Dirasat, Educational Sciences, Volume **32**, No. **1**, 2005

Using Distractors in Correcting for Guessing in Multiple-Choice Tests

***Hamzeh M. Dodeen* ***

ABSTRACT

This study proposes a new formula for correcting for guessing in multiple-choice items, the Distraction-Scoring Formula (DSF). DSF takes into account the option distracting level in calculating an examinee's final test score. It assumes that guessing is not always random and that the most distracting option for examinees in multiple-choice tests is the most nearly correct one. The effectiveness of DSF was compared with the common correcting for guessing formula, Scoring Formula (SF) using data from two standardized tests that were used at the United Arab Emirates University (UAEU). Results supported the use of DSF to award students for their selection of options based on the level of correctness when not sure about the correct answer. Also, the use of DSF increased test reliability for subsamples drawn from the total test sample.

INTRODUCTION

Guessing in multiple-choice tests has been observed as a problem that affects test scores. Test reliability and validity are reduced when examinees respond to test items by randomly selecting answers. Rogers (1999) mentioned three types of guessing: random guessing which occurs when an examinee responds blindly to a test item, cued guessing where an examinee responds to a stimulus in a test item, and informed guessing where an examinee responds based on partial knowledge or misinformation. Although the three types involve some guessing, they are different from a psychometrical point of view. While clued and informed guessing are used to measure partial knowledge and are usually encouraged by teachers and examiners, random guessing is undesirable.

Generally, test scores are used to measure the examinee's ability on some trait(s). If blind guessing is used in answering the tests questions, particularly in multiple-choice tests, some of the scores will be obtained only by chance. Consequently, an important source of error is introduced, which reduces the validity and accuracy of scores. To overcome this problem, test scores need to be corrected. Correcting for guessing is one of the testing issues that has received attention from researchers and educators for the last few decades. However, to date, no single solution is acknowledged as best (e.g. Rogers, 1999; Wang, 1995). Chevalier (1998:3) reviewed several scoring algorithms for correction for guessing and concluded that while some procedures indicate a slight increase in reliability and validity of tests, these increases "do not justify the additional complexity, time, and cost involved in developing, administering, scoring, and interpreting test results".

Scoring Formula

Scoring Formula (SF) is one of the most common procedures used to correct for guessing in multiple choice-tests. The formula reduces an examinee's total score by a proportion of the number of wrong answers. This deduction is assumed to compensate for the scores obtained by random guessing. The calculation of that magnitude depends on the number of options for each test item. A commonly used formula for SF is:

$$F = R - \frac{W}{A-1}$$

Where:

F is the examinee corrected score

R is the number of items answered correctly

W is the number of items answered incorrectly, and

A is the number of options in items.

The utility of SF and its impact on test results and on reliability and validity of tests have been studied and evaluated by many researchers. (See for example: Abu-

* Faculty of Educational Sciences, The United Arab Emirates University. Received on 9/3/2003 and Accepted for Publication on 16/12/2003.

Sayf, 1979; Chevalier, 1998; Frary, 1980; Jaradat and Tollefson, 1988; Kurz, 1999; and Tollefson and Chung, 1986).

Advocates of SF claim that it discourages students from guessing (Oosterhof, 1994); it reduces scores variance caused by guessing (Frary, Cross, and Sewell, 1985), and it is more appropriate for speeded tests and tests with low passing score. In speeded tests time is limited, and with low passing score there is lack of basis for answering many items, therefore in these two situations many examinees will guess at random; hence, using SF will reduce the massive amount of random guessing (Frary, 1988).

In contrast, many shortcomings of the SF have been observed. First, SF assumes that an examinee tends to guess at random whenever he/she does not know the correct answer (Rogers, 1999). This is an incorrect assumption about the examinee's behavior. It ignores the clued guessing and guessing based on partial knowledge or misinformation. Usually, examinees evaluate options before deciding on selecting one or eliminating one or more. Ignoring the examinee's evaluation of the wrong answer is an issue that is worth consideration in multiple-choice testing (Tollefson and Chung, 1986).

Second, SF penalize examinees with some specific personal traits. When SF is used as a grading procedure for a test, examinees are instructed to omit rather than guess when they do not know the answer. The omit response is considered neither right nor wrong (Frary, 1988). The tendency to follow these instructions is a personal trait that differs among examinees. Studies have shown that students who follow instructions (low-risk takers) are penalized by SF, while students who ignore the instructions (high-risk takers) increase their chances of getting more items correct by so doing and guessing (Albanese, 1988; Angoff, 1989; and Frary et al., 1985).

Third, a negative effect of SF is increasing the test time. This is observed when there are omissions after the examinees' last response which suggests that the test time was not adequate (Frary, 1988). In terms of improving test validity and reliability by using SF, there is no consistent conclusion in the literature about how SF affects validity and reliability of tests. (See for example: Oosterhof, 1994; Rogers, 1999; Tollefson and Chung, 1986; Zin and Williams, 1991).

Partial Credit Models

The main shortcoming of the SF is that it ignores partial knowledge in answering an item. Partial knowledge is assumed to be present when an examinee is able to eliminate some of the incorrect alternatives. Partial credit procedures overcome this limitation of SF by awarding some credit for the selection of a wrong option. The amount of the credit gained is supposed to represent the level of correctness of the wrong option. Crocker and Algina (1986) classified partial knowledge models into three groups: Confidence Weighting, Answer-Until Correct, and Option Weighting.

When a confidence weighting procedure is used, examinees are instructed to provide their level of confidence in selecting an answer. Credits are given; hence, based on whether the selection is correct or not and on the level of confidence stated (Kurz, 1999). In answer-until-correct procedure, feedback is immediately given to the examinee about his/her selection (answer). The examinee, then, is informed to move on to the next item if the selection he/she made was correct or to select another option if the selection was incorrect. Some drawbacks of this method are the administration costs and the involvement of some personal factors such as test anxiety (Chevalier, 1998).

Another scoring procedure that assumes partial knowledge is option weighting. In this procedure, a special weight is given to the selection of an option based on its level of correctness. Expert judgment of the weights of option correction is one of these methods. Another method gives greater weight to options selected by a high total score among examinees. Studies on these methods have shown a slight increase in the reliability and validity of tests (Crocker and Algina, 1986).

Purpose

The purpose of this study is to propose a new formula of correction for guessing and to demonstrate its utility in sample set data. The new formula is the Distraction-Scoring Formula (DSF). In multiple-choice tests, DSF takes into account the Alternative Distraction Level (ADL) in calculating the examinee's final test score. ADL is determined by the percentage of students selecting each alternative. The formula is:

$$F = R - \frac{\sum_{r=1}^{K} M_r}{A-1}$$

Table 1.: Distraction of Options and Examinee's Response on the Ten Wrong Items.

Item	1	2	3	4	5	6	7	8	9	10
Order of options	a	c	a	d	c	d	a	c	c	d
	b	b	b	a	a	b	b	b	a	b
	c	d	d	b	d	a	c	d	d	a
Examinee's response	a	b	d	d	c	a	a	b	c	a
M_r	1/3	2/3	1	1/3	1/3	1	1/3	2/3	1/3	1

Note. $\sum_1^{10} M_r$ = 1/3 + 2/3 + 1 + 1/3 + 1/3 + 1 + 1/3 + 2/3 + 1/3 + 1 = 6.

Where:

F is the final corrected score

R is the number of items answered correctly

K is the number of items answered incorrectly

A is the number of options, and

M_r = 1/ A-1, if the examinee selects the most frequently chosen distractor, 2/A-1, if the examinee selects the second most frequently chosen distractor, …, and 1, if the examinee selects the least frequently chosen distractor.

In the case of getting equal percentages for the wrong options of an item, the average of the values 1/A-1, 2/A-1, …, 1 is given for the examinee's response to that item.

The DSF weights options by their distraction. The level of distraction is determined by the percentage of students who select each option. The most distracting option is the one that is selected by the highest percentage of examinees. Although a high percentage of examinees selecting an option does not necessarily mean a high correctness level of that option (Tollefson and Chung, 1986), the relationship between the level of correctness of an option and the percentage of examinees who select it needs to be determined. If this relationship is positive and consistent, then using the DSF can be of practical use for estimating the partial knowledge that is assumed in an examinee's response to a multiple-choice item. However, if no consistent relationship is observed, the formula might still have impact on test characteristics.

Theoretical Base of DSF

Under the framework of the Classical Test Theory (CTT), test characteristics are determined with respect to a particular group of examinees (Hambleton et al., 1991). Test characteristics include item indices (difficulty and discrimination) and distractors analysis. This means that how a group of students responds to a test determines the test psychometric features in CTT.

In multiple-choice tests, distraction of options can be a measure of examinees' achievement level. Not only does this enhance the utility of using distractors, but it also emphasizes the need for having attractive alternatives. In addition, most classroom and standardized tests are interpreted as norm-referenced rather than criterion-referenced tests. In norm-referenced tests, examinees' scores are compared with each other. In other words, scores are not absolute measures of examinees' knowledge or ability (Frary, 1988). Scores are often used to give relative positions for the test group. Therefore, the distraction level of an option, which is determined by the percentage of students who select the option, can be an appropriate procedure to distinguish among the examinees' levels and to locate each in a relative position.

Example of Distraction-Scoring Formula

As an example of using DSF, assume an examinee responded to a 30-item multiple-choice test with four options for each item. Assume further that the examinee answered 20 items correctly, 10 items incorrectly, with no omitted items. To calculate the final test score for this examinee, both the distracting level of each option and then M_r are calculated for each of the wrong items. Since there are four options for each item, M_r = 1/3 when the examinee selected the most distracting option, 2/3 for the moderate distracting one, and 1 for the least distracting option. Table (1) shows the rank order of each option on each wrong item and the examinee's hypothetical response.

By applying the DSF formula the final examinee's score is 20 - 6/(4-1) = 18. When the SF is applied, the same examinee's final score is 20 - 10/(4-1) = 16.67. The difference between the two scores is 1.33 points (4.4%),

the award that the examinee got for selecting more distraction options for some items. Notice here that M_r in DSF is less than or equal to W in SF. This means that an examinee score on DSF will never be less than his/her score in SF. Both formulas give the same score when an examinee selects the least likely distraction option for each item answered incorrectly.

A Comparison of Scoring Formula and Distraction-Scoring Formula

Although both formulas (DSF and SF) use the number of wrong items in the test and the number of options in the item in the calculation of the final examinee's score, DSF overcomes many of the shortcomings observed when SF is used. First, SF assumes that all guessing is random (blind). In contrast, DSF assumes that guessing is not always random. The examinee selects the best answer from the item options when he/she is not sure about the correct answer. This selection is based on some clues or partial knowledge rather than random guessing. When the examinee has more probability than chance only to get the answer correct, DSF awards this by giving some credit for a specific selection. It assumes that the most distracting options for examinees in multiple-choice tests are the more nearly correct ones. If this assumption is correct then DSF could be used to credit for partial knowledge that is assumed when responding to multiple-choice items.

Second, SF ignores the use of distractions in the scoring process. Hence, any answer in the SF framework is either correct or incorrect. It makes no difference whether the examinee selected the most or the least distracting option. In contrast, DSF uses distraction of options in scoring such that more credits are awarded for selecting more distracting options. This is especially important when there is the "correct answer" and the "most correct one". In this way, the distraction of options can be used as a measure of achievement.

Third, DSF links the final grading process with the initial process of developing or writing test items. This relationship increases the importance of developing or selecting alternatives to be attractive to students. Attractive means related or closed to the correct answer. This issue is important in DSF scoring because the selection of the wrong alternative will be accounted for students.

Empirical Examples

To illustrate the use of DSF in real testing examples, data sets of two standardized tests that have been used in the United Arab Emirates University (UAEU) were analyzed. The first test was the English Test for Level Three students in the English Program at the University General Requirements Unit. The sample size of this test was 227 male freshmen students. The second was the Quantitative Test which is used to measure students' quantitative ability. The sample size of this test was 148 senior students (71 females, 77 males). Both tests were administrated during the 2nd semester of 2000-2001. Each test consisted of 40 multiple-choice questions with four options. Students were instructed to guess when they were not sure about the correct answer. There was no penalty for guessing.

Two analyses were conducted on the results of each test. First, the relationship between the level of correctness of each distractor and its attractiveness to students was estimated. The level of correctness was determined by having five teachers who conducted the test or who taught in the program, evaluate each distractor of each question for each test, and then rank order distractors based on their level of correctness.

The level of attractiveness of each distractor for each question was determined by calculating the percentage of students who selected each option, so that the higher the percentage, the more attractive the distractor. The level of agreement between the two rank orderings of the distractors was estimated by calculating the Spearman correlation and the Cohen's Kappa measure of agreement (Kappa values ranged from –1 to 1. A value of –1 indicates perfect disagreement while a value of 1 indicates perfect agreement).

The second analysis used DSF formula in scoring both tests. The effect of the formula on the test scores was analyzed and compared with the results if SF had been used in scoring. Finally, the possible effect of using the DSF on the test reliability was examined.

Results

Rank Ordering of Distractors: Five teachers in the English Program at UAEU reviewed each item of the English Test and rank ordered each distractor based on the level of correctness of each. The same was done for the items of the Quantitative Test. Only items for which

at least three teachers agreed on the level of correctness of their distractors were included in the analysis. Based on that, 8 items from the Quantitative Test and 7 items from the English Test were excluded from this analysis because of disagreement among the teachers on the level of correctness. This suggests that it is difficult to write items with clearly discernable option rankings. The other classification of the distractors was based on the level of attractiveness of each. This was determined by the percentage of students who selected each option. The Spearman correlations between the two classifications were significant ($p < .001$), and were .69 and .65 for the Quantitative Test and the English Test, respectively. The results of Cohen's Kappa were .62 and .58 ($p < .01$) for the Quantitative Test and the English Test, respectively. These significant agreements between the two classifications support the use of distractors in the estimation of the level of correctness for both exams.

Results of Using the Two Scoring Formulas: Table (2) shows the descriptive statistics for the results of the two tests using SF and DSF in scoring.

Table 2.: Descriptive Statistics of the Scores on each Test by Using SF and DSF.

Test	Sample size	Formula [a]	Mean	SD
English Test	227	SF	20.7	7.0
		DSF	23.2	6.3
Quantitative Test	148	SF	26.2	10.0
		DSF	28.8	8.9

a: SF: Scoring Formula and DSF: Distraction-Scoring Formula.

The gains in the mean of the students' scores were 2.5 points (5%) for the English Test and 2.6 points (5.2%) for the Quantitative Test (maximum score is 40 in both tests). These values represent the gains that the students were awarded for selecting options with a higher level of attraction. As for the standard deviation of the scores, there was a reduction of about 1 point (2.5%) in both tests when the DSF was used.

The difference in each student's score when using SF and DSF in scoring was calculated. The distribution of the differences was checked for normality using the Kolomogorove-Smirnov (K-S) Test. The results showed that the distributions of the differences in both tests were not close to the normal distribution. The two distributions were negatively skewed. This means that students with high scores gained less than other students. In another analysis, correlations were calculated between the differences (gain scores) and the total scores (measured by the number of correct answers) on each test. The Pearson correlations were significant ($p < .05$) and were -.57 and -.84 for the English Test and the Quantitative Test, respectively. The negative correlation means that as scores go high, the gain decreases. In the English Test, the students with totals 28, 15, 10, 19, and 20 gained the highest five differences: 6.47, 6.13, 6.06, 5.54, and 5.43, respectively. Similarly, in the Quantitative Test, the five students who got the highest five differences: 6.20, 5.87, 5.83, 5.67, and 5.13 had the total scores: 20, 22, 15, 25, and 25, respectively. All these students are of medium ability as measured by the total score (between the 25^{th} and the 75^{th} percentile).

Reliability Analysis: Alpha coefficients were calculated for the two tests using the complete samples. The results showed no difference when DSF was used for scoring as compared with SF. The effect of the DSF on the high-ability students is, on average, not large because these students usually miss only a few items, and the effect of DSF on low-ability students is small too because these students could not distinguish the level of correctness of options. Considering that, the total score was used to create three subsamples with decreased total scores. These subsamples represented individuals who scored less than 30, less than 25, and less than 20 (Table 3). As the total score drops, the sample size drops as well. In addition, students' homogeneity in the created subsamples increases because of narrowing the range (variation) of the scores. Consequently, test reliability is reduced (Linn and Gronlund, 2000). This explains the reduction in the reliability value for both tests when smaller and more homogenous samples were used. However, when comparing the reliability values obtained by using each scoring formula, it was observed that the reliability values with DSF were slightly higher than their counterparts with SF. This result was consistent over the three subsamples and for both tests. For example, in the second subsample of the Quantitative Test the reliability of the test when using SF was .59 while it was .64 when using DSF.

Table 3.: Reliability of Scores for the Total Sample and for Three Subsamples Drawn from Each Test.

Test	Formula [a]	All students	Total scores <30	Total scores < 25	Total scores < 20
		(N=227)	(N =176)	(N =92)	(N =33)
English	SF	.74	.53	.22	.15
Test	DSF	.74	.56	.26	.18
		(N=148)	(N =65)	(N=51)	(N=30)
Quantitative Test	SF	.89	.72	.59	.39
	DSF	.89	.75	.64	.40

a: SF: Scoring Formula and DSF: Distraction Scoring Formula.

DISCUSSION

This study investigates the effect of DSF in scoring multiple-choice tests and the estimation of scores reliability. The generalizeability of the results of the current study is limited to the examples used and the conditions assumed. More specifically, the correlation between distractor correctness and distractor popularity may vary from test to test, and students were instructed to guess. Of course, tests with different instructions regarding guessing yield different results. Further examples using tests with different subject matters, samples, and under different testing conditions need to be investigated.

In both multiple-choice tests used in this study, a positive relationship was found between the distractor level of correctness and the level of attractiveness. This relationship indicates that when answering multiple-choice questions, students tend to select the best answer based on its level of correctness rather than a random or blind selection. As long as students evaluate the options by using clues or partial knowledge, there is a probability more than chance only to get the answer correct. In this case, not only should the examinees' evaluation of the wrong answer not to be ignored (Tollefson and Chung, 1986), but students should also be awarded for non-random selection. This is an advantage offered by using DSF as compared with the common scoring formula (SF).

DSF awards all students for their selection of options when they are not sure about the correct answer. However, students of moderate ability gain the maximum scores. Another advantage of using DSF in scoring which has been observed from the results of this study is the reduction in score variances. This was considered one of the important advantages of SF (Frary et al., 1985). However, results showed that DSF reduced standard deviation of scores more than SF.

When the total test sample was used, DSF had the same effect as SF on estimating test reliability. However, DSF influenced the estimation of the test reliability on subsamples drawn from the total sample using the total score. This occurred because students with different ability levels benefit differentially from the use of DSF in scoring. This causes the change in the impact of DSF on test reliability estimation. As compared with SF and across the drawn subsamples from both tests, DSF gives slightly higher reliability values.

Two important issues are related to the application of the DSF formula in real testing situations. First, students must be trained on the use of the formula, so they know how it is used in scoring. This may help them learn how to deal wisely with options in multiple-choice questions. Second, the advantage of using DSF will be increased if tests are constructed with attention to the use of the formula in scoring. In other words, options must be selected carefully to represent different levels of correctness. Bearing these in mind, DSF can provide a promising formula to correct for guessing on multiple-choice tests.

REFERENCES

Abu-Sayf, F. K. 1979. The Scoring of Multiple-choice Tests: A Closer Look. *Educational Technology*, 19: 5-15.

Albanese, M. A. 1988. The Projected Impact of the Correction for Guessing on Individual Scores. *Journal of Educational Measurement*, 25: 149-157.

Angoff, W.H. 1989. Does Guessing Really Help? *Journal of Educational Measurement*, 26: 323-336.

Chevalier, S.A. 1998. *A Review of Scoring Algorithms for Ability and Aptitude Tests*. Paper Presented at the Annual Meeting of Southwest Psychological Association, New Orleans.

Crocker, L. and Algina, J. 1986. *Introduction to Classical and Modern Test Theories*. Orlando: Harcourt Brace Jovanovich, Inc.

Frary, R. 1980. The Effect of Misinformation, Partial Information, and Guessing on Expected Multiple-choice Test Item Scores. *Applied Psychological Measurement*, 4: 79-90.

Frary, R. 1988. Formula Scoring of Multiple-choice Tests (Correction For Guessing). *Educational Measurement: Issues and Practice*, 7: 33-38.

Frary, R., Cross, L. and Sewell, E. 1985. *Partial Information and the "Correction" for Guessing*. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. 1991. *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Jaradat, D. and Tollefson, N. 1988. The Impact of Alternative Scoring Procedures for Multiple-choice Items on Test Reliability, Validity, and Guessing. *Educational and Psychological Measurement*, 48: 627-635.

Kurz, T. B. 1999. *A Review of Scoring Algorithms for Multiple-choice Tests.* Paper Presented at the Annual Meeting of Southwest Educational Research Association, San Antonio.

Linn, R. and Gronlund, N. E. 2000. *Measurement and Assessment in Teaching*. New Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Oosterhof, A. 1994. *Classroom Applications of Educational Measurement*. New York, NY: Macmillan College Publishing Company.

Rogers, H. J. 1999. Guessing in Multiple-choice Tests. In: Masters, G. N. and Keeves, J. P. (Eds). *Advances in Measurement in Educational Research and Assessment*. 235-243, Oxford, UK: Pergamon.

Tollefson, N. and Chung, J. 1986. *A Comparison of Two Methods of Assessing Partial Knowledge on Multiple-choice Tests*. (ERIC Document Reproduction Service No. ED 299 281).

Wang, J. 1995. *Critical Values of Guessing on True-false and Multiple-choice Tests.* Paper Presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Zin, T. and Williams, J. 1991. *Searching for Better Scoring of Multiple-choice Tests: Proper Treatment of Misinformation, Guessing, and Partial Knowledge*. (ERIC Document Reproduction Service No. ED 339 744).

طرح معادلة جديدة للتصحيح لأثر التخمين في اختبارات أسئلة الاختيار من متعدد

حمزة محمد دودين*

ملخص

تهدف هذه الدراسة إلى طرح معادلة جديدة للتصحيح لأثر التخمين في أسئلة الاختيار من متعدد Distraction-Scoring Formula (DSF). ومما يميز المعادلة الجديدة (DSF) أنها تأخذ بعين الاعتبار مستوى جاذبية البديل في حساب درجة السؤال النهائية، وتفترض أن التخمين ليس عشوائياً دائماً، وأن الاختيار الأكثر جاذبية هو الاختيار الأقرب إلى الصواب من ضمن قائمة البدائل في سؤال الاختيار من متعدد. وقد تمت في هذه الدراسة مقارنة المعادلة الجديدة (DSF) بالمعادلة التقليدية للتصحيح لأثر التخمين الشائعة الاستعمال Scoring Formula (SF)، كذلك تمت مقارنة فعالية المعادلتين خصوصاً من حيث تأثير كل منهما على درجات الاختبار و ثباتها من خلال تطبيق المعادلتين على نتائج اختبارين معياريين في جامعة الإمارات العربية المتحدة.

وقد دلت النتائج على أفضلية استخدام المعادلة الجديدة (DSF) في مكافأة الطلبة على حسن اختيارهم للبديل الأقرب إلى الصواب أو الأكثر صحة عندما يكونون غير متأكدين تماماً من الجواب الصحيح، وبالإضافة إلى ذلك فقد أدى تطبيق هذه المعادلة إلى زيادة ثبات نتائج الاختبارين عندما حسبت على عينات فرعية مشتقة من العينة الكلية لكل اختبار.

* كلية العلوم التربوية، جامعة الامارات العربية المتحدة، العين، دولة الامارات العربية المتحدة. تاريخ استلام البحث 2003/3/9، وتاريخ قبوله 2003/12/16.