

The Effect of Guessing on Item Reliability under Answer-Until-Correct Scoring

Michael Kane
National League for Nursing, Inc.

James Moloney
State University of New York at Brockport

The answer-until-correct (AUC) procedure requires that examinees respond to a multiple-choice item until they answer it correctly. The examinee's score on the item is then based on the number of responses required for the item. It was expected that the additional responses obtained under the AUC procedure would improve reliability by providing additional information on those examinees who fail to choose the correct alternative on their first attempt. However, when compared to the zero-one (ZO) scoring procedure, the AUC procedure has failed to yield consistent improvements in reliability. Using a modified version of Horst's model for examinee behavior, this paper compares the effect of guessing on item reliability for the AUC procedure and the ZO procedure. The analysis shows that the relative efficiency of the two procedures depends strongly on the nature of the item alternatives and implies that the appropriate criteria for item selection are different for each procedure. Conflicting results reported for empirical comparisons of the reliabilities of the two procedures may result from a failure to control for the characteristics of the items.

During the past twenty years, several testing procedures have been proposed for assessing partial information on multiple-choice test items. Among the procedures that have been suggested are elimination scoring (Coombs, Millholland, & Womer, 1956), response weighting

(Wang & Stanley, 1970), and confidence testing (DeFinetti, 1965; Shuford, Albert, & Massengill, 1966). The aim of all of these procedures is to obtain more information from each test item by allowing examinees who do not recognize the correct answer to indicate their level of partial knowledge.

The answer-until-correct (AUC) procedure (Gilman & Ferry, 1972) employs a very simple and intuitive mechanism for making use of partial knowledge. When the AUC procedure is used on a test, examinees are required to respond to each item until they choose the correct alternative for the item. The score on the item is a function of the number of responses required to choose the correct alternative. The standard zero-one (ZO) scoring procedure allows only one response to each item; the additional responses required by the AUC procedure are intended to provide information on the level of partial knowledge for those examinees who cannot identify the correct alternative on their first try. The assumption can be made that the additional responses collected under the AUC procedure might have an effect similar to that of increasing the number of items and, therefore, increasing reliability.

In a test taken under AUC instructions, after responding incorrectly to an N -alternative item, the examinee responds to an item with the same stem but with the $N-1$ remaining alternatives.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 2, No. 1 Winter 1978 pp. 41-49
© Copyright 1978 West Publishing Co.

The content of the item remains essentially the same; however, the residual item is clearly not independent of the original item. Since the Spearman-Brown formula does not apply, a more detailed analysis of examinee response behavior is necessary if the effect of the AUC procedure on reliability is to be predicted.

The few empirical studies that have compared the reliability of AUC scoring to that of ZO scoring have led to contradictory results. Gilman and Ferry (1972) found that the split-half reliability for AUC scoring was substantially higher than it was for ZO scoring. However, Hanna (1975) and Taylor, West, and Tinning (1975) report little or no improvement in reliability when the AUC procedure was used. The empirical results, therefore, do not provide consistent support for the hypothesis that the additional responses obtained under the AUC procedure generally increase test reliability.

Previous studies (Brown, 1965; Merwin, 1959) have indicated that the properties of the items determine whether or not the AUC procedure will have a higher reliability than the ZO scoring procedure. The items used in the empirical studies of the AUC procedure were not selected specifically for use with the AUC procedure; therefore, it is likely that some of the items used in these studies were more suitable for AUC scoring and others were more suitable for ZO scoring.

The present study uses a simple model of examinee behavior to examine the effect of item characteristics and scoring rules (AUC or ZO) on item reliability. The results indicate that the two scoring rules require different item characteristics in order to minimize the influence of guessing on item reliability.

Assumptions of the Model

The model for examinee response behavior makes two assumptions for a multiple-choice item with one correct alternative and $N-1$ distractors:

1. The examinee recognizes n of the distrac-

tors to be incorrect and eliminates them from consideration, leaving a residual set of $k = (N-n)$ alternatives.

2. The examinee selects alternatives from the residual set randomly and without replacement, until the correct alternative is chosen.

The first assumption allows for the existence of degrees of partial information, indicated by the number of distractors an examinee can eliminate. However, this assumption precludes the possibility of misinformation; it does not allow for the possibility that an examinee might eliminate the correct alternative.

Although the model does not allow misinformation, it is possible within the model for examinees to possess different partial information. Examinees with a given level of partial information (defined by the number of distractors eliminated) may eliminate different subsets of distractors. Two examinees, each eliminating n distractors, do not necessarily recognize the same distractors as being incorrect.

The model used in this study is based on a model proposed by Horst (1932) to examine the effect of guessing when multiple-choice items are scored on the basis of a single examinee response. The second assumption is included to make the model applicable to AUC scoring. The first assumption is similar to Horst's main assumption, but differs in one important respect. Horst's model assumes that all of the distractors are uniquely ordered from the most plausible to the least plausible; it also assumes that the alternatives are scalable. All examinees who eliminate n distractors are assumed to eliminate the same set of n alternatives. Lord and Novick (1968, p. 312) have pointed out that this is the least plausible of Horst's assumptions; the model used in this study does not make this assumption.

The simple model used here makes it possible to predict the proportion of item variance resulting from guessing for both the AUC procedure and the ZO procedure. These results can then be used to select items for which the effect of guess-

ing is minimized when either scoring rule is used.

Implications of the Model

An expression for the proportion of item variance attributable to guessing can be derived from the model. This expression will be general enough to apply to any scoring rule that could be used with either the AUC procedure or the ZO procedure. In subsequent sections, scoring rules typically used with each of these procedures will be specified.

Examinees who can eliminate all but k of the alternatives are said to be in the state C_k ($k=1, \dots, N$). Let α_k be the probability that a randomly selected examinee is in state C_k . The vector of probabilities,

$$\alpha = (\alpha_1, \dots, \alpha_N), \quad [1]$$

describes the structure of the item relative to some population and will be called the item characteristic vector. Let s_j be the score given to any examinee who requires j responses to select the correct alternative. The vector,

$$S = (s_1, \dots, s_N), \quad [2]$$

specifies the numerical score to be assigned to any response sequence and is called the scoring rule or score vector.

The probability that an examinee in state C_k will respond correctly on the first attempt is merely the probability of randomly selecting the correct alternative from a set of k alternatives. If X is a random variable designating the examinee's observed score, the probability that an examinee in state C_k would answer correctly on the first attempt is given by

$$P(X=s_1 | C_k) = \frac{1}{k}. \quad [3]$$

The probability that an examinee in state C_k ($k \geq 2$) will require two responses on an item is equal to the probability of selecting a distractor

on the first attempt, multiplied by the probability of selecting the correct alternative from the set of $k-1$ alternatives remaining after this attempt,

$$P(X=s_2 | C_k) = \left(\frac{k-1}{k}\right)\left(\frac{1}{k-1}\right) = \frac{1}{k}. \quad [4]$$

Since examinees in C recognize the correct answer, the probability that they would require more than one response is zero. In general, the probability that an examinee in C_k will require j responses is given by

$$P(X=s_j | C_k) = \frac{1}{k}, \quad [5]$$

if $j \leq k$, and is zero for $j > k$. Thus, the observed scores for the examinees in any state have a uniform distribution over the set of possible scores for that state. The probability π_j that an examinee, chosen randomly from the population, will require j responses to select the correct alternative is

$$\pi_j = \sum_{k=1}^N P(X=s_j | C_k) [P(C_k)] \quad [6]$$

$$= \sum_{k=j}^N (1/k)\alpha_k.$$

In order to estimate α , the observed proportion of examinees requiring j responses to answer an item may be substituted for π_j in Equation 6.

The expected value of the observed score over the population of examinees is

$$E(X) = \sum_{j=1}^N s_j \pi_j = \sum_{j=1}^N s_j \sum_{k=j}^N \left(\frac{1}{k}\right)\alpha_k. \quad [7]$$

Changing the order of summation,

$$E(X) = \sum_{k=1}^N \alpha_k A_k, \quad [8]$$

where A_k is the expected score for an examinee in state C_k :

$$A_k = \frac{1}{k} \sum_{j=1}^k s_j. \quad [9]$$

Similarly, the expected value of X^2 is

$$E(X^2) = \sum_{k=1}^N \alpha_k B_k, \quad [10]$$

where B_k is the expected value of the squared score for examinees in state C_k

$$B_k = \left(\frac{1}{k}\right) \sum_{j=1}^k s_j^2. \quad [11]$$

From Equations 8 and 10, the observed score variance for an item is given by

$$\sigma^2(X) = \sum_{k=1}^N \alpha_k B_k - \left(\sum_{k=1}^N \alpha_k A_k\right)^2. \quad [12]$$

Given C_k , the true score (T_k) for an examinee in state C_k is defined to be the expected value of X ,

$$T_k = \sum_{j=1}^k s_j (1/K) = A_k. \quad [13]$$

The true score variance, over the population of examinees, is

$$\sigma^2(T) = \sum_{k=1}^N \alpha_k A_k^2 - \left(\sum_{k=1}^N \alpha_k A_k\right)^2. \quad [14]$$

The error (e) for any examinee is the difference between his/her observed score on the item and his/her true score. The variance in e over independent trials for an examinee in state C_k is the observed score variance for examinees in C_k :

$$\sigma_k^2(e) = B_k - A_k^2. \quad [15]$$

Therefore, the error variance for an examinee is a function of the state of partial knowledge of the examinee (C_k). Those examinees who know least about the item do the most guessing, while examinees who recognize the correct answer never guess. The error variance for C_1 is zero; the error variance over the population is merely a weighted average of the error variances for the particular states:

$$\sigma^2(e) = \sum_{k=1}^N \alpha_k \sigma_k^2(e) \quad [16]$$

$$= \sum_{k=1}^N \alpha_k B_k - \sum_{k=1}^N \alpha_k A_k^2.$$

From Equations 12, 14, and 16 it follows that

$$\sigma^2(X) = \sigma^2(T) + \sigma^2(e). \quad [17]$$

The observed score variance is the sum of the true score variance and the error variance. Since the true scores for all examinees are equal to their expected observed scores, the expected value of e for all examinees is zero. Therefore, the correlation between true score and error is given by

$$E(eT) = E[(X-T)T] \quad [18]$$

$$= E(XT) - E(T^2).$$

The first term in Equation 18 can be expanded as

$$E(XT) = \sum_{k=1}^N \sum_{j=1}^k s_j T_k (1/k) \alpha_k \quad [19]$$

$$= \sum_{k=1}^N T_k A_k \alpha_k$$

$$= E(T^2).$$

Substituting Equation 19 in Equation 18, the correlation of true score with error is found to be zero. With an appropriate definition of true score, the model becomes a special case of the classical test theory model.

For $\alpha_1 < 1.0$, reliability can be defined as the ratio of the true score variance to the observed score variance:

$$\rho = \frac{\sigma^2(T)}{\sigma^2(X)} \quad [20]$$

$$= \frac{\sum_{k=1}^N \alpha_k A_k^2 - \left(\sum_{k=1}^N \alpha_k A_k \right)^2}{\sum_{k=1}^N \alpha_k B_k - \left(\sum_{k=1}^N \alpha_k A_k \right)^2} .$$

The assumption that $\alpha_1 < 1.0$ is necessary to assure that the observed score variance is different from zero. Because guessing is the only source of error considered, Equation 20 is an upper bound for the usual indices of stability, equivalence, and internal consistency.

The item reliability given by Equation 20 depends on both the vector of probabilities (α), describing the structure of the item relative to some population, and the score vector (S), specifying the item score as a function of the number of responses needed to select the correct alternative.

Score Vectors

Where the AUC procedure has been used (Gilman & Ferry, 1972; Hanna, 1975; Taylor et al., 1975), the scoring rule has had the following simple form:

$$s_k = N - k , \quad [21]$$

where k is the number of responses needed to choose the correct alternative.

Dalrymple-Alford (1970) has examined an alternate scoring rule for the AUC procedure that uses an average uncertainty measure based on information theory. However, this measure of uncertainty depends strongly on the number of alternatives in the item. For this reason Dalrymple-Alford suggests that the scoring rule in Equation 21 is preferable to the average uncertainty.

Because Equation 9 is invariant under linear

transformations of the scoring rule, an equivalent but more convenient form of this rule will be used:

$$s_k = k . \quad [22]$$

The rule specified by Equation 22 will be employed to define the AUC scoring rule, although it is possible to examine the reliability of items for any monotone scoring rule for the AUC procedure by appropriate substitutions in Equation 20.

For the ZO procedure, the examinee responds only once to each item, and an appropriate scoring rule is given by

$$s_k = \begin{cases} 1 & k = 1 \\ 0 & 2 \leq k \leq N \end{cases} . \quad [23]$$

It is not known how many responses are required for an examinee to choose the correct alternative when the ZO procedure is used. However, this is not a serious problem because the same score is assigned to all examinees requiring more than one response to choose the correct alternative.

Two Special Cases

Two special cases for the form of the item characteristic vector (α) illustrate the implications of Equation 20.

The first case considers the N -alternative items for which each examinee can eliminate at least $N-2$ of the alternatives. This implies that there is only one effective distractor for each examinee. However, it does not imply that the same distractor is effective for all examinees. The item characteristic vector has the form

$$\alpha = (\alpha_1, \alpha_2, 0 \dots 0) . \quad [24]$$

Substituting these values in Equation 20 and simplifying the resulting expressions, the same result is obtained for both the AUC scoring rule and the ZO scoring rule:

$$\rho = \alpha_1 / (\alpha_1 + 1) . \quad [25]$$

It is not surprising that both scoring rules lead to the same expression for ρ , because the item in this case is effectively a two-alternative item for each examinee. The AUC scoring rule is then a linear transformation of the ZO scoring rule.

The second case assumes that all of the examinees fall into one of two mutually exclusive states. They can either eliminate all of the distractors and choose the correct alternative on their first response, or they cannot eliminate any of the distractors and guess randomly until they choose the correct alternative. The item characteristic vector corresponding to these assumptions is

$$\alpha = (\alpha_1, 0, \dots, 0, \alpha_N) . \quad [26]$$

Item reliabilities for AUC-scored items that satisfy these assumptions can be derived from Equations 20, 22, and 26. The result after algebraic simplification is given by

$$\rho = \frac{(N-1)\alpha_1}{(N-1)\alpha_1 + (1/3)(N+1)} . \quad [27]$$

Reliabilities for ZO-scored items can be derived similarly from Equations 20, 23, and 26:

$$\rho = \frac{(N-1)\alpha_1}{(N-1)\alpha_1 + 1} . \quad [28]$$

For this case, then, the reliability for ZO scoring is greater than it is for AUC scoring, provided $N > 2$. As Horst's original paper (1932) showed, the item characteristic vector given in Equation 26 minimizes the effect of guessing for multiple-choice items scored in the traditional way. The probability that an examinee who does not know the correct answer will choose the correct alternative on the first response is a minimum, $1/N$, when the examinee cannot eliminate any of the distractors.

The reason for the superiority of ZO scoring when all of the distractors are equally plausible

is obvious. Under this condition, only the first response supplies any information about the examinees. All of the examinees having any information on the item answer it correctly on their first response, along with those guessing correctly. After the first response all responses on an item are guesses and supply no reliable information on the examinee. In this case, the additional responses required by the AUC procedure are not necessary.

Items With Three Alternatives

The relationship between item reliability and item characteristics will next be examined for items with three alternatives. The results will also apply to items with more than three alternatives for which each examinee is able to eliminate at least $N-3$ of the alternatives.

Table 1 presents the values of ρ for AUC scoring and ZO scoring for items with three alternatives (increments of .1 are used for α_1 and α_2 and $\alpha_3 = (1.0 - \alpha_1 - \alpha_2)$). The cells in Table 1 for given values of α_1 and α_2 consist of two entries. The first entry is the AUC reliability as computed from Equations 20 and 22, and the second entry is the ZO reliability as computed from Equations 20 and 23. The line in Table 1 separates it into two sections. In the upper section the reliability for ZO scoring is larger than it is for AUC scoring. In the bottom section, the reliability for ZO scoring is less than or equal to the reliability for AUC scoring.

The top row in Table 1 is an example of the second of the special cases discussed in the last section. If α_2 is zero, then all of the examinees fall into one of two categories; they either know the correct answer or they have no information about the item and guess randomly. ZO reliability is uniformly higher than the AUC reliability under these conditions.

For the last entry in each column, α_3 is zero; and the item is effectively a two-alternative item. As shown previously, the reliabilities for AUC scoring and ZO scoring are identical in this case.

For a fixed value of α_1 the ZO reliability is a

Table 1
AUC and ZO Reliabilities for Items with 3 Alternatives

α_2	α_1								
	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.00									
AUC	.130	.231	.310	.375	.429	.474	.512	.545	.574
ZO	.167	.286	.375	.444	.500	.546	.583	.615	.643
0.10									
AUC	.155	.248	.323	.383	.433	.474	.506	.528	.474
ZO	.166	.277	.360	.424	.474	.511	.536	.542	.474
0.20									
AUC	.176	.262	.331	.387	.432	.466	.485	.444	
ZO	.163	.267	.344	.402	.444	.472	.480	.444	
0.30									
AUC	.192	.272	.335	.385	.423	.443	.412		
ZO	.158	.255	.326	.377	.412	.427	.412		
0.40									
AUC	.202	.276	.333	.375	.398	.375			
ZO	.152	.241	.306	.350	.375	.375			
0.50									
AUC	.207	.274	.322	.348	.333				
ZO	.144	.226	.283	.320	.333				
0.60									
AUC	.204	.261	.294	.286					
ZO	.133	.208	.258	.286					
0.70									
AUC	.190	.231	.231						
ZO	.121	.189	.231						
0.80									
AUC	.158	.167							
ZO	.107	.167							
0.90									
AUC	.091								
ZO	.091								

maximum when α_2 equals zero. It decreases approximately linearly for increasing values of α_2 . On the other hand, the AUC reliability generally increases at first, reaches a maximum, and then decreases as a function of α_2 . For any number of alternatives and any item characteristic vector, the two scoring rules can be compared by evaluating Equation 20. When items with more than three alternatives are examined in this way, the relationship between AUC reliability and ZO

reliability becomes quite complicated. There are, however, two general characteristics of the AUC reliability that are very clear in Table 1; they apply to items with any number of alternatives.

First, the possible improvements to be derived from the substitution of AUC scoring for ZO scoring are most pronounced when α_1 is relatively small (i.e., for difficult items). When α_1 is large, a substantial proportion of the examinees

know the correct answer; and any procedure for assessing partial information will affect relatively few examinees. It is, therefore, not surprising that when α_1 is large, the substitution of AUC scoring for ZO scoring has little effect on item reliability.

Second, for a given value of α_1 , the AUC reliability tends to be highest when both α_2 and α_3 are appreciably different from zero. For items with N alternatives, the AUC reliability tends to be highest when $\alpha_2, \alpha_3, \dots, \alpha_N$ are all appreciably different from zero. Examinees can be discriminated on the basis of partial information only when they differ in their levels of partial information on an item. To the extent that examinees do not differ in their levels of partial knowledge, the additional responses called for by the AUC procedure only increase the variance due to guessing.

Conclusions

In general, the total number of responses for a set of items will be greater for the AUC procedure than for the ZO procedure. Gilman and Ferry (1972) have suggested that these additional responses tend to increase the reliability of a test by effectively lengthening it. The results reported in the present study can be interpreted as an analysis of the effect that these additional responses have on item reliability, where ZO scoring provides a standard of comparison.

The model predicts that the effect of guessing on item reliability depends on both the item characteristics and the scoring rule. The empirical research which has compared AUC scoring to ZO scoring has ignored the effect of item characteristics. This failure to control item characteristics may account for the fact that AUC scoring has not consistently produced improvements in reliability when it has been compared to ZO scoring.

The admission of additional responses for the AUC procedure increases the observed score variance. If the AUC procedure produces a larger proportional increase in true score vari-

ance than in error variance, its use will increase reliability; otherwise, it will tend to decrease item reliability. The present study has produced some model-based predictions about the item characteristics determining which of these conditions will hold.

Although the model considers guessing to be the only source of error in comparing the two scoring systems, this may not be a serious limitation. Other sources of error are likely to have approximately the same effect for both scoring rules, since the content of the items remains essentially the same independently of the scoring rule.

The accuracy of the predictions made in this paper will depend on how closely the assumptions of the model reflect actual test-taking behavior. Horst's model is clearly more realistic than the traditional model for guessing; the version used in the present study has relaxed the least plausible of Horst's assumptions—that the alternatives can be uniquely ordered. An empirical test reported by Dalrymple-Alford (1970) indicates that the model which has been used provides a good fit to actual test-taking behavior.

There are, however, a number of ways in which the assumptions of the model may be violated; further empirical testing is, therefore, needed. For example, it is unlikely that the $N-m$ alternatives in the residual set have equal probability of being chosen after n alternatives have been eliminated. The robustness of the model against violations of this assumption should be examined. Cronbach and Merwin (1955) have suggested a model which assumes that the alternatives have different probabilities of being chosen. Although their model is mathematically unwieldy, it could be employed in monte carlo studies of the model used here.

Another assumption which needs to be examined empirically is that the AUC procedure does not change what is being measured. The AUC procedure may introduce error, since examinees are not allowed to correct mistakes if they carelessly and unintentionally mark an alternative. If the test is not speeded and if examinees are in-

structed to be careful about marking alternatives, this source of error can be minimized.

The AUC procedure has the advantage of not introducing errors associated with the differential effects of risk-taking behavior. The possible influence of risk-taking is a serious threat to the validity of scores when some other methods of assessing partial information (e.g., elimination scoring and confidence testing) are used. Equation 20 makes it possible to select items that will minimize the effect of guessing when the AUC procedure is used. Alternately, items can be analyzed to judge whether any improvement in reliability can be expected if AUC scoring is substituted for ZO scoring.

References

- Brown, J. Multiple response evaluation of discrimination. *British Journal of Mathematical and Statistical Psychology*. 1965, 18, 125-137.
- Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial knowledge. *Educational and Psychological Measurement*. 1956, 16, 13-37.
- Cronbach, L. J., & Merwin, J. C. A model for studying the validity of multiple-choice tests. *Educational and Psychological Measurement*. 1955, 15, 337-352.
- Dalrymple-Alford, E. C. A model for assessing multiple-choice test performance. *British Journal of Mathematical and Statistical Psychology*. 1970, 23, 199-203.
- DeFinetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*. 1965, 18, 87-123.
- Gilman, D., & Ferry, P. Increasing test reliability through self-scoring procedures. *Journal of Educational Measurement*. 1972, 9, 205-207.
- Hanna, G. Incremental reliability and validity of multiple-choice tests with an answer-until-correct procedure. *Journal of Educational Measurement*, 1975, 12, 175-178.
- Horst, P. The chance element in the multiple-choice test item. *Journal of General Psychology*, 1932, 6, 209-211.
- Lord, F., & Novick, M. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley, 1968.
- Merwin, J. C. Rational and mathematical relationship of six scoring procedures applicable to three-choice items. *Journal of Educational Psychology*, 1959, 50, 153-161.
- Shuford, E. H., Albert, A., & Massengill, M. E. Admissible probability measurement procedure. *Psychometrika*, 1966, 31, 125-145.
- Taylor, J., West, D., & Tinning, F. *An examination of decision-making based on a partial credit scoring system*. Paper presented at the meeting of the National Council on Measurement in Education, Washington, DC, 1975.
- Wang, M. W., & Stanley, J. C. Differential weighting: a review of methods and empirical studies. *Review of Educational Research*, 1970, 40, 663-705.

Author's Address

Michael T. Kane, National League for Nursing, Inc.,
 10 Columbus Circle, New York, NY 10019