DOCUMENT RESUME

ED 263 221 TM 850 723

AUTHOR Albanese, Mark A.

TITLE Some Comments on the Correction for Guessing. A Further Analysis of Angoff and Schrader.

PUB DATE Apr 85

NOTE 35p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Chicago, IL, April 1-3, 1985).

PUB TYPE Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS *College Entrance Examinations; Data Interpretation; Fatigue (Biology); *Guessing (Tests); Higher Education; High School Students; Hypothesis Testing; Research Problems; Response Style (Tests); *Sampling; *Scoring Formulas; Secondary Education; Testing Problems; *Test Reliability; Test Wiseness; Volunteers

IDENTIFIERS *Differential Effects Hypothesis; Graduate Management Admission Test; *Invariance Hypothesis; Scholastic Aptitude Test; Test Instructions

ABSTRACT

This study reexamines results reported by Angoff and Schrader regarding formula directions and rights directions for standardized tests. In that study, it was concluded that the two scoring directions were essentially equivalent. In this study, methodological concerns are discussed and additional data analyses undertaken. Among various methodological concerns discussed are the potential problems in using volunteers for the College Board phase of the study and the likelihood of treatment contamination in the Graduate Management Admissions Test (GMAT) phase. Estimates of success rates of the rights directions group on items omitted and not reached by the formula group, were beyond chance by 3 percent to 13 percent for the College Board tests and at or below chance levels for the GMAT. Alternative interpretations of the data are made and suggestions for additional research proposed. (Author/LMO)

Some Comments on
The Correction for Guessing
A Further Analysis of Angoff and Schrader

By Mark A. Albanese

SOME COMMENTS ON THE CORRECTION FOR GUESSING
A Further Analysis of Angoff and Schrader

MARK A. ALBANESE
University of Iowa

This study reexamines results reported by Angoff and Schrader (1981, 1984) regarding formula directions and rights directions for standardized tests. In that study, it was concluded that the two scoring directions were essentially equivalent. In this study, methodological concerns are discussed and additional data analyses undertaken. Among various methodological concerns discussed are the potential problems in using volunteers for the College Board phase of the study and the likelihood of treatment contamination in the GMAT phase. Estimates of success rates of the rights directions group on items omitted and not reached by the formula group, were beyond chance by 3% to 13% for the College Board tests and at or below chance levels for the GMAT. Alternative interpretations of the data are made and suggestions for additional research proposed.

SOME COMMENTS ON

THE CORRECTION FOR GUESSING[1]

A Further Analysis of Angoff and Schrader

By Mark A. Albanese

Angoff and Schrader (1981, 1984) present results from a very carefully conceived and impressively large study to examine whether or not certain examinees were "penalized" when formula scoring was used. Their results seemed to indicate that students are not penalized when formula scoring is used. However, there are several methodological issues that could be raised regarding the results and the ability to generalize from the results. For instance, the subjects for the major portion of the study were volunteers who took an old form of the Scholastic Aptitude Test Verbal Subtest (SAT-Verbal). To what extent one can apply the results from a no risk situation to an operational test is questionable. Because of the size of the study and the prestige of the authors, the study is likely to have a sizable impact on practitioners. It is important, therefore, that the study be carefully examined and any limitations on generalizations of the results and weaknesses of the study be thoroughly presented.

In examining the formula scoring issue, Angoff and Schrader coined terms for two competing hypotheses to be tested. The first hypothesis, the Differential Effects hypothesis, states that some students, when tested under formula directions, omit items about which they have useful partial knowledge. This implies that such directions are not as fair as rights directions, especially to those students who are less inclined to guess. The

[1] The author wishes to acknowledge Robert A. Forsyth for his helpful review and critique of earlier drafts of this paper.

alternative hypothesis, the Invariance hypothesis, states that examinees would perform no better than chance expectation on items that they would omit under formula directions but would answer under rights directions.

To test these hypotheses, Angoff and Schrader conducted a two phase study. The first phase involved volunteer high school students taking two forms of the SAT-Verbal test and another group taking the Chemistry Achievement test. The SAT-Verbal test had two sections, either of which could be administered under rights directions or formula directions. Four groups were formed for the SAT-Verbal test based on how the instructions were given (both sections under rights directions; both sections under formula directions; section one--rights, section two--formula; section one--formula, section two--rights). For the Chemistry Achievement test only two groups were formed (rights directions, formula directions). The number of students in each of these groups ranged from 1026 to 1155. Angoff and Schrader compared the formula scores obtained on each section of the tests and concluded that few differences in results were found, providing support for the Invariance hypothesis. They further analyzed the results by dividing the groups according to: performance on a separate section of the test, ethnic group, and the number of items not answered on a section. Once again, they concluded that the formula scores were equivalent, but in this analysis they generalize across ethnic groups, ability groups and tendency to omit items.

In the second phase of the study, an experimental section (the last of eight separately-timed sections) of the regular operational administration of the Graduate Management Admissions Test (GMAT) was studied over a period just short of three years. A total of 55,780 examinees received one of 10 different experimental forms with from 5408 to 5739 examinees taking each form. The ten different forms were composed of five different subtests comparable to those on the regular GMAT but administered under the two different types of directions (5 subtests x 2 different types of directions = 10 forms). Mean formula scores for rights and formula directions were interpreted as being quite similar for all five subtests. Angoff and Schrader concluded that the results from both the College Board studies and the GMAT studies yielded results that appear to be consistent with the Invariance hypothesis and that formula scoring has the effect of compensating for differences in guessing strategies, so that it is not necessary to require every examinee to answer every item to equalize guessing strategies.

In examining the Angoff and Schrader study, the discussion will be divided into two major sections. The first section will deal with methodological issues while the second section will present results from a further analysis of the Angoff and Schrader data.

Section I: Methodological Issues

There are a number of methodological issues that may be raised concerning the Angoff and Schrader study. In this section they are grouped according to three issues: 1) issues related to the nature of the populations studied; 2) issues related to the testing situation; and 3) issues related to the interpretation of the data that were obtained. Within each part, both phases of the study will be discussed. It should be noted that much of the data to be discussed is found in the 1981 research report published by ETS (Angoff and Schrader, 1981) and may not be in the 1984 <u>Journal of Educational Measurement</u> article.

1) Issues related to the nature of the population studied

As one examines the populations involved in the Angoff and Schrader study, it becomes readily apparent that the groups studied cannot be considered representative of a very broad population. This is a problem faced by almost all studies. It is raised as an issue only to more clearly articulate the bounds to which the Angoff and Schrader study results can be generalized. In Phase I, the study population consisted of college-bound juniors in high school who volunteered to participate. The percentage of college bound juniors in each school who volunteered is not reported (and may not be known), however, of 109 schools included in the SAT-Verbal sample, only 52 provided usable data. Thus it is difficult to determine to what extent the participants are representative of college-bound juniors at their respective schools let alone all college-bound juniors. In addition, because the college-bound junior population tends to contain a disproportionately greater number of high academic achievers, it is likely to be unrepresentative of the larger population of high school juniors. For instance, the self-report class rank of the students in the SAT part of the study showed 63% in the top 2/5 of the class and only 7% in the bottom 2/5. The distribution for the group taking the Chemistry Achievement test was even more extreme with 72% in the highest 2/5 and 5% in the lowest 2/5. Because of this documented academic achievement bias, generalization to the larger population of juniors in high school must be done cautiously and most certainly only with additional data that would suggest such a generalization is warranted.

In Phase II, the examinee population was even more specialized--college graduates (or soon to be graduates) who were applying for admission to graduate school in business. It may be possible to generalize from the Phase II results to students who are taking the Graduate Record Examination (GRE), however, GRE examinees are a much broader representation of prospective graduate students than are examinees taking the GMAT. Thus, there may be idiosynchracies of the GMAT population that may make them different from the larger group of aspiring graduate students. As a result, care should be exerted in any generalization of Phase II results to other examinee populations.

2) Issues related to the nature of the testing situation studied

There are several aspects of the testing situations studied that are of interest. First, it should be noted that in both Phase I and II, the tests represented broad samplings of content that students should have been exposed to over a period of years. This is quite different from a test developed for a single class or course. In addition, for Phase I, the items were derived from old forms of the SAT test and were, therefore, a very carefully scrutinized set of items. They were most likely to be of much higher quality than would be the typical instructor-made examination. This suggests that the results from Angoff and Schrader should not be generalized to classroom examinations.

Second, the tests were administered under atypical circumstances. Phase I was administered under totally volunteer conditions. It was a special testing that involved gaining the cooperation of schools selected to conform to some very exacting standards in terms of the number of eligible examinees and minority enrollment. Directions to students made it clear that those who participated were there on a voluntary basis. (Angoff and Schrader, 1981).

Whether or not examinees perceived the tests to be operational has important implications for interpreting the results. One of the more serious arguments against formula scoring is that it adds an element of risk taking into the examinee's test taking task. Because points are lost for a wrong answer, students who are not completely certain of an answer, must decide whether they are confident enough of their response to risk the point loss. The concept of risk is to a large degree dependent on what is at stake. In earlier studies of the Differential Effects/Invariance hypotheses test performance contributed toward course grades (see, for example, Sherriffs and Boomer, 1954; Slakter et al. 1968; Cross and Frary, 1977). Thus, course grades were at stake. In Phase I of the Angoff and Schrader study, participation was totally voluntary. There was nothing at stake for the participating students except perhaps pride. With nothing at stake, some students may not take the testing experience as serious as they might otherwise. Thus, the efforts they make to answer items which are difficult for them may not be as great as they would be under conditions where the consequences of poor performance are perceived to be more serious. This could lead to performance under rights directions on items that students would omit under formula directions being closer to chance levels than they might be otherwise. It is an unusual test where there is nothing at stake for the examinee. In course examinations students grades are at stake. In admissions tests such as the SAT, students' admission to college and/or financial assistance may be at stake. It is doubtful that Angoff and Schrader's results from Phase I can be generalized to such situations but must be limited to the situation where students are voluntarily taking an examination.

While Phase I participation was unquestionably voluntary, Phase II participation occurred during the course of operational administrations of the GMAT. Thus, one might expect the examinees to have treated the experimental subtest as seriously as all of the remainder of the test. However, this may not necessarily have been the case. For the rights directions subtest, it was the last subtest and the only one of eight subtests administered under rights directions. The remaining seven subtests were under formula directions. It is possible that students in this case recognized that the subtest was experimental. The formula directions group, however, had no such clue and most likely answered items on the experimental subtest as though they were operational. Thus, comparison of the rights directions with formula directions in Phase II may have confounded volunteer participation with non-volunteer participation. It is possible therefore that the formula directions examinees took the experience more seriously than the rights directions examinees.

Besides the potential for confounding volunteer with non-volunteer participation, it should be noted that in all cases the experimental GMAT subtest was the last in the sequence of subtests. Thus, examinees encountered the experimental test after three hours of intensive testing. It is, therefore, very likely that examinees were fatigued by the time they encountered the experimental subtest. Fatigue would most likely have the effect of reducing the care with which the examinees considered their responses. It might also either dull or accentuate the examinee's concern as to the risk of making an incorrect response. For instance, some students may just want to get the test over with, throw caution to the wind and respond somewhat cavalierly to questions. Other students may recognize their diminished thought capacity and respond quite conservatively to items, making many omissions.

Another concern regarding the administration conditions involves the degree of risk as perceived by the examinee. The directions given by Angoff and Schrader are vague in describing the amount of points that will be subtracted for a wrong answer. They say that a fraction will be subtracted for each wrong answer and then suggest that if one option can be eliminated that it will probably be to the examinee's advantage to guess from among the rest. Under such directions, examinees are left to guess what they risk by a wrong response. This could lead some students to be overly cautious and others to be somewhat reckless depending on how these directions are interpreted. Also, even if the formula is provided, it may be misconstrued. The traditional formula subtracts from the number right the number wrong divided by one less than the number of options. Experience with examinees suggests that at least a few think that the number of options is the divisor that would adjust for chance success. When one less than the number of options is applied it gives the impression that the adjustment is in excess of chance. This could possibly enhance the perception of risk felt by examinees leading them to be more cautious than they might be otherwise.

A final concern, and perhaps the most serious, is that because the experimental test occurred after seven subtests administered under formula directions, and since the directions were self administered, it is likely that at least a few examinees in the rights directions group failed to read the directions and took the test as though it were under formula directions. If this were the case, one would expect to see comparatively high rates of omitted items for the rights directions group in phase II, perhaps even approaching the rate for the formula directions group as the upper limit. Table 1 shows the mean percentage of omitted and not reached items for the rights directions group and formula directions group averaged over all of the five subtests studied in phase II as well as similar data for phase I. The phase I data is reported for comparative purposes since the directions in phase I were less likely to be overlooked.

Insert Table 1 About Here

In phase II, examinees taking the test under rights directions on the average omitted two items for every three items omitted by the formula directions group. This compares with a one to four ratio in Phase I. In both phase I and phase II, examinees taking the test under rights directions failed to reach an average of approximately three items for every four items not reached by the formula directions group. Thus, on the average, examinees in phase II in the rights direction group omitted items at a rate much more comparable to that of the formula directions group than did examinees in phase I. No such affect was noted for the number of items not reached. These results would suggest that a fairly large number of examinees in the phase II rights directions group took the test as though it were under formula directions.

3) Issues related to the interpretation of the data

One might wonder that given the methodological issues raised whether a consideration of the results from the Angoff and Schrader study is of value. A problem with methodological issues like those just raised is that one is never quite certain what the extent of their effects may be on the results. Consideration of the results may provide some indication of the seriousness of the methodological concern. Also, the results reported by Angoff and Schrader raise other issues not implicated in the methods, but of importance to practitioners and future research.

Two issues related to the interpretation of the data provided in Angoff and Schrader will be discussed: 1) influence of directions on the number of test items attempted; and 2) use of formula scores as an indicator of the relative merits of the Invariance and Differential Effects hypotheses. Angoff and Schrader (1984) present results that they contend support the position that items answered under rights directions that would be left blank under formula directions are for all practical purposes answered at chance success levels (see Table 2 in Angoff and Schrader). They report the means and standard deviations of the rights scores, formula scores and number of items omitted and show that formula scores, computed under both directions, are almost identical. Table 2 shows the various means reported by Angoff and

Schrader but, in addition, the mean number of items reached (RCH)[2] is included.

Insert Table 2 about here

In ALL CASES the mean RCH was greater under rights directions than under formula directions. This suggests that for speeded tests students' scores will be based on a larger sample of performance under rights directions than will be the case under formula directions.

Angoff and Schrader compared the formula scores obtained under rights and formula directions to assess the relative merits of the Invariance and Differential Effects hypotheses. While this seems like a very pragmatic approach, there are two reasons why it may not be the most sensitive method of evaluating the validities of the Invariance and Differential Effects hypotheses.

First, formula scores treat omitted items and items students failed to reach identically. Thus, two students with the same formula score could have a dramatically different distribution of omitted and not reached items, particularly if the test is difficult for the group. For example, suppose a test had 50 items and two students both had 25 right and 5 wrong, but one had 20 omitted items and the other had 20 not reached items. These results could have dramatically different meaning regarding the students' ability since the first student considered all 50 items while the second only considered 30 items. Had the latter student considered all 50 items his/her score may have markedly improved. The second reason is that formula scores tend to obscure

[2] The number of items students reached (RCH), is equal to the difference between the total number of items on the test and the number of items not reached (NR). Angoff and Schrader define NR as the number of items left unmarked beyond the last item marked.

the success rates on unanswered items, since the items left unanswered are usually few in number compared to the total number of items on a test. It seems that the fundamental issue in the Differential Effects versus Invariance hypotheses controversy lies in whether or not success rates on items omitted or not reached under formula directions would exceed chance levels. This leads to the next major section in which two success rate indices will be proposed and data in Angoff and Schrader will be re-analyzed to estimate these success rates.

Section II: Re-analysis of the Angoff and Schrader data

The matter at issue in the Differential Effects/Invariance controversy is whether or not examinees would respond at chance success levels on items they would leave unanswered under formula directions. Performance at or below chance levels would support the Invariance hypothesis while performance beyond chance levels would support the Differential Effects hypothesis. The ideal approach to testing the relative validities of the two hypotheses would be to have two sets of test results for each examinee: one based on the test administered under formula directions and the other based on rights directions. Then one could obtain various success rates, including the success rate for all items examinees left unanswered under formula directions that they answered under rights directions (SR_{NA}), the success rate for items examinees did not reach under formula directions, but reached under rights directions (SR_{NR}), and the success rate for items omitted under formula directions, but answered under rights directions (SR_{O}). This discussion will limit itself to the first two success rates SR_{NA} and SR_{NR}.

The formulas for computing SR_{NA} and SR_{NR} would then appear as in equations 1 and 2.

1. $SR_{NA} = (R_O + R_{NR})/(N_O + N_{NR}) \times 100$

2. $SR_{NR} = R_{NR}/N_{NR} \times 100$

Where

R_O = # right under rights directions on items omitted under formula directions

R_{NR} = # right under rights directions on items not reached under formula directions

N_O = # items omitted under formula directions but answered under rights directions

N_{NR} = # items not reached under formula directions but reached under rights directions.

The values are multiplied by 100 in order to express the success rates as percentages.

Unfortunately, it is almost never possible to administer an operational test under two different administrative conditions to the same examinees. The examinees would not tolerate such a study. Previous studies have attempted to approach this ideal by administering the test under formula directions and then asking the examinees to answer all unanswered items using a different colored pencil (see Sherriffs and Boomer, 1954; Slakter et al. 1968, Cross and Frary, 1977). However, as Angoff and Schrader (1981) note; such a methodology suffers from the examinees having more time to work on the unanswered items than they had when the items were left blank. Given this additional time, it is possible that the examinees would have answered the items.

Although the ideal cannot be had, it is possible to use the means for the different groups reported in Angoff and Schrader to estimate SR_{NA} and SR_{NR}. In order to make these estimates, it is necessary to make the following assumptions:

1. The groups receiving the different directions in Angoff and Schrader were randomly equivalent;
2. All else being equal, examinees attempt more items in a given time period under rights directions than under formula directions; and
3. Under rights directions, the number of items answered correctly (R_R) is equal to the sum of three quantities: a) the number of items that would have been answered correctly under formula directions (R_F); b) the number correct on items that would have been answered under rights directions but omitted under formula directions (R_O); and c) the number correct on items that would have been answered under rights directions but would not have been reached under formula directions(R_{NR}).

Equation 3 shows this third assumption in equation form.

3. $R_R = R_F + R_O + R_{NR}$,

These are actually quite plausible assumptions. The support for the first assumption rests with the method in which examinees were assigned to the two administration conditions. In both phases of the study, the tests containing the various directions were "spiraled." That is, since there were six different sets of test booklets (in Phase I), the booklets were "in the order: 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, . . ., and the distribution of the books to the students in that order with the result that every sixth student received the same book" Angoff and Schrader (1981, p. 21). Angoff and Schrader state further that "the groups formed with this method of sampling were more nearly equivalent than would have been obtained with random sampling methods" (p. 22). A corresponding spiralling occurred in Phase II.

The support for the second assumption, that examinees answer more items under rights directions, can be based on both logical grounds and empirical results. The logical argument for this assumption is that under rights directions, the examinees have one less decision to make than they do under formula scoring (i.e., whether making a response is worth the risk). With fewer decisions to make, it should take examinees less time to complete a test under rights directions. The empirical support for this assumption is provided by the additional analysis of the Angoff and Schrader data shown in Table 2. For all tests in both Phase I and Phase II, more items were omitted and not reached under formula directions than under rights directions.

The support for assumption 3 (equation 3) also has both logical and empirical support. The logic of the third assumption derives from its relationship to the second assumption. If examinees answer more items under rights directions than under formula directions, given everything known about multiple choice testing, it is to be expected that examinees will have higher number right scores under rights directions. The only issue is how much higher. If the Invariance hypothesis holds, number right scores under rights directions would be higher by chance expectation. If the Differential Effects hypothesis holds, the scores would be higher by more than chance expectation.

It is reasonable to presume further that items students would answer correctly under formula directions would also be answered correctly under rights directions. Thus, if the number right scores are higher under rights directions, the increase is most likely to come from responses to items that were either omitted or not reached under formula directions.

Empirical support for assumption 3 can be found by examining the number right scores in Table 2. With only one exception (GMAT-Practical Business Judgment), the mean number right scores for examinees taking the test under rights directions were higher than those for examinees taking the test under formula directions.

Now, given the three assumptions, it is possible to use the group means reported by Angoff and Schrader to compute estimates of SR_{NA} and SR_{NR} (estimates will be denoted by bold-face print). To avoid a lengthy digression, the derivation of these estimates can be found in Appendix A.

Equation 4 shows the formula for $\mathbf{SR}_{NA}$ modified for using estimates from Angoff and Schrader.

$$4. \quad \mathbf{SR}_{NA} = \begin{cases} (\mathbf{R}_R - \mathbf{R}_F)/(\mathbf{N}_O + \mathbf{N}_{NR}) \times 100, \text{ if } (\mathbf{R}_R - \mathbf{R}_F) \text{ and } (\mathbf{N}_O + \mathbf{N}_{NR}) \text{ are both} > 0. \\ -[(\mathbf{R}_R - \mathbf{R}_F)/(\mathbf{N}_O + \mathbf{N}_{NR})] \times 100, \text{ if } (R_R - \mathbf{R}_F) \text{ and } (\mathbf{N}_O + \mathbf{N}_{NR}) \text{ are both} < 0. \\ \text{Not Applicable, if } (\mathbf{N}_O + \mathbf{N}_{NR}) = 0. \end{cases}$$

Where $\mathbf{R}_R$ is the mean right for the rights directions group, $\mathbf{R}_F$ is the mean right for the formula directions group, $\mathbf{N}_O$ is the mean number of items omitted by the formula group minus the mean number omitted by the rights group. $\mathbf{N}_{NR}$ is the mean number of items not reached by the formula group minus the mean number not reached by the rights directions group.

Under the Invariance hypothesis, $\mathbf{SR}_{NA}$ would be expected to be 20%--the reciprocal of the number of options for each item (all items had 5 options). However, it would not be unusual to see $\mathbf{SR}_{NA}$ values as low as 10% because, as Lord has noted, lower asymptotes estimated from latent trait studies have often had values less than the expected chance rate. If success rates appreciably exceed the 20% chance success rate, support for the Differential Effects hypothesis would be indicated.

However, negative values for $\mathbf{SR}_{NA}$ would seem extraordinary--particularly in light of the large number of subjects involved in the study--and should serve as a caution. Negative values can arise in only two ways: if $(\mathbf{N}_O + \mathbf{N}_{NR})$ is negative or if $(\mathbf{R}_R - \mathbf{R}_F)$ is negative. Negative quantities for

$(\mathbf{N}_O + \mathbf{N}_{NR})$ would indicate that the formula directions group answered more items than the rights directions group. This never occurred in the Angoff and Schrader study and would be an illogical outcome. Since some students taking the tests under rights directions omitted items in the Angoff and Schrader study it is possible that some students will omit items even under the strongest guessing instructions with rights directions; however, to have them leave more items unanswered than the formula group would not make sense. This would suggest that the two groups were not operating with the same motivational conditions or at least that the rights group was not operating with as high a motivation as the formula group. Because one would expect more items to be answered, one would also expect the # right to be greater under rights directions. Negative values for $(\mathbf{R}_R - \mathbf{R}_F)$ would therefore also be counter-intuitive. (However, it did occur for one subtest in the Angoff and Schrader study--the GMAT Practical Business Judgment subtest. This result will be discussed in more detail in a later section.)

From equations 1-4 and data supplied in Angoff and Schrader an estimate of SR_{NR} can be obtained if one substitutes the chance performance values for omitted items into $\mathbf{SR}_{NA}$. (Appendix A shows the derivation of this estimate.) Equation 5 shows the formula for $\mathbf{SR}_{NR}$ modified for using estimates from Angoff and Schrader.

$$5. \quad \mathbf{SR}_{NR} = \begin{cases} (\mathbf{R}_{NR}/\mathbf{NR}) \times 100, \text{ if } \mathbf{NR} \text{ is } > 0, \\ -[(\mathbf{R}_{NR}/\mathbf{NR}) \times 100, \text{ if both } \mathbf{R}_{NR} \text{ and } \mathbf{NR} \text{ are } < 0, \text{ and} \\ \text{Not Applicable, if } \mathbf{NR} = 0. \end{cases}$$

Under the Invariance hypothesis, $\mathbf{SR}_{NR}$ should function similar to SR_{NA}. However, values greater than 1.00 would seriously call into question the validity of using chance values for performance on omitted items. More likely, it would mean that success rates on both the omitted items and not-reached items were in excess of chance.

If the success rate is negative, it indicates that the formula directions group either answered more items correctly than the rights directions group or they reached more items than the rights directions group. Either case would seem irregular for reasons corresponding to SR_{NA}.

Table 3 shows the percent correct for the groups taking the tests under both formula and rights directions, and SR_{NA} and SR_{NR} for each of the tests for Phase I.

INSERT TABLE 3 ABOUT HERE

Note that the tests were rather difficult for the examinees. The mean correct never exceeded 50 percent and went as low as 35.8 percent on the Chemistry Achievement test.

The two success rate indices both show success rates beyond chance for omitted and/or not-reached items. The second index is especially enlightening as it shows estimated mean success rates on items not reached under formula directions that were greater than the mean percent correct for four of the five cases. For part 2 of the Verbal subtest SR_{NR} is in excess of 100 percent indicating that the use of chance values for the omitted items is suspect.

To explore the relationship between the effects of the directions and ability, Angoff and Schrader report formula scores on part 2 of the test broken down by score ranges on part 1 of the test. Table 4 shows a similar breakdown for percent correct and the two success rate indices.

INSERT TABLE 4 ABOUT HERE

SR_{NA} was greatest for the lowest ability group for both sets of groups. This finding suggests that the lowest ability examinees have a comparatively high rate of success on the items that they would leave unanswered under formula directions. With the exception of the second highest ability group, the SR_{NR} values were also far beyond chance levels. Thus, if the Invariance hypothesis is accepted for the items that students omit under formula directions, the student appears to be penalized fairly severely by virtue of not reaching test items on which they would achieve a fairly high degree of success.

Table 2 shows subtest statistics for the five GMAT subscores in Phase II. Differences in the mean right scores between the rights and formula groups ranged from 0.13 items (Practical Business Judgment) to 0.68 item (Problem solving). For the Practical Business Judgment Subtest, the mean number right for the formula directions group exceeded that for the rights directions group. However, the difference was not at all large (.13). As might be expected, the mean number of omits and items not reached were very few and were almost identical for the two groups (0.64 item and 0.57 item respectively for the formula and rights directions groups). Of all the subtests studied, this was the least speeded based on the number of items not reached and the very small number of items omitted. Thus, the difference in the number of items either omitted or not reached was so small as to suggest any differences in the success rates studied would be spurious.

Table 5 shows the mean percent correct and the two success rate indices for the GMAT subtests.

INSERT TABLE 5 ABOUT HERE

As was the case with the College Board Tests, the GMAT tests were difficult for the examinees. Mean percent correct scores ranged from 36% to 59% for the various subtests. However, unlike the findings from the College Board Tests, the success rate indices were at best at chance success levels. With the exception of the Practical Business Judgment subtest, SR_{NA} ranged from 13.6% (Data Sufficiency) to 20.7% (Sentence Correction). The values for SR_{NR} were similar to those for SR_{NA} with the exception of the Data Sufficiency Subtest for which it was -38.8%. The value was negative because the expected chance difference in rights scores based on differences in the number of omitted items was greater than the actual difference in rights scores.

In contrast to the findings for the SAT-verbal, results from the GMAT test showed, at best, chance levels of success on items the rights directions group answered that the formula group did not. To examine this result in more detail, the percent correct and success rate indices for each experimental subtest except the anomalous Practical Business Judgment subtest were computed for different subgroups based on scores on the corresponding operational segment of the GMAT (recall that the experimental subtest was an alternate form of one of the six subtests that comprised the operational portion of the GMAT).

For each subtest, the success rate estimates ranged from negative to positive for the various performance levels. The least variable subtest was Reading Comprehension with SR_{NA} values ranging from -4.9% to 3.43% and SR_{NR} values ranging from -97.3% to 53.7%. The most variable subtest was Sentence Completion with SR_{NA} values ranging from -88.2% to 68.4% and SR_{NA} values ranging from -3660% to 115.2%. Unlike in phase I, the GMAT results showed no discernable pattern across the subtests that would suggest that success rates for any of the different subgroups were systematically different from any other subgroup. Thus, contrary to the results from the College Board examinations, the results from the GMAT tended to support the Invariance hypotheses.

Discussion

Why is there such a discrepancy in the results obtained from the two phases of the study and why do the results from the College Board phase conflict with the formula scores interpreted by Angoff and Schrader?

At one level, one might argue that the mere difference in the examinee populations might explain the difference in the results obtained in Phase I and II. The examinees in Phase II were college graduates in business who were vying for admission to graduate school. Thus, they are likely to be highly motivated and have extraordinarily sophisticated examination taking skills. This group might know very well when they do and do not have sufficient knowledge to guess on a test item. On the other hand, the College Board group consisted of high school juniors who were planning to go on to college after graduation. In addition, care had been taken during the school selection process to over-sample from minority populations--a group that is not noted for high levels of test taking skills. Thus, this group is almost certainly far less sophisticated in their test taking skills than the Phase II group.

At another level, one might argue that the test administration conditions were the cause of the difference. In Phase I, it was made perfectly clear that participation was voluntary in both the school selection process and the directions to students. In Phase II, participation was made to appear as much as possible to be part of the operational test. However, there is good reason to believe that at least some students may have realized that the experimental subtest was not part of the regular test--particularly those students under rights directions. Specifically, since the operational test was given under formula directions, it seems likely that students encountering a rights directions subtest as the last section would be clued, in no little degree, that the subtest was experimental. Thus, some of the GMAT examinees may have perceived themselves as being conscripted volunteers. Such examinees may not have given as careful consideration to their responses as examinees who did not come to this realization. Because the rights directions group would be more likely to be clued to the experimental nature of the last subtest, there may be a systematic bias in the results. This may account for the negative indices for the Practical Business Judgment Subtest and the below chance level performance on omitted and not-reached items for four of the five subtests.

Another explanation may be the differences in the fatigue level between the examinees. The College Board examinees took only the experimental tests. The GMAT examinees encountered the experimental test after 2 1/2 hours of taking a grueling standardized test. It is without doubt that the latter group was suffering from greater mental fatigue. Under mental fatigue, examinees may not be as alert to making fine distinctions among responses as they are when they are fresh. Thus, plausible distractors may be more attractive. This might explain the lower than chance performance on omitted and not reached items. The situation might be aggravated if the students are not only fatigued, but realize they are participating in an examination that will not contribute toward their test score.

A final explanation may be that treatment contamination occurred in phase II because some students failed to read the directions for the experimental subtest. If this occurred, students would take the experimental subtest under formula directions since the remainder of the test was under formula directions. While the formula group performance would be unaffected by this, the rights directions group may have been seriously affected. The comparatively high rates of omission of items by the rights directions group in phase II and the counter intuitive findings for the Practical Business Judgement Subtest (i.e., students in the formula directions group answered fewer items but had higher number right scores than the rights direction group) suggest this was a likely occurence. Thus, although phase II would seem on first inspection to be more valid than phase I because it was part of an operational test, the methodological problems cited would render its results suspect.

Another point deserving mention is that in all cases, examinees answered more questions under rights directions than under formula directions. At least for the Phase I study, the examinees answered the added items correctly at a rate beyond chance. This suggests that the scores for examinees under rights directions are likely to be based on a broader representation of content than under formula scoring. The content validity of rights directions scores would therefore seem to be greater than for formula directions scores.

In some senses, it is not surprising that it takes students more time to take a test under formula directions since students must make the added decision of whether to risk a response to items in which they are not certain which answer is correct. Under rights directions all students should answer all items if the directions are working properly. In the results presented by Angoff, a fair number of students in the rights directions group in both phases either omitted or failed to reach items. This suggests that the rights directions were not working properly. Too often, rights directions are not stated sufficiently strongly. It is the author's opinion that the directions should state "Your score on this test will be the number of items you answer correctly. Mark the best answer you can to every item even if you must guess. You put yourself at a disadvantage if you leave even one item unanswered."

At the very least, it can be concluded from the results reported in this article that examinees answer more items under rights directions than under formula directions and that performance on items that examinees answer under rights directions and would not answer under formula directions may vary depending on the sophistication of the examinee, fatigue and whether or not the examination is an operational test. While it would be interesting to study how these factors affect such scores, from a practical standpoint, it seems that the prudent approach would be to use rights directions.

The interpretations of Phase I results reported in this study differ from those reported by Angoff and Schrader. The difference clearly lies in the different indices used to make the interpretations. Which of these indices is most appropriate? It depends on the question being asked. If the question of concern is whether students would perform beyond chance levels on items they would leave unanswered under formula directions, the two success rates SR_{NA} and SR_{NR} would seem to provide the most direct answer. The interpretation of formula scores for this purpose is not as satisfactory because the number of items not attempted is small compared with the total number of items answered, with the result that formula score means and standard deviations obscure the effects of success rates on not-attempted items. If, on the other hand, the question relates to the practical consequences of using formula directions, formula scores would provide a more interpretable index than the success rates.

Thus, if one considers the results from previous studies and the success rates in Phase I, it appears that the Differential Effects hypothesis is valid. However, the practical consequences of its operation may be open to question. Angoff and Schrader's results suggest that for volunteer groups taking standardized tests, the practical consequences are negligible. It remains to be seen if the same results would be obtained in an operational testing situation.

The research into the whole issue of formula scoring has really only scratched the surface. Since a major tenet of the Differential Effects hypothesis involves risk, behavior under various risk situations needs to be explored as well as the response of different examinee groups. For instance, none of the studies of formula scoring have involved elementary school children taking either standardized achievement tests or classroom tests. Neither have similar studies been conducted with high school students, with the exception of the Angoff and Schrader study. A better understanding of the Differential Effects and Invariance hypotheses could be had if the design of the Angoff and Schrader studies could be used in an operational version of the SAT.

Table 1

Mean Percentage of Items Omitted and Not Reached

in Phase I and II

Phase	Directions	Mean Percentage omits	Mean Percentage not reached	Mean Percentage omits & not reached
I	Rights	1.96	3.06	5.02
	Formula	7.78	4.00	11.77
	Ratio*	0.25	0.77	0.43
II	Rights	7.84	5.63	13.46
	Formula	11.89	7.20	19.09
	Ratio*	0.66	0.78	0.70

*The ratio was formed by dividing the mean for the rights group by the mean of the formula group.

Table 2

Test Results

A. Phase I -- SAT

					Mean	Mean	Mean	Mean
Part	Group	# Items	Directions	Cases	Right	Formula	Omitted	Reached
			I. SAT Verbal, Form A					
1	1+2	45	Rights	2094	22.18	17.06	0.78	43.44
1	3+4	45	Formula	2092	21.65	17.04	2.86	42.95
2	1+3	40	Rights	2080	18.24	13.25	0.83	39.03
2	2+4	40	Formula	2106	17.46	12.86	2.94	38.80
Total	1	85	Rights	1026	40.47	30.30	1.41	82.56
Total	4	85	Formula	1038	38.99	29.70	5.55	81.70
			II. SAT Verbal, Form B					
Total	5	85	Rights	1040	40.30	30.08	1.34	82.52
Total	6	85	Formula	1034	39.03	29.76	5.76	81.87
			III. Chemistry Test					
Total	7	90	Rights	1151	34.50	22.06	2.49	86.75
Total	8	90	Formula	1155	32.21	21.52	10.71	85.68

B. Phase II -- GMAT

			Mean	Mean	Mean	Mean	Mean
Subtest	# Items	Directions	Cases	Right	Formula	Omitted	Reached
Reading	29	Rights	5658	15.03	12.35	1.32	27.06
Comprehension	29	Formula	5739	14.68	12.37	2.44	26.34
Problem	25	Rights	5501	9.67	7.56	4.76	22.85
Solving	25	Formula	5594	8.99	7.56	7.41	22.10
Practical	32	Rights	5738	18.82	15.67	.44	31.87
Judgment	32	Formula	5408	18.95	15.85	.50	31.86
Data	40	Rights	5590	16.38	12.31	4.31	36.98
Sufficiency	40	Formula	5657	16.18	12.42	5.62	36.82
Sentence	30	Rights	5409	17.05	14.44	1.03	28.53
Correction	30	Formula	5486	16.79	14.43	1.73	27.95

Table 3
Percent Correct and Success Rates for
College Board Tests

				Mean		
			Sample	Percent	Success Rates	
Part (Groups)	# Items	Directions	Size	Correct	SR_{NA}	SR_{NR}
A. SAT verbal, Form A						
1 (1 & 2)	45	rights	2094	49.3%	20.6%	23.3%
1 (3 & 4)	45	formula	2092	48.1%		
2 (1 & 3)	40	rights	2080	45.6%	33.3%	155.7%
2 (2 & 4)	40	formula	2106	43.7%		
Total (1)	85	rights	1026	47.6%	25.0%	75.8%
Total (4)	85	formula	1038	45.9%		
B. SAT Verbal, Form B						
Total (5)	85	rights	1040	47.4%	25.0%	59.4%
Total (6)	85	formula	1034	45.9%		
C. Chemistry Achievement Test						
Total (7)	90	rights	1151	38.3%	24.7%	60.4%
Total (8)	90	formula	1155	35.8%		

Table 4

SAT - Verbal Part II Scores Stratified

by Scores on Part I (45 items)

I. Right directions on Part I (Groups 1 and 2)

	Mean Percent Correct		Success Rates	
Score Range	Rights Directions	Formula Directions	SR_{NA}	SR_{NR}
>25	60.3	56.7	46.2	940.0
20-24	44.7	43.3	17.7	0.0
15-19	37.8	35.4	34.3	102.0
<14	30.9	27.8	55.2	87.1

II. Formula directions on Part I (Groups 3 and 4)

	Mean Percent Correct		Success Rates	
Score Range	Rights Directions	Formula Directions	SR_{NA}	SR_{NR}
>21	60.3	58.6	26.9	91.7
15-20	46.2	44.6	23.1	*
9-14	37.6	36.0	31.0	53.3
<8	32.2	28.8	82.2	171.3

* The number of items reached was greater for the formula directions group than the rights directions group.

APPENDIX A

Derivation of Estimates for SR_{NA} and SR_{NR}

I. Derivation of SR_{NA}—The success rate of examinees taking the test under rights directions on items examinees omitted and/or failed to reach under formula directions.

To obtain an estimate of SR_{NA}, estimates are necessary for the two values $(R_O + R_{NR})$ and $(N_O + N_{NR})$. Ideally, N_O would be estimated by simply using the mean number of items omitted by the formula directions group because examinees taking the test under rights directions should omit no items. A similar estimate would be ideal for N_{NR}. However, the rights directions group omitted some items and failed to reach some items on every subtest in both Phases. Thus, it is necessary to subtract the values derived from the rights directions group from the estimates obtained from the formula group in order to adjust for this less than ideal situation. Therefore, N_O is estimated by subtracting the mean number of items omitted for examinees taking the test under rights directions ($N_{O/R}$) from the mean number of items omitted by examinees taking the test under formula directions ($N_{O/F}$). Similarly, N_{NR} can be estimated by subtracting the mean number of items reached by examinees taking the tests under rights directions ($N_{NR/R}$) from that of the formula directions group ($N_{NR/F}$). These quantities are all provided in Angoff and Schrader.

The mean values for R_O and R_{NR} are not provided; however, it can be shown from equation 3 that

A-1 $\quad (R_O + R_{NR}) = R_R - R_F$

where

R_O = mean right under rights directions on items that would be omitted under formula directions and

R_{NR} = mean right under rights directions on items that would not be reached under formula directions

Thus, the numerator of SR_{NA} can be estimated by subtracting the mean number correct for the formula directions group (R_F) from the mean number correct for the rights directions group (R_R). Equation A-2 shows SR_{NA}.

$$\text{A-2} \quad SR_{NA} = \begin{cases} (R_R - R_F)/(N_O + N_{NR}) \times 100, \text{ if } (R_R - R_F) \text{ and } (N_O + N_{NR}) \text{ are both } > 0, \\ -[(R_R - R_F)/N_O + N_{NR})] \times 100, \text{ if } (R_R - R_F) \text{ and } (N_O + N_{NR}) \text{ are both } < 0, \text{ and} \\ \text{Not Applicable, if } (N_O + N_{NR}) = 0 \end{cases}$$

where,

$R_R - R_F$ = the mean number right for the rights directions group minus the mean number right for the formula directions group.

$$N_O + N_{NR} = (N_{O/F} - N_{O/R}) + (N_{NR/F} - N_{NR/R})$$
$$= (N_{O/F} + N_{NR/F}) - (N_{O/R} + N_{NR/R})$$

The term ($N_O + N_{NR}$) is algebraically equivalent to a more intuitively appealing quantity, the mean number of items answered under rights directions minus the mean number answered under formula directions.

II. Derivation of SR_{NR}--The success rate of examinees taking a test under rights directions on items examinees failed to reach under formula directions.

In equation (A-1), while both R_O and R_{NR} are unknown, the mean number of omitted items for the rights directions group and formula directions group are reported (O_R and O_F respectively). If one substitutes chance values for omtted items, the expected mean right on omitted items should be equal to the difference in the mean number of omitted items divided by 5 (the number of options/item).

A-2 $R_O = (O_F - O_R)/5.$

Equation A-3 shows the formula for estimating R_{NR} in which the R_O estimate consists of omits responded to at chance success levels.

A-3 $R_{NR} = (R_R - R_F) - [(O_F - O_R)/5]$

In equation A-3, O_R is subtracted from O_F because examinees omitted items under rights directions (albeit fewer than under formula directions) in spite of directions to the contrary.

Equation A-4 shows SR_{NR}.

$$\text{A-4}\quad SR_{NR} = \begin{cases} (R_{NR}/NR) \times 100, \text{ if } NR \text{ is} > 0 \\ -\,[(R_{NR}/NR) \times 100], \text{ if both } R_{NR} \text{ and } NR \text{ are} < 0 \\ \text{Not Applicable, if } NR - 0 \end{cases}$$

where

$NR = NR_F - NR_R$, and

NR_F = # of items not reached under formula directions, and

NR_R = # of items not reached under rights directions.

Determining an estimate for SR_O (success rate on omitted items) by substituting chance values for R_{NR} would not be appropriate as it would not be reasonable to assume examinees would perform at chance levels on items for which they had never given consideration.

REFERENCES

Angoff, W.H., & Schrader, W.B. (1981). A study of alternative methods for equating rights scores to formula scores (Research Rep. No. 81-8). Princeton, NJ: Educational Testing Service.

Angoff, W.H., & Schrader, W.B. (1984). A study of hypotheses basic to the use of rights and formula scores. Journal of Educational Measurement, 21, 1-17.

Cross, L.H., & Frary, R.B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 14, 313-321.

Sherriffs, A.C., & Boomer, D.S. (1954). Who is penalized by the penalty for guessing? Journal of Educational Psychology, 45, 81-90.

Slakter, M.J. (1968b). The effect of guessing strategy on objective test scores. Journal of Educational Measurement, 5, 217-226.