FORMULA SCORING, NUMBER-RIGHT SCORING, 
AND TEST-TAKING STRATEGY

GLENN L. ROWLEY
La Trobe University
ROSS E. TRAUB
The Ontario Institute for Studies in Education

Opponents and proponents of the "correction for guessing" have rarely been able to convince one another of the wisdom of their respective beliefs. Lord (1975) noted that formula scoring, like religion and politics, is an area "where two informed people often hold opposing views with great assurance" (p. 7). The present paper represents an attempt to resolve these differences by focussing on the assumptions typically made about the behavior of students in an examination situation. Attention will also be given to the value judgments which are made, but rarely spelled out, in defending one position or another.

WHY DISCOURAGE GUESSING?

If the purpose of formula scoring is to discourage examinees from guessing at questions to which they do not know the answer, then it is important to consider why such guessing should be discouraged. There appear to be two possible reasons. The first involves a value judgment; it could be argued, for instance, that an examinee who responds with a guess is trying to deceive the examiner by concealing his lack of knowledge, and that such behavior needs to be penalized whenever it can be detected. A similar argument could be made on the grounds that to reward successful guessing has the effect (even if unintended) of encouraging the examinee to gamble, and the value judgment (that gambling is an undesirable behavior) is used to justify its discouragement by the use of a "guessing penalty." The weakness of such a justification lies in its total dependence on the initial value judgment. If that value judgment is rejected out of hand (as it almost certainly will be, even by many of the strongest supporters of formula scoring), then the argument has no further force. ¹

Secondly, the use of a guessing penalty can be justified on the grounds that it improves the psychometric properties of the test (or, more properly, of the scores obtained on the test). It can be argued on theoretical grounds that the use of formula scoring should increase both reliability (Mattson, 1965) and validity (Lord, 1963). However empirical studies in this area have (understandably) not been unanimously in agreement. The differences being sought are typically small (.05 or less), and while some researchers have found that the instruction to guess results in decreased reliability (Blommers & Lindquist, cited in Ebel, 1965; Keislar, 1953; Ruch & De Graff, 1926; Swineford & Miller, 1953), other results (e.g., Sabers & Feldt, 1968; Traub, Hambleton, & Singh, 1969) indicate that this is not always the case.

The effect of "guess" versus "do-not-guess" instructions on validity is considerably

¹Arguments like these have greatest force when applied to testing situations in which the examinee stands to benefit from an open revelation of ignorance. An example is diagnostic testing where guessing interferes with the identification of weaknesses in learning.
less clear, and it is extremely doubtful that the typical “validity study” (in which different sets of scores are correlated with a criterion measure; e.g., Hakstian & Kansup, 1975) can be of much help in this regard. This is because a major concern about validity relates to the unknown extent to which personality factors influence scores obtained under penalty instructions. Research in this area has produced results which provide reasons for concern, rather than answers and prescriptions. Votaw (1936) found that the use of the “do-not-guess” instruction favoured students having ascendant, as opposed to submissive personalities; the ascendant students, it seems, profiting from their lesser willingness to omit items. Wiley and Trimble (1936) found not only that a personality trait (confidence in responses) was influencing scores, but that tests scored with this intent measured the personality trait more reliably than they measured knowledge of subject matter. Soderquist (1936) found that reliability of scores could be increased by confidence weighting. Swineford (1938) used Soderquist’s weighting scheme to construct “gambling scores” which were uncorrelated with achievement, and more reliable than either the weighted or the unweighted achievement scores. Sherriffs and Boomer (1954) found that the use of a guessing penalty conveys an extra disadvantage upon students who are “characterized by introversion, rumination, anxiety, low self-esteem, and undue concern with the impression they make on others.” On these grounds they argued that the use of the penalty for guessing could no longer be justified. Slakter’s (1968) study indicated that the use of modified “do-not-guess” instructions (“Omit those items that you could answer only by pure guess . . .”) penalized those examinees whose tendencies to take risks in the examination situation were lowest.

Such studies raise questions about validity which cannot be answered by the traditional criterion-related approach. If, indeed, the choice of scoring method and instructions about guessing can result in scores which are influenced substantially by personality factors which it was not intended to measure, then the nature of the criterion used becomes all-important. A criterion variable contaminated by the same unwanted personality factors may be expected to yield an enhanced validity coefficient, but this would not be evidence of validity in the fundamental sense of the test “measuring what it is intended to measure.” Indeed, a judicious choice of criterion measures could surely produce results favoring whatever method of scoring and administration the investigator set out to defend. The fact that the criterion measure may be ill-defined and its scoring method unspecified (e.g., GPA) can only make the findings of traditional validity studies more difficult to interpret.

Considerations such as these cast doubt also on the usefulness of comparative reliability studies. A finding that formula scoring yields increased reliability can be of little comfort to us if it is suspected that the scores measure, at least partially, an unwanted (but reliable) personality factor. Viewed in this light, the findings of Wiley and Trimble (1936) and Swineford (1938) provide food for thought; each found that achievement tests could be used to measure personality traits associated with gambling behavior more reliably than they measured achievement. If, indeed, the use of “do-not-guess” instructions, with accompanying penalties, does result in the test measuring something quite distinct from what it was intended to measure, then enhanced reliability could be an expected result.

In summary, the question of whether to encourage or to discourage guessing has yet to be satisfactorily resolved. The dilemma one faces in connection with this decision may be summarized as follows: if one encourages students to answer all questions,
whether they know the answer or not, a source of random variance is introduced (Lord, 1963) which decreases both reliability and validity; on the other hand if one attempts to discourage students from guessing, it is apparent that some students will comply to a greater extent than others, causing the test results to be contaminated by personality factors which the test was not intended to measure.

UNDERLYING ASSUMPTIONS

The assumption behind the derivation of standard formula scoring is both well known, and widely disbelieved. It is assumed that, on any given question, an examinee has either complete knowledge, or total ignorance. If (s)he has complete knowledge, then, naturally (s)he will check the correct answer; otherwise (s)he will either omit the question, or guess randomly amongst the alternatives offered. This assumption rules out the possibility that examinees sometimes answer on the basis of partial information (e.g., by ruling out as impossible one or more of the distractors), or from misinformation (choosing a distractor in the genuine belief that it is correct). To the extent that partial information is present, and being used, the scoring formula will under-correct, since a candidate’s guessing will be more successful than random choices; to the extent that misinformation is present, it will over-correct, since the candidate loses not only the point for the question about which (s)he is misinformed, but a fraction of another from questions for which (s)he is entitled to credit. Whether the formula over-corrects or under-corrects will be a function of the relative amounts of partial information and misinformation used by the examinee, and will vary not only from test to test, but also from examinee to examinee on the same test.

Lord (1975) argued that the obvious falsehood of the assumption need not invalidate the conclusion derived from it. He then proceeded to outline an alternative assumption which leads to the same conclusion. His assumption is “that the difference between an answer-sheet obtained under formula-scoring directions . . . and the same answer sheet obtained under number-right scoring directions . . . is only that omitted responses, if any, on the former answer-sheet are replaced by random guesses on the latter” (Lord, 1975, p. 8). Under this assumption, it is demonstrated that the formula score and the number-right score are unbiased estimators of the same quantity, and it is argued that the formula score is to be preferred because it has a smaller sampling variance. In situations where the assumption accurately describes examinee behavior, the conclusion is unchallengeable.

A good example of such a situation would probably be a highly speeded test. Consider two examinees of equal ability, who work at the same pace, and run out of time after having completed the same number of items. Examinee A turns in his paper without responding to the remaining items, while examinee B responds randomly to them. Then the two sets of responses would differ only in that omits (in the case of examinee A) would be replaced by random responses (in the case of examinee B), and Lord’s conclusion would most certainly be justified. But if speed is to be an important factor on the test, one could hardly contemplate not correcting for guessing, since (in the above example) it would give examinee B an unfair advantage over examinee A. What is a matter of debate is the use of formula-scoring when the test is primarily a power test, and it is for unspeeded tests that Lord believes his assumption to hold best (Lord, 1975, p. 8).
EXAMINEE STRATEGY

Underlying Lord’s assumption there seems to be the belief that examinees are capable of accurately distinguishing informed from random guesses; a belief which is not supported by research results reported to date. Contrary findings (Sherriffs & Boomer, 1954; Slakter, 1968; Votaw, 1936) were discounted on the grounds that the examinees were given only information about the scoring method, but no guidance as to the proper strategy for responding. Results reported in passing by Rowley (1972) may help to throw some light on this question. Responses were obtained from 177 grade nine students to an 80-item, 4-alternative, multiple-choice test. Students were asked to claim 1, 2, or 3 marks for each question, according to their degree of certainty in responding. Penalties of 0, 1, 2 marks respectively were imposed for wrong answers. Instructions to the students included the following advice on strategy:

What is the best strategy to use in a test like this?
If you think you have the correct answer .......................................................... claim 3
If you are not sure, but think that one answer looks
more likely to be correct than the others ....................................................... claim 2
If you have no idea which answer is correct, and are
just guessing wildly .................................................................................. claim 1

Remember:
If you claim one mark for a question, there is no penalty if you are wrong. Therefore there is no
point in leaving any question unanswered. If you have no idea at all on a question, guess the
answer and claim 1 mark only. But ANSWER EVERY QUESTION. (Rowley, 1972, p. 92).

The numbers of responses falling into each category and the numbers correct within
each, are given in Table 1.

In the context of this discussion, we note that 40.5 percent of the 4595 responses

<table>
<thead>
<tr>
<th>Degree of certainty</th>
<th>Marks claimed</th>
<th>Number of responses</th>
<th>Percentage of all responses</th>
<th>Number of correct responses</th>
<th>Percentage of responses correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Confident</td>
<td>3</td>
<td>5878</td>
<td>41.51</td>
<td>4587</td>
<td>78.0</td>
</tr>
<tr>
<td>Uncertain</td>
<td>2</td>
<td>3601</td>
<td>25.43</td>
<td>2034</td>
<td>56.5</td>
</tr>
<tr>
<td>Guessing</td>
<td>1</td>
<td>4595</td>
<td>32.45</td>
<td>1859</td>
<td>40.5</td>
</tr>
<tr>
<td>No response</td>
<td>–</td>
<td>86</td>
<td>0.61</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>TOTAL</td>
<td></td>
<td>14160</td>
<td>100.00</td>
<td>8480</td>
<td>59.9</td>
</tr>
</tbody>
</table>
marked "1" were in fact correct, compared to the 25 percent which would be expected under the presumption of random guessing. The small proportion of omits, and the higher proportions correct in the other two categories were consistent with the students having understood, and done their best to comply with, the instructions given. (The 86 omits came from 29 examinees, one of whom omitted the last 17 items, two the last nine, and one the last six. Seventeen of the 29 omitted one item only.) The administration of the test was supervised by the investigator, and the few questions asked were dealt with to the apparent satisfaction of the examinees concerned. The most plausible explanation seemed to be that many of the examinees were unable to distinguish pure guessing from informed guessing, and it was concluded that "to have asked these students . . . to omit the items for which they thought they could only make random guesses would have resulted in the omission of items about which they apparently did have some knowledge" (Rowley, 1972, p. 44).

Of course, it must be conceded that examinees could have completely ignored the advice given (orally, as well as on the test form) and planned their strategies entirely on the basis of the rewards and penalties offered. Such a strategy would have led to an examinee claiming one mark on any item upon which his probability of answering correctly was judged to be less than 0.5, and the data are consistent with this possibility.

However, it is not contended that this study is conclusive in any sense, since no study could in fact be conclusive about this question. Along with the four studies cited by Lord (1975) and a recent study by Crehan, Candor and Beckett (1976) these results are consistent with the proposition that many examinees can perform at much higher than chance level, on items for which they consider themselves to be "just guessing." This does not prove that Lord's assumption is universally untrue (indeed no study, or studies could). What must be realized is that Lord's justification of formula-scoring is only valid if the assumption is universally true. If there are even a few examinees whose judgment is faulty, and who omit items upon which they could have done better than guess at random, then these examinees will be placed at a disadvantage. To the extent that this happens, the validity of the test will be lessened, since it will be measuring, at least in part, something quite distinct from what was intended.

**THE TEACHING OF EXAMINEE STRATEGY**

Lord suggests (1975, p. 8) that it is probably time that children in school were taught how to behave effectively when taking a test. Without disputing this point, we may note that it would provide the conditions under which formula-scoring was justified only if the instruction was perfectly effective for all of the students. Experience suggests that this may be difficult to achieve. (The authors dearly wish that their own teaching met with this degree of success, not least because all of their students would then be relieved of the temptation to guess, making questions of scoring method and guessing strategy irrelevant.)

But let us suspend our disbelief, at least temporarily, and consider how such an examinee would behave under formula-scoring conditions. We assume a rational, well-informed examinee, capable of planning a strategy which is in his own best interest (in the sense of yielding the highest expected score). If (s)he thought (s)he knew the answer to an item, (s)he would respond with that answer. If (s)he was unsure, but thought his probability of answering correctly was greater than $1/n$ (where $n$ is the number of al-
ternatives) the expected score from a response would be greater than that from an omit, and so (s)he would respond. Only on items upon which (s)he thought (s)he could do no better than guess randomly would (s)he even consider not responding. But then (s)he might argue as follows. Published research suggests that, in this situation, examinees typically have a probability of "guessing correctly" which is greater than 1/n. Hir own probability can only be 1/n or greater; it can never be less. The expected rewards from responding and omitting are the same only if the guess is truly random; otherwise the expected reward from responding will be the greater. Being unable to distinguish with certainty which situation (s)he is in, (s)he will reason that the expected reward for responding is probably higher, and certainly no lower than from omitting, and will respond to the item.

It is apparent, then, that examinees who had successfully learned to behave effectively in the testing situation would omit no items, even though the usual penalty is applied. If there are no omits, of course, it matters not at all which method of scoring is used, since the two yield scores which correlate perfectly. Formula-scoring can only fulfill its stated intention (the discouragement of random guessing), if examinees are persuaded to adopt a strategy which is not in their own best interests.

CONCLUSIONS

On an unspeeded test, the usual arguments in support of formula-scoring are found wanting. Empirical evidence in favour of higher validity proves very little if the criterion measure is itself (in whole or in part) a multiple-choice test. Increased reliability could come about as a result of the influence of unwanted (but reliable) personality factors, and would not provide evidence in favour of formula-scoring unless it could be demonstrated that such personality factors were not operating.

The set of assumptions from which formula-scoring is derived are so obviously wrong that they cannot be used to provide a credible justification of formula-scoring. The alternative assumption used by Lord (1975) to justify formula-scoring could only suffice if it were universally true for all examinees, and the available evidence suggests that it is not.

In the final analysis, on an unspeeded test, the choice between the use of formula-scoring and number-right scoring (like one's attitude toward religion or politics) will be decided by a value judgement. Number-right scoring has the advantage of unbiasedness, in the sense that no examinees are favoured or penalized by the irrelevant personality factors which determine their willingness to guess. Formula-scoring has the advantage of reduced error variance. Those who place a higher value on unbiasedness will favour number-right scoring, and those who place a higher value on reducing the error variance will favour formula-scoring. In practical terms, we should note that, if we opt for unbiasedness, we still have the option of reducing the error variance (in relative terms) by lengthening the test. The reverse does not apply.

A third possibility may warrant consideration as a possible avenue for future research. If we were able to estimate independently the contributions of both random error and bias (in the sense used here) to total error, we could seek out a strategy, which would minimize a quantity such as the sum of these two components, or a weighted combination of the two (de Groot, 1970, provides details of procedures for optimizing such decisions). The problem of course, is that there is, at present, no known way of inde-
pendently estimating the two quantities involved. A breakthrough in this direction
would make possible a fresh approach to a well-worn problem.
Meanwhile, we are left in much the same position as Maurice Kendall’s Hiawatha:

In a corner of the forest. . . .
Wondering in idle moments
Whether an increased precision
Might perhaps be rather better
Even at the risk of bias
If thereby one, now and then, could
Register upon the target. (Kendall, 1959, p. 24.)

REFERENCES

CREHAN, K. D., CANDOR, C. A., & BECKETT, G. W. Utilization of partial knowledge of ob-
jective exams under formula- and number-right scoring directions. Paper presented at the
HAKSTIAN, A. R., & KANSUP, W. A comparison of several methods of assessing partial
knowledge in multiple-choice tests: II. Testing procedures. Journal of Educational Meas-
urement, 1975, 12, 231–240.
KEISLAR, E. R. Test instructions and scoring method in true-false tests. Journal of Exper-
LORD, F. M. Formula-scoring and validity. Educational and Psychological Measurement, 1963,
23, 663–672.
LORD, F. M. Formula-scoring and number-right scoring. Journal of Educational Measurement,
1975, 12, 7–12.
MATTSON, D. The effects of guessing on the standard error of measurement and the reliability
ROWLEY, G. L. Which examinees are most favoured by the use of multiple-choice tests?
RUCH, G. M., & DE GRAFF, M. H. Corrections for chance and ‘Guess’ vs. ‘Do not Guess’
instructions in multiple-response tests. Journal of Educational Psychology, 1926, 17, 368–375.
SABERS, D. L., & FELDT, L. S. An empirical study of the effect of the correction for chance
success on the reliability and validity of an aptitude test. Journal of Educational Measurement,
1969, 6, 93–96.
SHERIFFS, A. C., & BOOMER, D. S. Who is penalized by the penalty for guessing? Journal
of Educational Psychology, 1954, 45, 81–90.
SLAKTER, M. J. The effect of guessing strategy on objective test scores. Journal of Educational
SODERQUIST, H. O. A new method of weighting scores in a true-false test. Journal of Educa-
tional Research, 1936, 30, 290–292.
SWINEFORD, F. The measurement of a personality trait. Journal of Educational Psychology,
1938, 29, 295–300.
SWINEFORD, F., & MILLER, P. M. Effects of directions regarding guessing on item statistics
TRAUB, R. E., HAMBLETON, R. K., & SINGH, B. Effects of promised reward and
threatened penalty on performance on a multiple-choice vocabulary test. Educational and

WILEY, L. N., & TRIMBLE, O. C. The ordinary objective test as a possible criterion of certain personality traits. *School and Society*, 1936, 43, 446–448.

AUTHORS
