The Effect of Misinformation, Partial Information, and Guessing on Expected Multiple-Choice Test Item Scores

Robert B. Frary
Virginia Polytechnic Institute and State University

Six response/scoring methods for multiple-choice tests are analyzed with respect to expected item scores under various levels of information and misinformation. It is shown that misinformation always and necessarily results in expected item scores lower than those associated with complete ignorance. Moreover, it is shown that some response/scoring methods penalize all conditions of misinformation equally, and others have varying penalties according to the number of wrong choices the misinformed examinee has categorized with the correct choice. One method exacts the greatest penalty when a specific wrong choice is believed correct; two other methods provide the maximum penalty when the examinee is confident only that the correct choice is incorrect. Partial information is shown to yield substantially different expected item scores from one method to another. Guessing is analyzed under the assumption that examinees guess whenever it is advantageous to do so under the scoring method used and that these conditions would be made clear to the examinee. Additional guessing is shown to have no effect on expected item scores in some cases, though in others it is shown to lower the expected item score. These outcomes are discussed with respect to validity and reliability of resulting total scores and also with respect to test content and examinee characteristics.

An examinee may be defined as misinformed on a multiple-choice test item whenever the (single) correct choice has been excluded from consideration. For example, the examinee may simply believe that one of the incorrect choices is correct, having by default categorized the correct choice with the remaining wrong choices. In contrast, the examinee may categorize the correct choice with one or more incorrect choices and guess among two or more remaining incorrect choices. Clearly, these differing circumstances represent some variation in examinee knowledge. Moreover, different scoring and response methods yield markedly different expected item scores for the same misinformation condition.

It will be shown that the effect is *always* to penalize for misinformation as compared with total ignorance, in which the examinee has no basis for categorizing any of the choices as incorrect. Some response/scoring methods penalize all conditions of misinformation equally; others have varying misinformation penalties according to the number of incorrect choices that the misinformed examinee has included in the same category with the correct choice. Some methods exact the greatest penalty when a single incorrect choice is believed correct (correct choice categorized with all remaining incorrect

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 1 Winter 1980 pp. 79–90

choices). This condition was referred to by Davis (1964) as complete mininformation. Davis referred to the conditions leading to guessing among two or more incorrect choices as degrees of partial misinformation. Although partial misinformation may sound more desirable than complete misinformation, there are response/scoring methods that penalize it more heavily than so-called complete misinformation.

Depending on the response/scoring method used, an examinee misinformed on a moderate proportion of items could receive scores varying considerably from the scores of an examinee totally ignorant on the same proportion of items. This outcome has substantial implications with respect to validity. Some criterion measures may be especially sensitive to misinformation; for other criterion measures the effect may be the same as that of ignorance. Thus, there is a need for consideration of misinformation in selection of response/scoring methods for predictors.

Partial information on a multiple-choice test item is defined as the ability to eliminate some, but not all, the incorrect choices, thus restricting guessing to a proper subset of choices that includes the correct choice. It will be shown that various response/scoring methods give more or less credit for partial information relative to the expected potential score range of the test. This effect also needs to be considered with respect to validity in selection of response/scoring methods, since partial information may be more valuable with respect to some criteria than to others.

Guessing on a multiple-choice item can occur under a variety of circumstances, depending on the response/scoring method. It will be shown that in most cases, with proper attention to scaling, this phenomenon has no adverse effect on expected item scores, though some additional total score variance does result. However, there are circumstances in some response/scoring methods when guessing results in score penalties (not corrections). These situations will be outlined and the extent of the penalties derived.

The scoring/response methods covered in this paper are ordinal, i.e., methods under which the examinee marks one or more choices in an effort to identify the correct choice or, at least, to include it in a proper subset of the choices. Probabilistic response/scoring methods are not covered, though misinformation and guessing are also of concern with these methods.

Number-Right and Corrected-for-Guessing Scoring

For total ignorance the expected item score for an n-choice item under number-right (NR) scoring is $1/n$, assuming that the examinee answers every item in keeping with appropriate directions. In contrast, for any level of misinformation, the expected item score is 0, since the examinee has eliminated the correct choice from consideration. Hence, there is a uniform $1/n$ point penalty for misinformation.

In the case of scores determined under the conventional correction-for-guessing (CG), the expected item score for total ignorance is 0, even if the examinee guesses contrary to the instructions. However, under any level of misinformation, the expected item score is $-1/(n-1)$, since the item would necessarily be answered incorrectly and the correction applied. Of course, it is assumed that examinees receive and follow appropriate instructions to the effect that an item should be answered whenever any choice can be eliminated as incorrect.

Since $1/(n-1) > 1/n$, it would appear at first glance that a greater penalty is exacted for misinformation under CG scoring. Actually, the penalties are equal when the effect of the penalty on the expected range of scores is taken into consideration. For an N-item test (n choices per item), the range of expected NR scores is from N/n to N, or $N - N/n$. Under CG, the range is 0 to N, or N. Viewing each penalty as a fraction of the expected score range, they are equal:

$$\frac{1/n}{N - N/n} = \frac{1/(n-1)}{N} = \frac{1}{N(n-1)} \quad . \qquad [1]$$

Accordingly, penalties under all scoring methods will be expressed as fractions of expected total score range. As will be seen, the penalty just derived is the minimum for any of the scoring methods surveyed.

In order to analyze the effects of partial information, it is necessary to define levels of information as follows:

i_1, the correct answer is known.
i_2, guessing between two choices including the answer.
⋮
i_p, guessing among p choices including the answer.
⋮
i_n, guessing among all n choices.

If it is assumed that examinees do guess under CG whenever at least one incorrect choice can be eliminated, expected item scores are as shown in Table 1.

CG expected item scores are a linear function of NR expected item scores, since

$$\frac{n-p}{p(n-1)} = \frac{n}{n-1} \cdot \frac{1}{p} - \frac{1}{n-1} \quad . \qquad [2]$$

As a result, CG expected total scores are a linear function of NR expected total scores, so that the two scoring methods are essentially equivalent as long as the scaling difference is taken into account.

The Arnold Scoring/Response Method

Arnold and Arnold (1970) reported an interesting and potentially valuable scoring/response method that eliminates score fluctuation due to chance when partial information is employed to eliminate some, but not all, incorrect choices. Under Arnold and Arnold's (AA) method, examinees are instructed to mark all choices they firmly believe to be incorrect, with the admonition that they should not guess in the absence of knowledge, since, on the average, score gains due to guessing will be removed by the scoring process. Responses are scored by assigning the CG expected item score values according to the knowledge level inferred from the number of incorrect choices eliminated, as-

Table 1
Expected NR and CG Item Scores for Levels of Information i_1 to i_n

Information Level	Expected Item Scores NR	CG
i_1	1	1
i_2	1/2	$1/2 - (1/2)[1/(n-1)] = (n-2)/2(n-1)$
i_3	1/3	$1/3 - (2/3)[1/(n-1)] = (n-3)/3(n-1)$
⋮	⋮	⋮
i_p	1/p	$1/p - (1-1/p)[1/(n-1)] = (n-p)/p(n-1)$
⋮	⋮	⋮
i_n	1/n	$1/n - (1-1/n)[1/(n-1)] = 0$

suming no guessing. However, should the correct choice be eliminated (possibly due to misinformation) a score deduction of $-1/(n-1)$ is made, the same penalty as under CG. Thus, scoring and penalties are equivalent to NR and CG, with the advantage that assigned item scores are essentially the same as expected item scores in the absence of completely random guessing; this should reduce total score variance, coincidentally enhancing reliability.

Even if guessing does occur, under AA there is, on the average, no advantage or penalty to the examinee. To see that this outcome is the case, consider an examinee who has eliminated $n-p$ incorrect choices (i_p) and decides to eliminate one more choice by guessing. The item score earned before guessing was $(n-p)/p(n-1)$. Guessing, the probability is $1/p$ that the correct choice will be marked and $1-1/p$ that it will not. The earned item score when the correct choice is marked is $-1/(n-1)$ and $(n-p+1)/(p-1)(n-1)$ if another incorrect choice is successfully marked. Then the expected item score is

$$\frac{1}{p}\cdot\frac{-1}{n-1}+(1-1/p)\frac{n-p+1}{(p-1)(n-1)}=\frac{n-p}{p(n-1)}\quad, \qquad [3]$$

exactly what had been earned previously.

The Answer-Until-Correct Method

For an answer-until-correct (AUC) test, examinees are directed to continue selecting choices until the correct choice has been identified. Usually, correctness of a choice is revealed when the examinee erases an answer shield corresponding to the desired choice. This procedure gives immediate feedback to the examinee and is highly regarded from the standpoint of its potential for enhancing learning (Pressey, 1950).

A scoring rule for this response method was first proposed by Brown (1965). In practice (Gilman & Ferry, 1972; Hanna, 1975), Brown's rule results in scoring $n-1, n-2, \ldots, 0$ points for $1, 2, \ldots, n$ erasures per item. It is easy to show that this scoring rule gives credit proportional to the number of incorrect choices eliminated and, when properly transformed, yields an expected gain of 0 from guessing.

In general, if an examinee guesses repeatedly among p choices (i_p), the average number of erasures will be the sum of the integers 1 to p divided by p or

$$\frac{1+2+\cdots+p}{p}=\frac{(p+1)(p/2)}{p}=\frac{p+1}{2}\quad. \qquad [4]$$

If informed guessing on an n-choice item is among p choices, $n-p$ choices have been correctly identified as incorrect, and credit may be awarded as

$$(n-p)/(n-1)\quad, \qquad [5]$$

the ratio of identified incorrect choices to the total number of incorrect choices. Then, to obtain the best estimate of credit due for an observed number of erasures, simply set Equation 4 equal to the number of erasures, solve for p, and substitute this value in Equation 5. For an n-choice item the following item score values result:

1 erasure, $(n-1)/(n-1)=1$.

2 erasures, $(n-3)/(n-1)$.

3 erasures, $(n - 5)/(n - 1)$.

$\vdots$

k erasures, $(n - 2k + 1)/(n - 1)$.

$\vdots$

n erasures, $(n - 2n + 1)/(n - 1) = -1$.

That this scoring is a linear transformation of Brown's rule may be seen from the fact that

$$\frac{n - 1}{2} \cdot \frac{n - 2p + 1}{n - 1} + \frac{n - 1}{2} = n - p \ . \qquad [6]$$

The expected item scores under the derived rule for different information levels are, from Equation 5, simply the proportion of incorrect choices eliminated. Under total ignorance (i_n) the average number of erasures will be $(n + 1)/2$ (from Equation 4). In this case $p = n$, and the expected item score from Equation 5 is 0.

Resulting scores are *not* linear transformations of NR or CG item scores. Under AUC, information levels i_1 and i_n correspond to expected item scores of 1 and 0 as under CG; but i_2 has an expected item score of $(n - 2)/(n - 1)$ under AUC and $(n - 2)/2(n - 1)$ under CG. Moreover, there is a substantial difference with respect to misinformation.

To understand this situation, it is necessary to define $n - 1$ levels of misinformation for an n-choice item:

m_1, the examinee has categorized only the correct choice as incorrect and guesses among the $n - 1$ remaining wrong choices.

m_2, the examinee has categorized the correct choice and one other as incorrect and guesses among the $n - 2$ remaining choices.

$\vdots$

m_p, the examinee has categorized the correct choice and $p - 1$ wrong choices as incorrect and guesses among the $n - p$ remaining choices.

m_{n-1}, the examinee has categorized the correct choice and p - 1 wrong choices as incorrect and hence believes that the remaining incorrect choice is correct.

The average number of erasures under AUC and corresponding expected item score for each of these levels in shown in Table 2. Of course, the formulas for the expected number of erasures were determined under the assumption that the misinformed examinee received no advantage from finding out that the answer was not in the subset initially under consideration and thereafter had to guess at random to find the answer.

It may be observed from the above expected item scores under misinformation that the penalties range from $-1/(n - 1)$ to -1. The lowest penalty (for level m_{n-1}) is essentially the same as for CG, since the effective score range under AUC is 0 to N, as under CG. However, the maximum penalty of -1 (for level m_1) is $(n - 1)$ times as large.

The Coombs and Related Response/Scoring Methods

Coombs (1953) recommended a response/scoring method under which examinees mark all responses they firmly believe to be wrong, as under AA. However, the scoring is different. Under Coombs' (CBS) method the examinee receives one point for each correctly identified incorrect choice and $-(n - 1)$ points for inadvertently marking the correct choice. Unlike AUC scoring, feedback to

Table 2
Expected Number of Erasures and Item Scores Under AUC for Misinformation Levels m_1 to m_{n-1}

Misinformation Level	Expected No. of Erasures	Expected Item Score
m_1	n	-1
m_2	$(n - 2) + (1 + 2)/2 = n - 1/2$	$-(n - 2)/(n - 1)$
m_3	$(n - 3) + (1 + 2 + 3)/3 = n - 1$	$-(n - 3)/(n - 1)$
⋮	⋮	⋮
m_p	$(n - p) + (1 + 2 + \cdots + p)/p = (2n - p + 1)/2$	$-(n - p)/(n - 1)$
⋮	⋮	⋮
m_{n-1}	$1 + (1 + 2 + \cdots + n - 1)/(n - 1) = n/2 + 1$	$-1/(n - 1)$

the examinee is not possible, since the examinee could reduce the loss whenever a correct answer was revealed by marking any previously unmarked incorrect answers.

Table 3 gives item scores for CBS for different answer conditions resulting from different levels of information and misinformation. If each of the item scores is divided by $(n - 1)$, the scoring is exactly equivalent to the expected values for corresponding levels under AUC. Thus, CBS is equivalent to AUC in the same manner in which AA is equivalent to CG, i.e., the assigned scores directly reflect the knowledge (or misinformation) level inferred from the examinee's marks. As with AUC, CBS is not equivalent to CG and has varying penalties for misinformation, with an expected item score of 0 for total ignorance regardless of the number of marks.

However, CBS, unlike AA, does penalize the examinee with partial information who decides to guess after marking all choices known to be incorrect. Suppose $n - p$ incorrect choices have been marked before guessing, in which case the earned item score is $n - p$. If a guess is now made, the expected item score is

$$(1/p)[-(n - 1) + (n - p)] + (1 - 1/p)(n - p + 1) = n - p + (1 - n/p) \quad . \qquad [7]$$

Since $n/p > 1$, $n - p$ is always reduced by guessing.

A possible variant on CBS is to instruct examinees to mark as many choices as they believe necessary to include the correct answer. Item scores are then assigned on the basis of the number of incorrect responses left unmarked and are exactly equivalent to those under CBS.

Another variant on CBS is one that would permit feedback using erasable answer sheets designed for AUC. The examinee would erase the shields over choices firmly believed incorrect with a fixed item score of $-(n - 1)$ whenever the correct answer was uncovered. Expected and assigned item scores for the first $n - 1$ levels of information would be the same as for CBS. Misinformation would always yield an item score of $-(n - 1)$ equal to the most severe penalty under CBS. Unlike CBS and AUC, however, this method would yield an expected penalty if a totally ignorant examinee guessed on *more than one* choice per item.

To verify this interesting outcome, assume that p choices have been eliminated successfully (p could be 0). Then the probability of marking the correct answer is $1/(n - p)$ with a corresponding

Table 3
Item Scores Under CBS for Information Levels i_1 to i_n and Misinformation Levels m_1 to m_{n-1}

Examinee Response	Level	Item Score
All wrong choices marked, right unmarked	i_1	$n - 1$
All but one wrong choice marked, right unmarked	i_2	$n - 2$
⋮	⋮	⋮
$n - p$ wrong choices marked, right unmarked	i_p	$n - p$
⋮	⋮	⋮
No mark or n marks	i_n	0
Only the right choice marked	m_1	$-(n - 1)$
Right and one wrong choice marked	m_2	$-(n - 2)$
⋮	⋮	⋮
Right and $p - 1$ wrong choices marked	m_p	$-(n - p)$
⋮	⋮	⋮
Right and $n - 2$ wrong choices marked	m_{n-1}	-1

penalty of $-(n - 1)$. The probability of marking another incorrect is $1 - 1/(n - p)$, in which case the item score earned is $p + 1$. Then, the expected item score for marking another choice is

$$-(n - 1)\frac{1}{n - p} + (p + 1)\left(1 - \frac{1}{n - p}\right) = \frac{pn - p^2 - 2p}{n - p} \quad . \qquad [8]$$

Subtract from this result the previous gain of p to obtain

$$\frac{pn - p^2 - 2p}{n - p} - p = \frac{-2p}{n - p} \quad . \qquad [9]$$

Therefore, the expected result of guessing is a score loss, unless $p = 0$, in which case the result is neutral. Moreover, the loss becomes more severe as the number of prior successfully eliminated incorrect choices increases.

A New Response/Scoring Method

To enhance validity with respect to some specific criterion, it might be desirable to reverse the order of the misinformation penalty, which under CBS increases from m_{n-1} to m_1. To accomplish this reversal, while permitting feedback to examinees, the following response/scoring method has been proposed by Cross (see Cross & Thayer, 1979). Students are instructed to erase on a feedback answer sheet the shields over the answers they believe incorrect, with scoring as follows for an n-choice item:

Marking Condition	Item Score
$n - 1$ incorrect choices erased	$2n - 1$
$n - 2$ incorrect choices erased	$2n - 2$
⋮	⋮

1 incorrect choice erased	$n+1$
no erasures	n
correct choice only erased	$n-1$
correct and 1 incorrect choice erased	$n-2$
⋮	⋮
correct and $n-2$ incorrect choice erased	1

In this case the examinee who inadvertently erases the correct choice would lose points by erasing further. Moreover, the penalty increases from m_1 to m_{n-1}, the opposite of AUC. To compare penalty levels with other methods, it is necessary to perform a linear transformation on the item scores from tthe proposed new method (PNM) so that i_1 receives a score of 1 and total ignorance (with no guessing) a score of 0 as follows:

Marking Condition	PNM Transformed Item Score
$n-1$ incorrect choices erased	1
$n-2$ incorrect choices erased	$(n-2)/(n-1)$
⋮	⋮
1 incorrect choice erased	$1/(n-1)$
no erasures	0
correct choice only erased	$-1/(n-1)$
correct and 1 incorrect choice erased	$-2/(n-1)$
⋮	⋮
correct and $n-2$ incorrect choices erased	-1

Expected values for the various levels of knowledge are the same as the assigned score values and as those for AUC. However, expected values for various levels of misinformation must be derived as shown in Table 4. Since under PNM the penalty for m_1 is equivalent to that for m_{n-1} under AUC and $(-n/2)/(n \times 1) > -1$ (the penalty for m_1 under AUC) it would appear that PNM is conveniently similar to, if slightly less severe than, AUC, while reversing the direction of the misinformation penalty.

This outcome is essentially the case; but further clarification is needed, since guessing under total ignorance does not, in general, yield an expected item score of 0 under PNM as under AUC. To understand this situation, assume the examinee marks at random one choice on an n-choice item. Then the expected (transformed) item score is

Table 4
Expected Item Scores Under PNM for Misinformation Levels m_1 to m_{n-1}

Level	Expected Item Score (transformed)
m_1	$-1/(n-1)$
m_2	$[-1/(n-1)]/2 + [-2/(n-1)]/2 = (-3/2)/(n-1)$
⋮	⋮
m_p	$[-1/(n-1)]/p + [-2/(n-1)]/p + \ldots + [-p/(n-1)]/p = [-(p+1)/2]/(n-1)$
⋮	⋮
m_{n-1}	$[-1/(n-1)]/(n-1) + [-2/(n-1)]/(n-1) + \ldots + [-1]/(n-1) = (-n/2)/(n-1)$

$$(1/n)[-1/(n-1)] + (1-1/n)[1/(n-1)] = (n-2)/n(n-1) \ . \qquad [10]$$

This value is always positive for $n > 2$. Hence, it is always advantageous (or at least harmless, on the average) to guess at least once per item under PNM. At this point the question arises whether to guess further on those items where the correct answer was not revealed. The same question arises when partial information has resulted in the successful erasure of one or more incorrect choices. In general, if p erasures of incorrect choices have been made, the earned item score is $p/(n-1)$. If a guess is now made and the correct answer erased, the new earned score is $-(p+1)/(n-1)$. If, instead, another incorrect choice is erased, the new earned score is $(p+1)/(n-1)$. Then, the expected item score is

$$\frac{1}{n-p} \cdot \frac{-(p+1)}{n-1} + \left(1 - \frac{1}{n-p}\right)\frac{p+1}{n-1} = \frac{pn + n - 3p - p^2 - 2}{(n-p)(n-1)} \ . \qquad [11]$$

Now subtract the previously earned score from this expected score.

$$\frac{pn + n - 3p - p^2 - 2}{(n-p)(n-1)} - \frac{p}{n-1} = \frac{n - 3p - 2}{(n-p)(n-1)} \qquad [12]$$

For the result to be positive it is necessary that

$$n - 3p - 2 > 0 \ , \qquad [13]$$

or

$$3p + 2 < n \ . \qquad [14]$$

For $n \leq 5$ this inequality holds for integral values of p only if $p = 0$. For a 10-choice item, the inequality holds for $p \leq 2$. Thus, for a test consisting of items with five or fewer choices, examinees should be informed that they should erase at least one choice on each item but not erase further in the absence of information. For a 10-choice item, the examinee should guess on no more than the third erasure after up to two successful erasures of incorrect choices.

It is also of interest to determine the maximum expected score on an n-choice item attainable by guessing. If the examinee resolves to erase up to p choices on an n-choice item under total ignorance, the expected PNM item score may be derived as

$$\frac{1}{n}\left(\frac{-1}{n-1} + \frac{-2}{n-1} + \cdots + \frac{-p}{n-1}\right) + \left(1 - \frac{p}{n}\right)\frac{p}{n-1}$$

$$= \frac{p}{n-1} - \frac{3p^2}{2n(n-1)} - \frac{p}{2n(n-1)} \ . \qquad [15]$$

Although always nonnegative for $p = 1$, it is of interest to investigate the behavior of the function just derived in order to determine the potential for gaining over a range of n's by guessing under PNM. To this end, the derivative is taken with respect to p and is set equal to 0:

$$1/(n-1) - 3p/n(n-1) - 1/2n(n-1) = 0 \ . \qquad [16]$$

The second derivative, $-3/n(n-1)$, is a negative, so that solution for p yields a maximum, namely,

$$p = n/3 - 1/6 \ . \qquad [17]$$

Substituting this value for p in Equation 15 yields

$$(2n - 1)^2/24n(n - 1) \quad . \qquad [18]$$

This expression then represents the maximum expected item score from guessing under PNM for a given n. Of course, only integral values of p could be substituted into Equation 15, so the theoretical maximum of Equation 18 may be slightly above that attainable in practice. For example, if $n = 5$, Equation 17 yields $p = 1.5$. Substituting $p = 1$ and $p = 2$ in Equation 15 yields expected transformed item scores of .15 (equivalent to an expected raw item score of 5.6) for guessing once *or* twice.

With respect to Equation 18, it is important to note the extent to which the expected maximum guessing gain varies according to the number of choices per item. Substituting integral values of n from 2 through 15 into Equation 18 yields values within the narrow range of .17 to .19. Thus, PNM has the desirable characteristic that total ignorance tends to yield about the same expected item score regardless of the number of choices, and this value is lower than that for a five-choice item under NR.

Because a limited amount of guessing is favorable under PNM, the effective score range is reduced as for NR. Accordingly, to compare misinformation penalties, it is necessary to express them as a fraction of the effective score range on an N-item test. For $2 \leqslant n \leqslant 15$, the largest expected item score from guessing is 1/6. Therefore, a reasonable range for comparison is $(5/6)N$ for PNM, while the range is N for AUC. Then, misinformation penalties can be compared as shown in Table 5. Hence, the mimimum penalty under PNM is effectively larger than the minimum under AUC. At the same time, the maximum penalty under AUC is effectively larger than the maximum under PNM.

Discussion

The various scoring rules just reviewed may be categorized according to their treatment of misinformation. NR, CG, and AA all have effectively the same relatively small, uniform penalty for misinformation. Moreover, for all three methods, expected item scores are the same and guessing to any extent has no effect on expected item scores (provided resulting scores are transformed to the same scale). In use, the three methods differ in the amount of total score variation due to chance as a result of guessing. NR should be the poorest in this regard. Under CG, examinees are urged to report their actual state of knowledge when totally ignorant, hence reducing chance variation from corresponding items. Under AA, examinees are expected to report their true state of knowledge on every item, thus further reducing chance variation. Uncritical evaluation of these results would suggest that AA is inherently superior. However, the literature on the use of CG strongly suggests that examinees often cannot or will not determine or display their true level of information (see Cross & Frary, 1977). Of course, the examinee can hardly avoid displaying misinformation, and the three methods are equivalent in this respect. Choice of one of these would depend on appropriateness of the small misinformation penalty (from the standpoint of validity) and on the ability of the examinees to report their true information levels.

Table 5
Comparison of Maximum and Minimum Misinformation Penalties Under AUC and PNM

Maximum Penalty		Minimum Penalty	
AUC	PNM	AUC	PNM
$-1/N$ for m_1	$> -(n/2)/N(5/6)(n - 1) = -1/N(5n/3)(n - 1)$ for m_{n-1}	$-1/N(n - 1)$ for m_{n-1}	$< -1/N(5/6)(n - 1)$ for m_1

AUC and CBS are similar in terms of expected item scores under all levels of information and misinformation. Expected item scores under information differ from NR, CG, and AA in that they are relatively greater for the higher levels of partial information. For example, on five-choice items, an examinee who can eliminate three incorrect choices gains 3/4 point under AUC/CBS but only 3/8 point, on the average, under NR/CG/AA. The difference for misinformation is much more striking. When a single incorrect choice is firmly believed correct (m_{n-1}), the penalty for misinformation under AUC/CBS is the same as under NR/CG/AA. However, as fewer incorrect choices are categorized with the correct choice, the penalty increases until it is $(n - 1)$ times as great on an n-choice item (m_1). This characteristic may or may not be beneficial from the standpoint of validity with respect to a variety of criteria. Hence, effects on validity should be investigated before undertaking extensive use of AUC or CBS. In selecting between the two, it should be kept in mind that CBS expects the examinee to report the level of information on each item, but AUC virtually insures that this is done. The author's experience suggests that many examinees have difficulty using CBS properly. Hence, examinee characteristics also need to be taken into consideration.

If the direction of increasing misinformation penalty under AUC/CBS (m_{n-1} to m_1) is counterproductive, PNM may be desirable. Its misinformation penalty varies, with the least penalty for m_1 and greatest for m_{n-1}, though the range is not quite as great as for AUC/CBS. PNM has the added advantage of encouraging some response on every item, since a score gain from guessing is inherent for the first one or two erasures per item. To judge this characteristic as desirable is consistent with much reported research that examinees who believe themselves totally ignorant can, on the average, identify some incorrect choices (Cross & Frary, 1977).

Table 6 shows expected item scores for five-choice items under the various response/scoring methods and information/misinformation levels.

Conclusions

Within the response/scoring methods reviewed, three factors emerge:

1. The extent to which examinees are expected or required to report their true state of knowledge on an item. CBS, PNM, and AA would be most demanding in this respect, with CG markedly less

Table 6
Expected Raw and Standardized[1] Item Scores Under Various Information/Misinformation Levels for 5-Choice Items

Level	Raw NR	Raw CG/AA[2]	Standardized NR/CG/AA	Raw AUC	Raw CBS	Standardized AUC/CBS	Raw PNM[2]	Standardized PNM
i_1	1.00	1.00	1.00	4	4	1.00	9	1.00
i_2	.50	.38	.38	3.5	3	.75	8	.71
i_3	.33	.17	.17	3	2	.50	7	.41
i_4	.25	.06	.06	2.5	1	.25	6	.12
i_5	.20	.00	.00	2	0	.00	5.6	.00
m_1	.00	-.25	-.25	0	-4	-1.00	4	-.47
m_2	.00	-.25	-.25	0.5	-3	-.75	3.5	-.62
m_3	.00	-.25	-.25	1	-2	-.50	3	-.76
m_4	.00	-.25	-.25	1.5	-1	-.25	2.5	-.91

[1]Linear transformation to assign 1 to i_1 and 0 to i_5.
[2]Assuming average gain from guessing whenever appropriate.

demanding, and AUC and NR even less demanding. Selection of a method must be consistent with examinee characteristics such as motivation and ability to understand directions, taking into consideration the likely enhancement of reliability under AA, CBS, and PNM.

2. The value of partial information. NR, CG, and AA give less credit for partial information, while AUC and CBS give the most. This question is primarily related to predictive validity and the nature of the criterion. However, the content of the predictor test may also be of concern if logical analysis argues in favor of more or less credit for partial information. Content analysis might resolve these questions to some extent, but empirical research would probably be needed to determine the best response/scoring method for any particular criterion.
3. The function of misinformation with respect to validity. Since the various response/scoring methods penalize misinformation to varying degrees, empirical studies are likely to reveal validity differences according to the method used.

These conclusions demonstrate the importance of response/scoring method determination in test development. This aspect of the test development process often appears to be completely ignored. Evidence of this oversight is the fact that, other than NR and CG, the methods reviewed have seldom gained more than experimental use. AUC is particularly promising, but there seems to have been no effort to adapt it to optical scanning equipment. CBS or AA could presently be adapted to optically scanned response sheets, though, admittedly, these methods are quite demanding of examinees.

Unfortunately, the search for improved validity through use of response/scoring methods other than NR has not been fruitful in the past. Thus, it has been the purpose of this paper to offer principles on which such a search may be based.

References

Arnold, J. C., & Arnold, P. L. On scoring multiple-choice exams allowing for partial knowledge. *The Journal of Experimental Education,* 1970, *39*, 8-13.

Brown, J. Multiple response evaluation of discrimination. *The British Journal of Mathematical and Statistical Psychology,* 1965, *18*, 125-137.

Coombs, C. H. On the use of objective examinations. *Educational and Psychological Measurement*, 1953, *13*, 308-310.

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

Cross, L. H., & Thayer, N. F. *A new method for administering and scoring multiple-choice tests: Theoretical and empirical consideration.* Unpublished manuscript, Virginia Polytechnic Institute and State University, 1979.

Davis, F. B. *Educational measurements and their interpretation.* Belmont, CA: Wadsworth, 1964.

Gilman, D. A., & Ferry, P. Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement,* 1972, *9*, 205-207.

Hanna, G. S. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. *Journal of Educational Measurement*, 1975, *12*, 175-178.

Pressey, S. L. Development and appraisal of devices providing immmediate automatic scoring of objective tests and concomitant self-instruction. *The Journal of Psychology,* 1950, *29*, 419-447.

Author's Address

Send requests for reprints or further information to Robert B. Frary, Learning Resources Center, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.