2 | | |
| 2.000 | 320 | 0.7 | | | 1851 | 4.0 | | |
| 1.000 | 1497 | 3.3 | | | 4663 | 10.2 | | |
| 0.0 | 22306 | 48.7 | 32514 | 71.1 | 10912 | 23.8 | 24098 | 52.7 |
| -1.000 | 9357 | 20.4 | | | 317 | 0.7 | | |
| -2.000 | 41 | 0.1 | | | 671 | 1.5 | | |
| -3.000 | 0 | 0.0 | | | 532 | 1.2 | | |
| -4.000 | 25 | 0.1 | | | 1130 | 2.5 | | |
| -5.000 | 112 | 0.2 | | | 779 | 1.7 | | |
| -6.000 | 218 | 0.5 | | | 917 | 2.0 | | |
| -7.000 | 369 | 0.8 | | | 1777 | 3.9 | | |
| -8.000 | 464 | 1.0 | | | 4670 | 10.2 | | |
| -9.000 | 3828 | 8.4 | | | 2736 | 6.0 | | |
| -10.000 | 560 | 1.2 | 6979 | 15.3 | 160 | 0.3 | 10514 | 23.0 |
| -11.000 | 0 | 0.0 | | | 179 | 0.4 | | |
| -12.000 | 0 | 0.0 | | | 60 | 0.1 | | |
| -13.000 | 0 | 0.0 | | | 145 | 0.3 | | |
| -14.000 | 0 | 0.0 | | | 144 | 0.3 | | |
| -15.000 | 0 | 0.0 | | | 200 | 0.4 | | |
| -16.000 | 0 | 0.0 | | | 202 | 0.4 | | |
| -17.000 | 0 | 0.0 | | | 520 | 1.1 | | |
| -18.000 | 0 | 0.0 | | | 197 | 0.4 | | |
| -19.000 | 0 | 0.0 | | | 7 | 0.0 | | |
| -20.000 | 0 | 0.0 | | | 0 | 0.0 | 931 | 2.0 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | .01 | | -.15 | | -.03 | | -.39 | |
| S.D. | 4.47 | | 5.39 | | 6.85 | | 7.46 | |

Table 4

Distributions of Differences Between Re-equated(RE) and Original(OS) Rounded and Unrounded Scaled Scores Associated with Deletion of Item LD (RE - OS)

| Scaled Score Difference | Item LD 59-item unrounded | | Item LD 59-item rounded | | Item LD 58-item unrounded | | Item LD 58-item rounded | |
|---|---|---|---|---|---|---|---|---|
| 30.000 | 0 | 0.0 | | | 0 | 0.0 | 10 | 0.0 |
| 29.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 28.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 27.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 26.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 25.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 24.000 | 0 | 0.0 | | | 13 | 0.0 | | |
| 23.000 | 0 | 0.0 | | | 27 | 0.1 | | |
| 22.000 | 0 | 0.0 | | | 14 | 0.0 | | |
| 21.000 | 0 | 0.0 | | | 45 | 0.1 | | |
| 20.000 | 0 | 0.0 | 610 | 1.3 | 26 | 0.1 | 1080 | 2.4 |
| 19.000 | 0 | 0.0 | | | 38 | 0.1 | | |
| 18.000 | 105 | 0.2 | | | 109 | 0.2 | | |
| 17.000 | 126 | 0.3 | | | 376 | 0.8 | | |
| 16.000 | 534 | 1.2 | | | 553 | 1.2 | | |
| 15.000 | 178 | 0.4 | | | 367 | 0.8 | | |
| 14.000 | 116 | 0.3 | | | 239 | 0.5 | | |
| 13.000 | 6 | 0.0 | | | 345 | 0.8 | | |
| 12.000 | 20 | 0.0 | | | 245 | 0.5 | | |
| 11.000 | 0 | 0.0 | | | 466 | 1.0 | | |
| 10.000 | 10 | 0.0 | 2819 | 6.2 | 533 | 1.2 | 7791 | 17.0 |
| 9.000 | 293 | 0.6 | | | 971 | 2.1 | | |
| 8.000 | 642 | 1.4 | | | 2037 | 4.5 | | |
| 7.000 | 1532 | 3.3 | | | 1481 | 3.2 | | |
| 6.000 | 702 | 1.5 | | | 1058 | 2.3 | | |
| 5.000 | 303 | 0.7 | | | 1016 | 2.2 | | |
| 4.000 | 181 | 0.4 | | | 784 | 1.7 | | |
| 3.000 | 55 | 0.1 | | | 1487 | 3.2 | | |
| 2.000 | 22 | 0.0 | | | 1767 | 3.9 | | |
| 1.000 | 21 | 0.0 | | | 2741 | 6.0 | | |
| 0.0 | 9573 | 20.9 | 37088 | 81.1 | 5804 | 12.7 | 25055 | 54.8 |
| -1.000 | 21956 | 48.0 | | | 5938 | 13.0 | | |
| -2.000 | 6084 | 13.3 | | | 3173 | 6.9 | | |
| -3.000 | 2122 | 4.6 | | | 1826 | 4.0 | | |
| -4.000 | 645 | 1.4 | | | 1212 | 2.6 | | |
| -5.000 | 315 | 0.7 | | | 1434 | 3.1 | | |
| -6.000 | 145 | 0.3 | | | 998 | 2.2 | | |
| -7.000 | 31 | 0.1 | | | 1330 | 2.9 | | |
| -8.000 | 42 | 0.1 | | | 2924 | 6.4 | | |
| -9.000 | 0 | 0.0 | | | 2424 | 5.3 | | |
| -10.000 | 0 | 0.0 | 5242 | 11.5 | 1066 | 2.3 | 11484 | 25.1 |
| -11.000 | 0 | 0.0 | | | 502 | 1.1 | | |
| -12.000 | 0 | 0.0 | | | 200 | 0.4 | | |
| -13.000 | 0 | 0.0 | | | 93 | 0.2 | | |
| -14.000 | 0 | 0.0 | | | 31 | 0.1 | | |
| -15.000 | 0 | 0.0 | | | 52 | 0.1 | | |
| -16.000 | 0 | 0.0 | | | 14 | 0.0 | | |
| -17.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| -18.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| -19.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| -20.000 | 0 | 0.0 | | | 0 | 0.0 | 339 | 0.7 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | - .08 | | - .26 | | - .08 | | - .48 | |
| S.D. | 3.45 | | 4.78 | | 6.41 | | 7.38 | |

differences in this table. Deletion of this item produces even more zero differences than deletion of the hard item. The mean differences, however are more negative for this item. When Tables 3 and 4 are placed side by side, it becomes apparent that rounding and the fact that an item is being deleted have bigger effects than do the psychometric properties of the item. These facts are most evident in the summary statistics.

Tables 5-8 contain the results summarizing the distributional effects produced by deleting the AC, Ac, aC, and ac items, respectively. Table 6 is the most unique of these four tables. The other three tables reveal that deletion of the AC, aC or ac item has more effect on scaled scores than deletion of either the most difficult or least difficult item, a result consistent with the psychometric expectation that deletion of items of middle difficulty will have a greater effect on score distributions than deletion of very hard or very easy items. The standard deviations reported in these tables summarize this effect. Note, however, that the fact that items are to be deleted and scores rounded tend to have sizeable effects as well. In conjunction with Tables 1 and 2, these three tables provide evidence for the complex interaction of rounding rules, the act of deletion, and psychometric properties. In all six tables, the standard deviation of rounded differences for the 58-item test exceeds that of the unrounded differences. The same ordered relationship holds for the 59-item test. This consistent ordering reflects the impact of rounding. The act of item deletion accounts for the fact that the 58-item standard deviation exceed the 59-item standard deviations, which exceed zero. Finally, the impact of psychometric characteristics is evident in the differences across tables in the magnitudes of the standard deviations.

Table 5

Distributions of Differences Between Re-equated(RE) and Original(OS) Rounded and Unrounded Scaled Scores Associated with Deletion of Item AC (RE - OS)

| Scaled Score Difference | Item AC 59-item unrounded | | Item AC 59-item rounded | | Item AC 58-item unrounded | | Item AC 58-item rounded | |
|---|---|---|---|---|---|---|---|---|
| 30.000 | 0 | 0.0 | | | 0 | 0.0 | 90 | 0.2 |
| 29.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 28.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 27.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 26.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 25.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 24.000 | 0 | 0.0 | | | 25 | 0.1 | | |
| 23.000 | 0 | 0.0 | | | 57 | 0.1 | | |
| 22.000 | 0 | 0.0 | | | 32 | 0.1 | | |
| 21.000 | 0 | 0.0 | | | 88 | 0.2 | | |
| 20.000 | 0 | 0.0 | 466 | 1.0 | 80 | 0.2 | 1957 | 4.3 |
| 19.000 | 0 | 0.0 | | | 107 | 0.2 | | |
| 18.000 | 0 | 0.0 | | | 103 | 0.2 | | |
| 17.000 | 0 | 0.0 | | | 106 | 0.2 | | |
| 16.000 | 326 | 0.7 | | | 207 | 0.5 | | |
| 15.000 | 223 | 0.5 | | | 332 | 0.7 | | |
| 14.000 | 234 | 0.5 | | | 636 | 1.4 | | |
| 13.000 | 393 | 0.9 | | | 299 | 0.7 | | |
| 12.000 | 460 | 1.0 | | | 713 | 1.6 | | |
| 11.000 | 451 | 1.0 | | | 1101 | 2.4 | | |
| 10.000 | 1142 | 2.5 | 10096 | 22.1 | 6347 | 13.9 | 12429 | 27.2 |
| 9.000 | 5242 | 11.5 | | | 221 | 0.5 | | |
| 8.000 | 264 | 0.6 | | | 254 | 0.6 | | |
| 7.000 | 241 | 0.5 | | | 468 | 1.0 | | |
| 6.000 | 352 | 0.8 | | | 603 | 1.3 | | |
| 5.000 | 403 | 0.9 | | | 1031 | 2.3 | | |
| 4.000 | 614 | 1.3 | | | 585 | 1.3 | | |
| 3.000 | 698 | 1.5 | | | 1526 | 3.3 | | |
| 2.000 | 1986 | 4.3 | | | 2289 | 5.0 | | |
| 1.000 | 4284 | 9.4 | | | 7844 | 17.1 | | |
| 0.0 | 10815 | 23.6 | 23793 | 52.0 | 782 | 1.7 | 19839 | 43.4 |
| -1.000 | 1864 | 4.1 | | | 1456 | 3.2 | | |
| -2.000 | 890 | 1.9 | | | 1164 | 2.5 | | |
| -3.000 | 1124 | 2.5 | | | 763 | 1.7 | | |
| -4.000 | 1049 | 2.3 | | | 1077 | 2.4 | | |
| -5.000 | 1206 | 2.6 | | | 1196 | 2.6 | | |
| -6.000 | 1151 | 2.5 | | | 1683 | 3.7 | | |
| -7.000 | 2118 | 4.6 | | | 3345 | 7.3 | | |
| -8.000 | 5212 | 11.4 | | | 4857 | 10.6 | | |
| -9.000 | 3017 | 6.6 | | | 510 | 1.1 | | |
| -10.000 | 0 | 0.0 | 11404 | 24.9 | 211 | 0.5 | 10141 | 22.2 |
| -11.000 | 0 | 0.0 | | | 268 | 0.6 | | |
| -12.000 | 0 | 0.0 | | | 510 | 1.1 | | |
| -13.000 | 0 | 0.0 | | | 221 | 0.5 | | |
| -14.000 | 0 | 0.0 | | | 431 | 0.9 | | |
| -15.000 | 0 | 0.0 | | | 619 | 1.4 | | |
| -16.000 | 0 | 0.0 | | | 930 | 2.0 | | |
| -17.000 | 0 | 0.0 | | | 615 | 1.3 | | |
| -18.000 | 0 | 0.0 | | | 67 | 0.1 | | |
| -19.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| -20.000 | 0 | 0.0 | | | 0 | 0.0 | 1303 | 2.8 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | -.02 | | -.08 | | .07 | | .84 | |
| S.D. | 6.22 | | 7.15 | | 8.12 | | 8.88 | |

Table 6

Distributions of Differences Between Re-equated(RE) and Original(OS) Rounded and Unrounded Scaled Scores Associated with Deletion of Item Ac (RE - OS)

| Scaled Score Difference | Item Ac 59-item unrounded | | Item Ac 59-item rounded | | Item Ac 58-item unrounded | | Item Ac 58-item rounded | |
|---|---|---|---|---|---|---|---|---|
| 30.000 | 0 | 0.0 | | | 0 | 0.0 | 4 | 0.0 |
| 29.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 28.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 27.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 26.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 25.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 24.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 23.000 | 0 | 0.0 | | | 4 | 0.0 | | |
| 22.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 21.000 | 0 | 0.0 | | | 5 | 0.0 | | |
| 20.000 | 0 | 0.0 | 17 | 0.0 | 20 | 0.0 | 415 | 0.9 |
| 19.000 | 0 | 0.0 | | | 2 | 0.0 | | |
| 18.000 | 0 | 0.0 | | | 19 | 0.0 | | |
| 17.000 | 0 | 0.0 | | | 72 | 0.2 | | |
| 16.000 | 0 | 0.0 | | | 70 | 0.2 | | |
| 15.000 | 17 | 0.0 | | | 169 | 0.4 | | |
| 14.000 | 1 | 0.0 | | | 173 | 0.4 | | |
| 13.000 | 0 | 0.0 | | | 283 | 0.6 | | |
| 12.000 | 109 | 0.2 | | | 437 | 1.0 | | |
| 11.000 | 25 | 0.1 | | | 402 | 0.9 | | |
| 10.000 | 181 | 0.4 | 6019 | 13.2 | 842 | 1.8 | 9702 | 21.2 |
| 9.000 | 329 | 0.7 | | | 1315 | 2.9 | | |
| 8.000 | 1979 | 4.3 | | | 6432 | 14.1 | | |
| 7.000 | 5086 | 11.1 | | | 302 | 0.7 | | |
| 6.000 | 55 | 0.1 | | | 414 | 0.9 | | |
| 5.000 | 74 | 0.2 | | | 310 | 0.7 | | |
| 4.000 | 44 | 0.1 | | | 623 | 1.4 | | |
| 3.000 | 154 | 0.3 | | | 760 | 1.7 | | |
| 2.000 | 351 | 0.8 | | | 1407 | 3.1 | | |
| 1.000 | 358 | 0.8 | | | 1767 | 3.9 | | |
| 0.0 | 1164 | 2.5 | 30839 | 67.4 | 4180 | 9.1 | 24246 | 53.0 |
| -1.000 | 6721 | 14.7 | | | 12338 | 27.0 | | |
| -2.000 | 25939 | 56.7 | | | 358 | 0.8 | | |
| -3.000 | 121 | 0.3 | | | 701 | 1.5 | | |
| -4.000 | 114 | 0.2 | | | 370 | 0.8 | | |
| -5.000 | 136 | 0.3 | | | 745 | 1.6 | | |
| -6.000 | 119 | 0.3 | | | 746 | 1.6 | | |
| -7.000 | 242 | 0.5 | | | 869 | 1.9 | | |
| -8.000 | 195 | 0.4 | | | 1635 | 3.6 | | |
| -9.000 | 339 | 0.7 | | | 4098 | 9.0 | | |
| -10.000 | 997 | 2.2 | 8772 | 19.2 | 2883 | 6.3 | 10934 | 23.9 |
| -11.000 | 909 | 2.0 | | | 180 | 0.4 | | |
| -12.000 | 0 | 0.0 | | | 81 | 0.2 | | |
| -13.000 | 0 | 0.0 | | | 52 | 0.1 | | |
| -14.000 | 0 | 0.0 | | | 103 | 0.2 | | |
| -15.000 | 0 | 0.0 | | | 130 | 0.3 | | |
| -16.000 | 0 | 0.0 | | | 71 | 0.2 | | |
| -17.000 | 0 | 0.0 | | | 73 | 0.2 | | |
| -18.000 | 0 | 0.0 | | | 194 | 0.4 | | |
| -19.000 | 0 | 0.0 | | | 108 | 0.2 | | |
| -20.000 | 0 | 0.0 | 112 | 0.2 | 16 | 0.0 | 458 | 1.0 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | - .58 | | - .64 | | - .35 | | - .29 | |
| S.D. | 4.23 | | 5.75 | | 6.65 | | 7.26 | |

Table 7

Distributions of Differences Between Re-equated(RE) and Original(OS) Rounded and Unrounded Scaled Scores Associated with Deletion of Item aC (RE - OS)

| Scaled Score Difference | Item aC 59-item unrounded | | Item aC 59-item rounded | | Item aC 58-item unrounded | | Item aC 58-item rounded | |
|---|---|---|---|---|---|---|---|---|
| 30.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 29.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 28.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 27.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 26.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 25.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 24.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 23.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 22.000 | 0 | 0.0 | | | 8 | 0.0 | | |
| 21.000 | 0 | 0.0 | | | 39 | 0.1 | | |
| 20.000 | 0 | 0.0 | 555 | 1.2 | 114 | 0.2 | 1794 | 3.9 |
| 19.000 | 0 | 0.0 | | | 74 | 0.2 | | |
| 18.000 | 0 | 0.0 | | | 131 | 0.3 | | |
| 17.000 | 0 | 0.0 | | | 178 | 0.4 | | |
| 16.000 | 0 | 0.0 | | | 181 | 0.4 | | |
| 15.000 | 12 | 0.0 | | | 329 | 0.7 | | |
| 14.000 | 53 | 0.1 | | | 469 | 1.0 | | |
| 13.000 | 931 | 2.0 | | | 431 | 0.9 | | |
| 12.000 | 640 | 1.4 | | | 1091 | 2.4 | | |
| 11.000 | 877 | 1.9 | | | 3533 | 7.7 | | |
| 10.000 | 1843 | 4.0 | 9618 | 21.0 | 352 | 0.8 | 11825 | 25.8 |
| 9.000 | 10 | 0.0 | | | 372 | 0.8 | | |
| 8.000 | 0 | 0.0 | | | 537 | 1.2 | | |
| 7.000 | 15 | 0.0 | | | 647 | 1.4 | | |
| 6.000 | 89 | 0.2 | | | 804 | 1.8 | | |
| 5.000 | 900 | 2.0 | | | 1424 | 3.1 | | |
| 4.000 | 2667 | 5.8 | | | 1901 | 4.2 | | |
| 3.000 | 4302 | 9.4 | | | 4355 | 9.5 | | |
| 2.000 | 10846 | 23.7 | | | 6615 | 14.5 | | |
| 1.000 | 5349 | 11.7 | | | 1114 | 2.4 | | |
| 0.0 | 32 | 0.1 | 25497 | 55.7 | 780 | 1.7 | 20311 | 44.4 |
| -1.000 | 74 | 0.2 | | | 381 | 0.8 | | |
| -2.000 | 284 | 0.6 | | | 936 | 2.0 | | |
| -3.000 | 811 | 1.8 | | | 1687 | 3.7 | | |
| -4.000 | 1444 | 3.2 | | | 1872 | 4.1 | | |
| -5.000 | 2258 | 4.9 | | | 2764 | 6.0 | | |
| -6.000 | 6089 | 13.3 | | | 3728 | 8.1 | | |
| -7.000 | 4470 | 9.8 | | | 3524 | 7.7 | | |
| -8.000 | 1763 | 3.9 | | | 904 | 2.0 | | |
| -9.000 | 0 | 0.0 | | | 230 | 0.5 | | |
| -10.000 | 0 | 0.0 | 10089 | 22.0 | 468 | 1.0 | 10693 | 23.4 |
| -11.000 | 0 | 0.0 | | | 426 | 0.9 | | |
| -12.000 | 0 | 0.0 | | | 399 | 0.9 | | |
| -13.000 | 0 | 0.0 | | | 773 | 1.7 | | |
| -14.000 | 0 | 0.0 | | | 1132 | 2.5 | | |
| -15.000 | 0 | 0.0 | | | 518 | 1.1 | | |
| -16.000 | 0 | 0.0 | | | 420 | 0.9 | | |
| -17.000 | 0 | 0.0 | | | 118 | 0.3 | | |
| -18.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| -19.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| -20.000 | 0 | 0.0 | | | 0 | 0.0 | 1136 | 2.5 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | .06 | | .14 | | .07 | | .54 | |
| S.D. | 5.43 | | 6.92 | | 7.56 | | 8.63 | |

Table 8

Distributions of Differences Between Re-equated(RE) and Original(OS) Rounded and Unrounded Scaled Scores Associated with Deletion of Item ec (RE - OS)

| Scaled Score Difference | Item ec 59-item unrounded | | Item ec 59-item rounded | | Item ec 58-item unrounded | | Item ec 58-item rounded | |
|---|---|---|---|---|---|---|---|---|
| 30.000 | 0 | 0.0 | | | 0 | 0.0 | 64 | 0.1 |
| 29.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 28.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 27.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 26.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 25.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 24.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 23.000 | 0 | 0.0 | | | 13 | 0.0 | | |
| 22.000 | 0 | 0.0 | | | 65 | 0.1 | | |
| 21.000 | 0 | 0.0 | | | 92 | 0.2 | | |
| 20.000 | 0 | 0.0 | 899 | 2.0 | 138 | 0.3 | 1910 | 4.2 |
| 19.000 | 0 | 0.0 | | | 128 | 0.3 | | |
| 18.000 | 0 | 0.0 | | | 81 | 0.2 | | |
| 17.000 | 0 | 0.0 | | | 257 | 0.6 | | |
| 16.000 | 0 | 0.0 | | | 218 | 0.5 | | |
| 15.000 | 798 | 1.7 | | | 392 | 0.9 | | |
| 14.000 | 444 | 1.0 | | | 470 | 1.0 | | |
| 13.000 | 452 | 1.0 | | | 973 | 2.1 | | |
| 12.000 | 739 | 1.6 | | | 1227 | 2.7 | | |
| 11.000 | 1274 | 2.8 | | | 1606 | 3.5 | | |
| 10.000 | 799 | 1.7 | 8502 | 18.6 | 1285 | 2.8 | 10815 | 23.6 |
| 9.000 | 354 | 0.8 | | | 783 | 1.7 | | |
| 8.000 | 70 | 0.2 | | | 636 | 1.4 | | |
| 7.000 | 416 | 0.9 | | | 687 | 1.5 | | |
| 6.000 | 1578 | 3.4 | | | 1260 | 2.8 | | |
| 5.000 | 1779 | 3.9 | | | 1583 | 3.5 | | |
| 4.000 | 2313 | 5.1 | | | 2098 | 4.6 | | |
| 3.000 | 3223 | 7.0 | | | 2441 | 5.3 | | |
| 2.000 | 3715 | 8.1 | | | 2819 | 6.2 | | |
| 1.000 | 2512 | 5.5 | | | 2148 | 4.7 | | |
| 0.0 | 1483 | 3.2 | 25258 | 55.2 | 1932 | 4.2 | 20257 | 44.3 |
| -1.000 | 1702 | 3.7 | | | 1475 | 3.2 | | |
| -2.000 | 2043 | 4.5 | | | 1703 | 3.7 | | |
| -3.000 | 3069 | 6.7 | | | 2166 | 4.7 | | |
| -4.000 | 4896 | 10.7 | | | 2767 | 6.0 | | |
| -5.000 | 4902 | 10.7 | | | 3225 | 7.0 | | |
| -6.000 | 3412 | 7.5 | | | 2205 | 4.8 | | |
| -7.000 | 1883 | 4.1 | | | 2118 | 4.6 | | |
| -8.000 | 1307 | 2.9 | | | 1197 | 2.6 | | |
| -9.000 | 435 | 1.0 | | | 940 | 2.1 | | |
| -10.000 | 161 | 0.4 | 11100 | 24.3 | 726 | 1.6 | 11452 | 25.0 |
| -11.000 | 0 | 0.0 | | | 826 | 1.8 | | |
| -12.000 | 0 | 0.0 | | | 886 | 1.9 | | |
| -13.000 | 0 | 0.0 | | | 900 | 2.0 | | |
| -14.000 | 0 | 0.0 | | | 525 | 1.1 | | |
| -15.000 | 0 | 0.0 | | | 339 | 0.7 | | |
| -16.000 | 0 | 0.0 | | | 211 | 0.5 | | |
| -17.000 | 0 | 0.0 | | | 138 | 0.3 | | |
| -18.000 | 0 | 0.0 | | | 55 | 0.1 | | |
| -19.000 | 0 | 0.0 | | | 25 | 0.1 | | |
| -20.000 | 0 | 0.0 | | | 0 | 0.0 | 1261 | 2.8 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | .00+ | | - .17 | | .03 | | .19 | |
| S.D. | 5.74 | | 7.12 | | 7.70 | | 8.81 | |

The mean differences demonstrate the interaction between rounding and number of items deleted. With the exception of Table 6, most mean unrounded differences are close to zero. In contrast, the absolute value of the mean rounded differences for the 59-item tests are in the range .08 to .15, while those for the 58-item tests are in the range .19 to .84. Rounding exaggerates the item deletion effect.

Table 6 is the exception to the rule. Item Ac is the only item for which the psychometric properties have much of an impact on mean unrounded differences. This item is highly discriminating ($a_{Ac} = 1.48$), above average in difficulty ($b_{Ac} = 1.48$), and has a low c-parameter ($c_{Ac} = .05$). Recall that when discussing Figure 5, it was noted that the shortened tests were easier than the original 60-item test for most of the formula score distribution. As a consequence, many more people were affected negatively as a consequence of deleting this item than were affected negatively by deletion of the other items. If this item had been of middle difficulty, the resultant standard deviation of differences would have been larger than any of those observed, and the mean difference of unrounded scores would have been close to zero. In short, the high a-parameter and low c-parameter allow the difficulty parameter to have its maximum effect.

To re-equate or not to re-equate. Tables 9-14 parallel Tables 3-8 and illustrate what would happen to scaled score distributions if after the flawed items were scored all options correct, a decision was made to not re-equate. These six tables contain differences between scaled scores based on re-equating and scaled scores based on using the original conversions on the "all options correct" adjusted formula scores. In all six tables, all differences, rounded and unrounded, are

non-negative, indicating that use of the original conversions with the "all options correct" adjusted formula scores introduces a positive bias, i.e., these "converted" scores are always as high or higher than the appropriated converted score resulting from re-equating after deletion. Note that all non-negative differences are consistent with the upper panels in Figures 1-6.

The extent of the positive bias clearly is related to the psychometric properties of the deleted item, in particular its difficulty level: the more difficult the deleted item, the larger the positive bias. The mean differences in Tables 9-14 reflect the extent of positive bias.

The final point to note in Tables 9-14 is that the decision not to re-equate has enabled the deleted item's psychometric properties to have the dominant impact on reported scores. In contrast, re-equating put the item's properties on a par with the arbitrary rounding effects. Clearly, re-equating after item deletion is necessary from a psychometric viewpoint.

Table 9

Distributions of Differences Between Unrounded and Rounded Scores Associated with Scoring Item MD All Options Correct and Using Original Equating(OE) vs. Re-equating(RE): (OE - RE)

| Scaled Score Difference | Item MD 59-item[a] unrounded | | Item MD 59-item[a] rounded | | Item MD 58-item[a] unrounded | | Item MD 58-item[a] rounded | |
|---|---|---|---|---|---|---|---|---|
| 20.000 | 0 | 0.0 | | | 0 | 0.0 | 28854 | 63.1 |
| 19.000 | 0 | 0.0 | | | 133 | 0.3 | | |
| 18.000 | 0 | 0.0 | | | 9238 | 20.2 | | |
| 17.000 | 0 | 0.0 | | | 19969 | 43.6 | | |
| 16.000 | 0 | 0.0 | | | 4894 | 10.7 | | |
| 15.000 | 0 | 0.0 | | | 3052 | 6.7 | | |
| 14.000 | 0 | 0.0 | | | 1854 | 4.1 | | |
| 13.000 | 0 | 0.0 | | | 1575 | 3.4 | | |
| 12.000 | 0 | 0.0 | | | 610 | 1.3 | | |
| 11.000 | 0 | 0.0 | | | 1253 | 2.7 | | |
| 10.000 | 5754 | 12.6 | 41905 | 91.6 | 936 | 2.0 | 16052 | 35.1 |
| 9.000 | 35807 | 78.3 | | | 770 | 1.7 | | |
| 8.000 | 2050 | 4.5 | | | 565 | 1.2 | | |
| 7.000 | 1271 | 2.8 | | | 358 | 0.8 | | |
| 6.000 | 524 | 1.1 | | | 192 | 0.4 | | |
| 5.000 | 222 | 0.5 | | | 230 | 0.5 | | |
| 4.000 | 40 | 0.1 | | | 43 | 0.1 | | |
| 3.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 2.000 | 58 | 0.1 | | | 57 | 0.1 | | |
| 1.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 0.0 | 33 | 0.1 | 3854 | 8.4 | 30 | 0.1 | 853 | 1.9 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | 8.95 | | 9.16 | | 15.86 | | 16.12 | |
| S.D. | .78 | | 2.78 | | 2.69 | | 5.24 | |

[a]Although the 59-item and 58-item tests literally are both 60-items long when the problem item(s) is (are) scored all options correct, the headings remain 59-item and 58-item for two reasons: (1) To facilitate comparison of these results with those obtained when the deleted item(s) is (are) not scored; (2) These 60-item tests are figuratively 59-item and 58-item tests because individual candidate responses to the deleted item or items are ignored.

Table 10

Distributions of Differences Between Unrounded and Rounded Scores Associated with Scoring Item LD All Options Correct and Using Original Equating(OE) vs. Re-equating(RE): (OE - RE)

| Scaled Score Difference | Item LD 59-item[a] unrounded | | Item LD 59-item[a] rounded | | Item LD 58-item[a] unrounded | | Item LD 58-item[a] rounded | |
|---|---|---|---|---|---|---|---|---|
| 20.000 | 0 | 0.0 | | | 0 | 0.0 | 2381 | 5.2 |
| 19.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 18.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 17.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 16.000 | 0 | 0.0 | | | 121 | 0.3 | | |
| 15.000 | 0 | 0.0 | | | 261 | 0.6 | | |
| 14.000 | 0 | 0.0 | | | 227 | 0.5 | | |
| 13.000 | 0 | 0.0 | | | 901 | 2.0 | | |
| 12.000 | 0 | 0.0 | | | 1396 | 3.1 | | |
| 11.000 | 0 | 0.0 | | | 3522 | 7.7 | | |
| 10.000 | 0 | 0.0 | 6498 | 14.2 | 6004 | 13.1 | 33407 | 73.0 |
| 9.000 | 0 | 0.0 | | | 10147 | 22.2 | | |
| 8.000 | 76 | 0.2 | | | 9301 | 20.3 | | |
| 7.000 | 63 | 0.1 | | | 3355 | 7.3 | | |
| 6.000 | 227 | 0.5 | | | 2095 | 4.6 | | |
| 5.000 | 515 | 1.1 | | | 2575 | 5.6 | | |
| 4.000 | 1051 | 2.3 | | | 1452 | 3.2 | | |
| 3.000 | 3002 | 6.6 | | | 1689 | 3.7 | | |
| 2.000 | 7329 | 16.0 | | | 1129 | 2.5 | | |
| 1.000 | 23645 | 51.7 | | | 1140 | 2.5 | | |
| 0.0 | 9851 | 21.5 | 39261 | 85.8 | 444 | 1.0 | 9971 | 21.8 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | 1.23 | | 1.42 | | 7.98 | | 8.34 | |
| S.D. | 1.11 | | 3.49 | | 2.82 | | 4.92 | |

[a]Although the 59-item and 58-item tests literally are both 60-items long when the problem item(s) is (are) scored all options correct, the headings remain 59-item and 58-item for two reasons: (1) To facilitate comparison of these results with those obtained when the deleted item(s) is (are) not scored; (2) These 60-item tests are figuratively 59-item and 58-item tests because individual candidate responses to the deleted item or items are ignored.

Table 11

Distributions of Differences Between Unrounded and Rounded Scores Associated with Scoring Item AC All Options Correct and Using Original Equating(OE) vs. Re-equating(RE): (OE - RE)

| Scaled Score Difference | Item AC 59-item[a] unrounded | | Item AC 59-item[a] rounded | | Item AC 58-item[a] unrounded | | Item AC 58-item[a] rounded | |
|---|---|---|---|---|---|---|---|---|
| 20.000 | 0 | 0.0 | | | 0 | 0.0 | 18495 | 40.4 |
| 19.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 18.000 | 0 | 0.0 | | | 1283 | 2.8 | | |
| 17.000 | 0 | 0.0 | | | 10190 | 22.3 | | |
| 16.000 | 0 | 0.0 | | | 12827 | 28.0 | | |
| 15.000 | 0 | 0.0 | | | 5347 | 11.7 | | |
| 14.000 | 0 | 0.0 | | | 2555 | 5.6 | | |
| 13.000 | 0 | 0.0 | | | 1178 | 2.6 | | |
| 12.000 | 0 | 0.0 | | | 2429 | 5.3 | | |
| 11.000 | 0 | 0.0 | | | 1033 | 2.3 | | |
| 10.000 | 0 | 0.0 | 31351 | 68.5 | 882 | 1.9 | 20900 | 45.7 |
| 9.000 | 10314 | 22.5 | | | 1001 | 2.2 | | |
| 8.000 | 17606 | 38.5 | | | 897 | 2.0 | | |
| 7.000 | 5110 | 11.2 | | | 841 | 1.8 | | |
| 6.000 | 2355 | 5.1 | | | 707 | 1.5 | | |
| 5.000 | 2074 | 4.5 | | | 561 | 1.2 | | |
| 4.000 | 1725 | 3.8 | | | 1228 | 2.7 | | |
| 3.000 | 1604 | 3.5 | | | 510 | 1.1 | | |
| 2.000 | 1184 | 2.6 | | | 1030 | 2.3 | | |
| 1.000 | 2239 | 4.9 | | | 812 | 1.8 | | |
| 0.0 | 1548 | 3.4 | 14408 | 31.5 | 448 | 1.0 | 6364 | 13.9 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | 6.78 | | 6.85 | | 13.56 | | 12.65 | |
| S.D. | 2.54 | | 4.64 | | 4.54 | | 6.88 | |

[a]Although the 59-item and 58-item tests literally are both 60-items long when the problem item(s) is (are) scored all options correct, the headings remain 59-item and 58-item for two reasons: (1) To facilitate comparison of these results with those obtained when the deleted item(s) is (are) not scored; (2) These 60-item tests are figuratively 59-item and 58-item tests because individual candidate responses to the deleted item or items are ignored.

Table 12

Distributions of Differences Between Unrounded and Rounded Scores Associated with Scoring Item Ac All Options Correct and Using Original Equating(OE) vs. Re-equating(RE): (OE - RE)

| Scaled Score Difference | Item Ac 59-item[a] unrounded | | Item Ac 59-item[a] rounded | | Item Ac 58-item[a] unrounded | | Item Ac 58-item[a] rounded | |
|---|---|---|---|---|---|---|---|---|
| 20.000 | 0 | 0.0 | 2703 | 5.9 | 574 | 1.3 | 31506 | 68.9 |
| 19.000 | 0 | 0.0 | | | 9669 | 21.1 | | |
| 18.000 | 0 | 0.0 | | | 21481 | 46.9 | | |
| 17.000 | 0 | 0.0 | | | 3483 | 7.6 | | |
| 16.000 | 0 | 0.0 | | | 1982 | 4.3 | | |
| 15.000 | 0 | 0.0 | | | 1893 | 4.1 | | |
| 14.000 | 0 | 0.0 | | | 1584 | 3.5 | | |
| 13.000 | 0 | 0.0 | | | 622 | 1.4 | | |
| 12.000 | 0 | 0.0 | | | 689 | 1.5 | | |
| 11.000 | 17646 | 38.6 | | | 1122 | 2.5 | | |
| 10.000 | 23301 | 50.9 | 41872 | 91.5 | 424 | 0.9 | 13331 | 29.1 |
| 9.000 | 1789 | 3.9 | | | 390 | 0.9 | | |
| 8.000 | 879 | 1.9 | | | 397 | 0.9 | | |
| 7.000 | 748 | 1.6 | | | 313 | 0.7 | | |
| 6.000 | 302 | 0.7 | | | 242 | 0.5 | | |
| 5.000 | 349 | 0.8 | | | 122 | 0.3 | | |
| 4.000 | 233 | 0.5 | | | 247 | 0.5 | | |
| 3.000 | 194 | 0.4 | | | 202 | 0.4 | | |
| 2.000 | 140 | 0.3 | | | 137 | 0.3 | | |
| 1.000 | 94 | 0.2 | | | 102 | 0.2 | | |
| 0.0 | 84 | 0.2 | 1184 | 2.6 | 84 | 0.2 | 922 | 2.0 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | 10.07 | | 10.33 | | 16.81 | | 16.68 | |
| S.D. | 1.37 | | 2.90 | | 3.22 | | 5.12 | |

[a] Although the 59-item and 58-item tests literally are both 60-items long when the problem item(s) is (are) scored all options correct, the headings remain 59-item and 58-item for two reasons: (1) To facilitate comparison of these results with those obtained when the deleted item(s) is (are) not scored; (2) These 60-item tests are figuratively 59-item and 58-item tests because individual candidate responses to the deleted item or items are ignored.

Table 13

Distributions of Differences Between Unrounded and Rounded Scores Associated with Scoring Item aC All Options Correct and Using Original Equating(OE) vs. Re-equating(RE): (OE - RE)

| Scaled Score Difference | Item aC 59-item[a] unrounded | | Item aC 59-item[a] rounded | | Item aC 58-item[a] unrounded | | Item aC 58-item[a] rounded | |
|---|---|---|---|---|---|---|---|---|
| 20.000 | 0 | 0.0 | | | 0 | 0.0 | 14655 | 32.0 |
| 19.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 18.000 | 0 | 0.0 | | | 0 | 0.0 | | |
| 17.000 | 0 | 0.0 | | | 2188 | 4.8 | | |
| 16.000 | 0 | 0.0 | | | 7446 | 16.3 | | |
| 15.000 | 0 | 0.0 | | | 6759 | 14.8 | | |
| 14.000 | 0 | 0.0 | | | 11956 | 26.1 | | |
| 13.000 | 0 | 0.0 | | | 5065 | 11.1 | | |
| 12.000 | 0 | 0.0 | | | 2312 | 5.1 | | |
| 11.000 | 0 | 0.0 | | | 1946 | 4.3 | | |
| 10.000 | 0 | 0.0 | 28435 | 62.1 | 1817 | 4.0 | 27990 | 61.2 |
| 9.000 | 0 | 0.0 | | | 816 | 1.8 | | |
| 8.000 | 6685 | 14.6 | | | 1289 | 2.8 | | |
| 7.000 | 14771 | 32.3 | | | 1239 | 2.7 | | |
| 6.000 | 15948 | 34.9 | | | 861 | 1.9 | | |
| 5.000 | 4424 | 9.7 | | | 735 | 1.6 | | |
| 4.000 | 2312 | 5.1 | | | 545 | 1.2 | | |
| 3.000 | 1143 | 2.5 | | | 493 | 1.1 | | |
| 2.000 | 354 | 0.8 | | | 168 | 0.4 | | |
| 1.000 | 96 | 0.2 | | | 98 | 0.2 | | |
| 0.0 | 26 | 0.1 | 17324 | 37.9 | 26 | 0.1 | 3114 | 6.8 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | 6.30 | | 6.21 | | 13.06 | | 12.52 | |
| S.D. | 1.24 | | 4.85 | | 3.23 | | 5.70 | |

[a]Although the 59-item and 58-item tests literally are both 60-items long when the problem item(s) is (are) scored all options correct, the headings remain 59-item and 58-item for two reasons: (1) To facilitate comparison of these results with those obtained when the deleted item(s) is (are) not scored; (2) These 60-item tests are figuratively 59-item and 58-item tests because individual candidate responses to the deleted item or items are ignored.

Table 14

Distributions of Differences Between Unrounded and Rounded Scores Associated with Scoring Item ac All Options Correct and Using Original Equating(OE) vs. Re-equating(RE): (OE - RE)

| Scaled Score Difference | Item ac 59-item[a] unrounded | | Item ac 59-item[a] rounded | | Item ac 58-item[a] unrounded | | Item ac 58-item[a] rounded | |
|---|---|---|---|---|---|---|---|---|
| 20.000 | 0 | 0.0 | | | 0 | 0.0 | 14136 | 30.9 |
| 19.000 | 0 | 0.0 | | | 674 | 1.5 | | |
| 18.000 | 0 | 0.0 | | | 1479 | 3.2 | | |
| 17.000 | 0 | 0.0 | | | 3115 | 6.8 | | |
| 16.000 | 0 | 0.0 | | | 3482 | 7.6 | | |
| 15.000 | 0 | 0.0 | | | 4196 | 9.2 | | |
| 14.000 | 0 | 0.0 | | | 5000 | 10.9 | | |
| 13.000 | 0 | 0.0 | | | 6763 | 14.8 | | |
| 12.000 | 0 | 0.0 | | | 5297 | 11.6 | | |
| 11.000 | 0 | 0.0 | | | 3620 | 7.9 | | |
| 10.000 | 953 | 2.1 | 25092 | 54.8 | 2337 | 5.1 | 26046 | 56.9 |
| 9.000 | 2174 | 4.8 | | | 1869 | 4.1 | | |
| 8.000 | 4532 | 9.9 | | | 960 | 2.1 | | |
| 7.000 | 5126 | 11.2 | | | 1668 | 3.6 | | |
| 6.000 | 7593 | 16.6 | | | 1271 | 2.8 | | |
| 5.000 | 9287 | 20.3 | | | 1210 | 2.6 | | |
| 4.000 | 7899 | 17.3 | | | 824 | 1.8 | | |
| 3.000 | 4383 | 9.6 | | | 726 | 1.6 | | |
| 2.000 | 2581 | 5.6 | | | 800 | 1.7 | | |
| 1.000 | 1157 | 2.5 | | | 397 | 0.9 | | |
| 0.0 | 74 | 0.2 | 20667 | 45.2 | 71 | 0.2 | 5577 | 12.2 |
| N | 45759 | | 45759 | | 45759 | | 45759 | |
| Mean | 5.34 | | 5.48 | | 12.09 | | 11.87 | |
| S.D. | 2.04 | | 4.98 | | 3.98 | | 6.29 | |

[a] Although the 59-item and 58-item tests literally are both 60-items long when the problem item(s) is (are) scored all options correct, the headings remain 59-item and 58-item for two reasons: (1) To facilitate comparison of these results with those obtained when the deleted item(s) is (are) not scored; (2) These 60-item tests are figuratively 59-item and 58-item tests because individual candidate responses to the deleted item or items are ignored.

## Conclusions

This report has presented a formal analysis of the effects of item deletion on equating/scaling functions and on reported score distributions. The analysis, based on item response theory, was used to decompose the item deletion effect into its constituent elements. This analysis was supplemented by empirical illustrations drawn from the May 1982 administration of the SAT-Mathematical test that contained the circles item.

The item deletion effect can be separated into several components. Deletion introduces changes in the equating function that maps formula scores onto the reported score scale. The psychometric characteristics of item and rounding rules for scaled scores contribute to the change in equating function. An item's difficulty determines where the change in equating function occurs along the formula score continuum. Deletion of a very difficult item can have no substantial effect on the equating function when the item is not scored. Deletion of an easy item under the not score condition, however, can have a very noticeable effect. In contrast, scoring the item all options correct makes deletion of the easy item essentially transparent and deletion of the hard item quite noticeable. An item's discriminating power determines the abruptness and direction of the effect. Deletion of a highly discriminating item produces an abrupt change in the equating function near the item's difficulty parameter. In contrast, deletion of a poorly discriminating item produces a gradual shift that affects more of the scores centered around the item's difficulty level. Finally, the item's susceptibility to guessing modulates the effect. Deleting an item with a high lower asymptote precludes the occurrence of positive differences in equating

functions under the not score condition.. Deletion of an item with a very low lower asymptote will yield positive differences, the number of which increase with the difficulty and discriminating power of the item.

The illustrative data demonstrated the importance of rounding rules for scaled scores. While the formal analysis referred to the impact of the rules, the illustrations vividly portrayed their impact. In many cases under re-equating, rounding has a large if not a larger effect on reported scores than the psychometric properties of the item. The same can be said for the act of item deletion itself. Deleting an item in general will have an effect on reported score distributions, a greater effect, in fact, than the particular psychometric properties of the deleted item, provided that re-equating is performed.

The reason that the psychometric properties of the item tend to have a smaller effect on reported score distribution differences than either rounding the scaled scores or the act of item deletion is re-equating. Re-equating, particularly via item response theory, compensates for the loss of the deleted item's psychometric properties. As a corollary, deletion without re-equating allows the deleted item's properties to have a more substantial impact. This fact explains why re-equating is psychometrically desirable after an item is deleted.

## Reference Notes

1. Petersen, N. S. Effects of not scoring Math I item 17 on SAT-M Form 3ESA05. Unpublished memorandum, June 16, 1982.

2. Petersen, N. S. Effect on scores of rescoring items 62 and 63 in Biology Form XAC and re-equating. Unpublished memorandum, September 23, 1982.

3. Petersen, N. S. Effect on scores of giving everyone credit on items 62 and 63 in Biology Form XAC. Unpublished memorandum, September 27, 1982.

4. Wainer, H. The item influence function: A strategy for dealing with unusual items. Unpublished manuscript, 1981.

## References

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

Wainer, H. Pyramid power: Searching for an error in test scoring with 830,000 helpers. ETS Research Report RR-31-27. Princeton, NJ: Educational Testing Service, 1981.