Introducing a Non-Dichotomous Method for Scoring Selected-Response Tests

A.A. Rostami Abusaeedi [1] (PhD)
H. Zahedi[2] (PhD)

Abstract

The main purpose of this study was to introduce a new non-dichotomous scoring procedure for the multiple-choice test method, called Parledge. The research also aimed to determine Parledge characteristics through a psychometric study comparing it with the "Number Correct" and "Correction for Guessing" scoring methods. There were 35 students participating in the study who were given a vocabulary and a structure tests. The results revealed favorable characteristics for parledge although it involves penalty threat. Further, positive psychometric, educational and cognitive implications were provided.

Keywords: parledge, MC ,guessing, scoring methods, partial knowledge, multiple-choice, selected response, Non-dichotomous scoring.

Introduction

The objective of mental testing, including language testing, is to estimate test takers' underlying knowledge or abilities of a specific construct or trait (Bachman, 1990). Among the procedures used to measure cognitive ability is the multiple-choice method (MC). Although the MC enjoys objectivity, simplicity, and automatic scoring, it is susceptible to guessing and insensitive to differences among various levels of knowledge (Ben-Simon, Budescu, and Nevo, 1997). As contrasted with other testing procedures, the guessing effect appears to be qualitatively different in the MC to the effect that we can never know what part of any individual's score has come about via guessing (Hughes, 1989). The examinee may have guessed all or some of the answers which has a narrowing effect on the range of scores (Weir, 1990) resulting in an overestimate of the person's ability (Smith, 1990). This inflation of scores gives an unfair advantage to examinees who guess frequently as opposed to those who do not

1. Faculty Member of the Shahid Bahonar University of Kerman
2. Faculty Member of the Shahid Bahonar University of Kerman

(Choppin, 1990). Moreover, guessing reduces the reliability of the scores when it is random and jeopardizes the validity of the inferences drawn from them when it varies among test takers in a systematic way (Prieto and Delgado, 1999).

Dichotomous Scoring Methods

There are two widely-used scoring methods for the MC which treat performance dichotomously, namely, number-correct (NC) and correction for guessing (CG).

The NC procedure simply awards the examinee one point for each correct response with the justification that guesses might be informed. Although simple and straightforward, this method may encourage guessing of incorrect and potentially unsafe responses (Tweed and Wilkinson, 2009) and fails to award partial credit for partial knowledge (Chevalier, 1998). In this paper, *partial knowledge* has been used with two senses: (1) a broad meaning, that is, any level of knowledge between *full knowledge* and *absence of knowledge* and (2) a narrow meaning, within Abu-Sayf's (1979) and Frary's (1980) model of knowledge which will be presented below).

Psychometrically, encouraging guessing will have negative consequences on the reliability of the scores and validity of the measurement because the effect of chance on the scores is unequal from subject to subject and uncorrelated with the construct (Prieto and Delgado, 1999). Furthermore, from an educational point of view, it is an undesirable habit to guess the answer to unknown questions (Thorndike, 1971).

The CG procedure sometimes referred to as rights-minus-wrongs, takes into account three possible situations: (a) the examinee knows the correct option, (b) the examinee omits the item, or (c) the examinee guesses blindly (Kurz, 1999). The proponents of CG procedure favor penalty for wrong answers which they believe are based on blind selection. Put differently, CG was a weighting formula for points awarded for correct answers, incorrect answers, and unanswered questions so that the expected value of the increase in test score due to guessing was zero (Prihoda, Pinckard, McMahan, and Jones, 2006).

Although it is believed that CG increases test scores reliability due to reducing measurement error produced by guessing (Henning, 1987), its assumptions are believed to be too simplistic to be credible (Choppin, 1990). Mousavi (1999, 2002) and Kline (2000) list other problems associated with CG as follows: (1) CG assumes that all incorrect answers are the result of guessing; thus, the test takers who did not guess would be unreasonably penalized for their wrong responses (Harris, 1969). In this case, CG overcorrects for chance success, treating such individuals harshly. (2) CG assumes that there is an equal chance for each alternative to be selected. However, this is not true, since in some items the individual may have eliminated some of the distractors, narrowing down the number of possible correct answers before guessing. For these guessers, the correction formula is an underestimate, under-correcting for chance success. (3) CG assumes that test takers have an average luck in guessing. However, there is no means through which we can determine the adequacy of this presupposition. As a result, the guessing correction seems to introduce an unknown amount of error into the scoring.

CG fails to take into account partial knowledge and that the scoring formula leaves more room for computational miscalculation (Kurz, 1999). Besides, several studies have shown that test takers who are low risk takers are penalized by CG (see Angoff, 1989; Albanese, 1988; Bliss, 1980; Cross and Frary, 1977; Slakter, 1968).

Non-dichotomous Scoring Methods Based on Partial Knowledge Model

Abu-Sayf (1979) and Frary (1980) (cited in Ben-Simon, 2000) developed a model of partial knowledge with five distinct levels of knowledge: (1) Full knowledge: the examinee has full knowledge regarding the problem presented in a given item, and is able to choose the correct answer with full confidence. (2) Partial knowledge (correct): the examinee possesses some degree of (correct) knowledge with regard to the test item, but this knowledge is insufficient for choosing the correct answer with full confidence. (3) Partial misinformation: the examinee possesses some degree of incorrect knowledge regarding the test item, but this knowledge is insufficient

for choosing an incorrect answer with full confidence. (4) Full misinformation: the examinee possesses some incorrect knowledge regarding the test item and thus chooses an incorrect alternative with full confidence. (5) Absence of knowledge: the examinee has no knowledge whatsoever regarding the problem at hand. Ben-Simon, et al., (1997) present five methods accounting for partial knowledge as follows:

Elimination Testing (ET)

In ET, introduced by Coombs, Milholland, and Womer (1956) examinees are instructed to eliminate all distractors that they can identify as incorrect. Comparisons of the reliability and validity of ET with NC and CG methods slightly favor ET (e.g., Collet, 1971; Coombs et al., 1956; Hakstian and Kansup, 1975; Jaradat and Tollefson, 1988). Coombs et al. (1956) and Jaradat and Tollefson (1988) also reported that a majority of the examinees thought ET was a better and fairer method. Most versions of ET discourage random guessing, discriminate between all the levels of knowledge, and in most cases tend to improve the psychometric quality of the tests. However, the methods are relatively complex (in terms of instructions and scoring) and require longer administration time. There are some indications that examinees may apply ET ineffectively (i.e., too conservatively).

Probability Testing (PT)

PT is the most general and flexible method in this class. PT allows examinees to express partial knowledge in the most detailed and specific manner by reporting the probability that each option is the correct answer. There are 101 possible scores for each item (ranging from 0 to 1). Michael (1968) and Pugh and Brunza (1975) found higher reliabilities and Hambleton, Roberts, and Traub (1970) reported higher validities (but lower reliability) for PT. Koehler (1971) and Hakstian and Kansup (1975) found identical reliabilities and validities for NC and PT. In a second study by the same authors (Hakstian and Kansup, 1975), test validity was significantly higher

under PT, but this advantage vanished after adjusting for differential administration times.

Confidence Marking (CM)

CM, a simplified version of PT, was first suggested and applied by Dressel and Schmidt (1953). Examinees are asked to express their confidence only for the most correct option by using a confidence scale with typically 3 to 5 points of reference (e.g., *unsure, reasonably sure,* and so forth). Empirical studies comparing CM with NC show a slight advantage to the former, primarily in terms of validity (e.g., Dressel and Schmidt, 1953, with 5 levels of confidence; Hopkins, Hakstian, and Hopkins, 1973, with 3 levels).

Complete Ordering (CO)

CO is another special case of PT. Instead of assigning probabilities to each option, the examinee is required to rank order them according to probabilities (e.g., De Finetti, 1965; Poizner, Nicewander, and Gettys, 1978). This method trades off precision of measurement for simplicity of administration and application.

Partial Ordering (PO)

PO is a hybrid of ET and CO. Examinees are asked to rank (as in CO) only those options that cannot be totally eliminated (as in ET) from consideration. The method was proposed by De Finetti (1965) and applied by Diamond (1975). There is little empirical work on the effectiveness of CO and PO in improving reliability or validity (Diamond, 1975; Poizner et al., 1978).

The Present Study

The purpose of the present study was two-fold: (1) to introduce a non-dichotomous scoring procedure, called *parledge,* for the MC test method, and (2) to determine the characteristics of parledge through a psychometric study comparing it with NC and CG methods. Below parledge has been introduced briefly. The term *parledge* is a blend made from the ‘partial’ and ‘knowledge.’ In the parledge answer sheet (Figure 1, right), a specific slot has been provided for test takers to admit their uncertainty (U) indicating partial knowledge. (In case the

examinee does not mark the *U* slot, this will be interpreted as full confidence.)

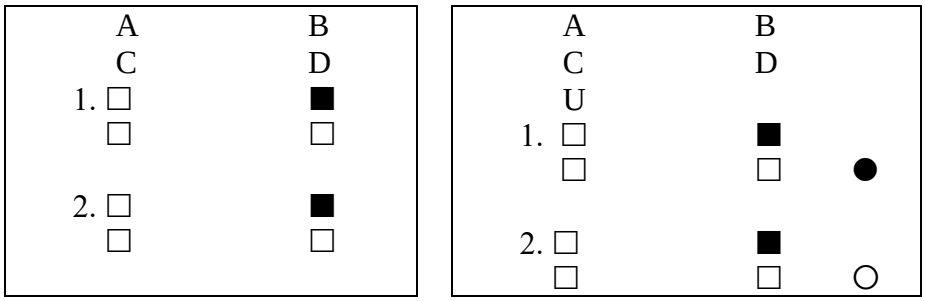

Figure 1. Conventional (left) and parledge (right) answer sheets

The general parledge formula is denoted below:

$S_p = \{R+[(K- \quad U_r\} - W+[(K- \quad U_w\}/(K-$

Where the symbols stand for the following:

S_p = the parledged score

R = the number of right responses that are given with certainty

W = the number of wrong responses that are given with certainty

U_r = the number of uncertain responses that are right (formally admitted)

U_w = the number of uncertain responses that are wrong (formally admitted)

K = the number of alternatives

Accordingly, parledge is capable of discriminating among all five levels of knowledge. Full knowledge is characterized by the selection of the correct answer with certainty. Partial knowledge is exhibited by the selection of the correct answer with uncertainty. Partial misinformation is characterized by the selection an incorrect options with uncertainty. Full misinformation is demonstrated by the selection of an incorrect answer with certainty. Absence of knowledge is characterized by omission of an item. Figure 2 compares the conventional and parledge answer sheets.

Methodology

A group of 35 sophomore students majoring in English literature participated in the study. Two tests of vocabulary and structure, each with 20 items, were composed from a pool of items drawn randomly from a number of standardized MC tests. The criterion for adequacy of test content was appropriateness of item facility to encourage maximum guessing. In a pilot study, the minimum and maximum item facility indexes for the vocabulary test were 0.10 and 0.70, with an average of 0.45. The indexes for the structure test were 0.10 and 0.60, with an average of 0.39. The reliability indexes for the vocabulary and structure tests were 0.84 and 0.88, respectively.

The examinees were instructed to answer the items on three different response sheets each presenting them with one of the following three instructions.

1) NC instruction: You may guess when you are not certain of your answer or you don't know the correct option. Your score will be the sum of the number of the correct responses (i.e., no penalty)

2) CG instruction: You may guess when you are not certain of your answer or you don't know the correct option. You will get one point for your correct guesses but lose one third of a point (1/3) for your wrong ones. However, if you omit the question, you will neither gain nor lose points. Your score will be the number of the correct responses minus one third (1/3) of the number of incorrect ones.

3) Parledge instruction: In the parledge answer sheet, specific slots have been provided for you to announce your probable uncertainties (U). You may guess when you are not certain of your answer or you don't know the correct option. Also you may choose to acknowledge your uncertainty of your answer by filling in the U-slots. Your score will be the sum of the correct responses (certain: 1 point, and uncertain: 3/4 points), subtracted by the sum of incorrect responses (certain: 1/3 point, and uncertain: 1/4 point).

Results

There were 35 test takers each responding to 20 vocabulary questions, altogether 700 MC items. The participants answered all the items under NC and parledge. However, performance under CG revealed

250 omissions out of 700 responses Moreover, NC and CG showed highest and lowest means, 12.11 and 7.87, respectively. Also, the parledge variance was greatest. Table 1 demonstrates the basic statistics results.

Table 1. Mean and SD indexes for the three measures (Vocabulary Test)

Scoring Methods	Omissions	Omissions%	Mean	SD
NC	-	-	12.11	2.61
Par ledge	-	-	9.20	2.94
CG	250 (from 700)	36%	7.87	2.68

To determine whether mean differences were real, a test of repeated measures was used. The Pillai's Trace value was .871 ($F_{(2, 33)}$ =111.73, p<001), and the F ratio for the test of within-subjects contrasts was 148.57, indicating highly significant differences at least between two of the performances. To find out the whereabouts of differences, the paired t test was used; the level of significance was set at 0.01 as a result of applying the Bonferroni adjustment. All the comparisons were highly significant as shown in Table 2.

Table 2. Pair t-test and correlations among the three measures

Scoring Methods	Pairs	Cor.	2-tail sig.	t-value	2-tail sig.
NC vs. CG	35	0.78	0.000*	14.32	0.000
NC vs. Parledge	35	0.88	0.000*	12.19	0.000
CG vs. Parledge	35	0.86	0.000*	5.22	0.000

Analogous to the findings in the vocabulary test, there were no omissions under NC and parledge in the structure test. However, performance under CG revealed 175 omissions out of 700 items on the whole. The NC and CG procedures produced the highest and lowest means, 12.49 and 7.20, respectively. Again parledge produced the highest variance. Table 3 shows the basic statistics for the test of structure.

Table 3. Mean and SD indexes for the three measures (Structure Test)

Scoring Methods	Omissions	Omissions%	Mean	SD
NC	-	-	12.49	3.11
Parledge	-	-	9.70	4.11
CG	175 (from 700)	25%	7.20	4.02

A test of repeated measures was used to determine within-subject differences. The Pillai's Trace value was .871 ($F_{(2, 33)} = 111.73$, $p<001$), and the F ratio for the test of within-subjects contrasts was 80.70, providing evidence for statistically significant differences at least between two of the means. The follow-up paired t tests with adjusted level of significance (0.01) demonstrated that the performance of the same test takers were meaningfully different, as displayed in Table 4 below.

Table 4. Paired t-test and correlations among the three measures

Instructions	Pairs	Cor.	2-tail sig.	t-value	2-tail sig.
NC vs. CG	35	0.95	0.000*	21.67	0.000*
NC vs. Parledge	35	0.91	0.000*	8.98	0.000*
CG vs. Parledge	35	0.90	0.000*	8.12	0.000*

Discussion

The assumptions behind parledge fall between NC and CG. Parledge does not treat all correct answers as representing full knowledge, neither all wrong answers as pure random guesses or signs of absence of knowledge. In this study, a number of features were found for parledge which have been discussed below in light of the literature.

Not Discouraging Guessing in Spite of Involving Penalty Threat

In a psychometric study, Ben-Simon, et al., (1997) compared NC, CG, ET, PT, CM, CO, and PO procedures. Four of the methods (CG, PO, CM and ET) penalized incorrect answers. They found that the threat of penalty induced higher levels of omission. The findings of their study were partially corresponding to those of this study. Under CG, a dichotomous procedure involving penalty threat, 36% and 25% of the vocabulary and structure items, respectively, were not performed due

to penalty threat. Moreover, all items under NC were performed. However, an important difference between parledge and other non-dichotomous scoring methods involving penalty (PO, CM and ET) was the rate of omission. While under these procedures, test takers avoided some items, all items were taken under parledge. Therefore, it was found that parledge did not discourage guessing—as opposed to CG or other non-dichotomous procedures like PO, CM and ET—although it involved penalty like the latter methods.

Increasing Response Rate Despite Penalty Threat

It is normally the case that examinees perform all items under NC. Based on similar findings, Ben-Simon, et al., (1997) proposed that penalizing methods should not be used if it is important to increase response rate (i.e., minimize omission rate). Betts, Elder, Hartley, and Trueman (2009) also found that fewer questions were unanswered, when there was no correction for guessing. However, this study showed somehow the contrary. The participants did not miss any questions under parledge although there was a penalty threat. In other words, there was no difference between the number of attempted items under NC and parledge. This might be to say that parledge acted like NC although it involved penalty.

Increasing Variance

Comparing NC, CG, ET, PT, CM, CO, and PO procedures, Ben-Simon, et al., (1997) determined that methods involving penalty induced higher variance for scores. This was true when parledge variance was compared with that of NC. More interestingly, in both tests of vocabulary and structure, parledge showed greater variance than CG. This shows that, holding penalty threat constant, non-dichotomous procedures produce higher variance than that of dichotomous methods.

Producing Better Discrimination

Under the NC and CG procedures, the test takers, when in doubt, were captive between two alternatives, performing or not. However, parledge encouraged a more sophisticated performance behavior by

giving the test takers two more choices: expressing their doubts or not, in addition to the previous alternatives. For this reason, parledge displayed the potentiality of producing better discrimination among test takers' levels of knowledge. This is corresponding to Ben-Simon, et al., (1997) position stating that non-dichotomous response methods discriminate between all levels of (partial) knowledge.

Producing Fairer Scores

Parledge offered the students a fairer opportunity of performing in a variety of ways. The data showed that in the vocabulary and structure tests 50% and 32% of the responses, respectively, were given within this specific feature of parledge. However, NC and CG deprived test takers from this obvious right, imposing a black-or-white response pattern onto their performance. Furthermore, informal talks with the testees, after the test, verified that they felt that parledge was a fairer procedure, giving them more freedom of performance. This is corresponding to Coombs et al. (1956) and Jaradat and Tollefson (1988) who found similar results for ET.

Encouraging Cautious Risk-taking

The examinees performing under the other two procedures were either encouraged to guess, no matter informed or wild, or were discouraged from guessing. However, parledge encouraged the examinees to take cautious risks by giving them a chance to acknowledge them. This was evidenced by 50% and 32% of the responses which were admitted as *uncertain* in the vocabulary and structure tests, respectively.

Revealing Information about Guessing Behavior

Parledge revealed some information about examinees' guessing behavior. The NC and CG procedures, however, openly failed to do so in which the rater never knew who guessed and which items were guessed at as well as how frequently guessing happened.

Conclusion

In their paper, Ben-Simon, et al., (1997) conclude that the epistemological and psychological foundations of the popular dichotomous model of knowledge are invalid. Thus, they propose that

the model should be rejected as a possible basis for developing accurate psychological and educational tests. In the same line, pedagogically speaking, parledge seems to have some potential diagnostic values because it can give teachers a pattern of test takers' degree of knowledge on each item. Moreover, from a real-life point of view, parledge appears to encourage a more realistic performance (or response) behavior. To elaborate, in the real world, people, besides making statements with certainty or avoiding making statement, normally express doubts when they do not possess full information about a specific issue. The parledge attempts to introduce this real-world feature of communication into the performance behavior of examinees on MC items by design. Finally, from a cognitive perspective, the parledge frame might be expected to gradually influence and re-shape students' strategic approaches toward responding MC test items, letting them be more principled, cautious, and rational guessers.

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*, 323-335.

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Ben-Simon, A. (2000). *Cognitive aspects of partial knowledge as expressed in multiple-choice tests*. Paper presented at the NCME annual meeting, New Orleans, LA.

Ben-Simon, A., Budescu, D. V., and Nevo, N. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*, 65-88.

Betts, L. R., Elder, T. J., Hartley, J., and Trueman, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? *Assessment and Evaluation in Higher Education, 34 (1),* 1-15.

Bliss, L. B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. *Journal of Educational Measurement, 17*, 147-153.

Chevalier, S. A. (1998). *A review of scoring algorithms for ability and aptitude tests*. Paper presented at the annual meeting of the southwestern psychological association, New Orleans, LA. (Eric Document no. ED 417220)

Choppin, B. H. (1990). Correction for guessing. In H. J. Walberg and G. D. Haertel, (Eds.), *The international encyclopedia of educational evaluation* (pp. 345-348). Oxford: Pergamon Press.

Collet, L. S. (1971). Elimination scoring: An empirical evaluation. *Journal of Educational Measurement, 8*, 209-214.

Coombs, C. H., Milholland, J. E., Womer, F. B. (1956). The assessment of partial knowledge. *Educational and Psychological Measurement, 16*, 13-37.

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

De Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology, 18*, 87-123.

Diamond, J. J. (1975). A preliminary study of the reliability and validity of a scoring procedure, based upon confidence and partial information. *Journal of Educational Measurement, 12*, 129-133.

Dressel, P. L. and Schmidt, J. (1953). Some modifications of the multiple-choice item. *Educational and Psychological Measurement, 13*, 574-595.

Frary, R. B. (1980). The effect of misinformation, partial information, and guessing on expected multiple-choice test item scores. *Applied Psychological Measurement, 4*, 79-90.

Hakstian, A. R. and Kansup, W. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedures. *Journal of Educational Measurement, 12*, 219-230.

Hambleton, P. K., Roberts, D. M., and Traub, R. E. (1970). A comparison of the reliability and validity of two methods for assessing partial knowledge on a multiple-choice test. *Journal of Educational Measurement, 7*, 75-82.

Harris, D. P. (1969). *Testing English as a second language*. New York: McGraw-Hill Book Company.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge: Newbury House Publishers.

Hopkins, K. D. and Hakstian, A. R., Hopkins, B. R. (1973). Validity and reliability consequences of confidence weighting. *Educational and Psychological Measurement, 33*, 135-141.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Jaradat, D. and Tollefson, N. (1988). The impact of alternative scoring procedures for multiple-choice items on test reliability, validity, and grading. *Educational and psychological Measurement, 48*, 627-635.

Kline, P. (2000). *Handbook of Psychological testing* (2nd ed.). London: Routledge.

Koehler, R. A. (1971). A comparison of the validities of conventional choice testing and various confidence marking procedures. *Journal of Educational Measurement, 8*, 297-303.

Kurz, T. B. (1999). *A review of scoring Algorithms for multiple-choice tests*. Paper presented at the annual meeting of the southwest educational research association, San Antonio, TX. (Eric Document no. ED 428076)

Michael, J. C. (1968). The reliability of a multiple-choice examination under various test-taking instructions. *Journal of Educational Measurement, 5*, 307-314.

Mousavi, S. A. (1999). *A dictionary of language testing* (2nd ed.). Tehran: Rahnama Publications.

Mousavi, S. A. (2002). *An encyclopedic dictionary of language testing* (3rd ed.). Taiwan: Tung Hua Book Company.

Poizner, S. B., Nicewander, W. A., and Gettys, C. F. (1978). Alternative response and scoring methods for multiple-choice items: An empirical study of probabilistic and ordinal response modes. *Applied Psychological Measurements, 2*, 83-96.

Prieto, G., and Delgado, A. R. (1999). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment, 15* (2), 143-150.

Prihoda, T. J., Pinckard, R. N., McMahan, C. A., and Jones, A.C. (2006). Correcting for Guessing Increases Validity in Multiple-Choice Examinations in an Oral and Maxillofacial Pathology Course. *Journal of Dental Education, 70 (4)*, 378-386.

Pugh, R. C. and Brunza, J. J. (1975). Effects of a confidence weighted scoring system on measures of test reliability and validity. *Educational and Psychological Measurement, 35*, 73-78.

Slakter, M. J. (1968). The penalty for not guessing. *Journal of Educational Measurement, 5*, 141-144.

Smith, R.M. (1990). Validation of individual test response patterns. In H. J. Walberg and G. D. Haertel, (Eds.), *The international encyclopedia of educational evaluation* (pp. 325-329). Oxford: Pergamon Press.

Thorndike, R. L. (Ed.) (1971). *Educational measurement.* Washington, DC: American Council on Education.

Tweed, M and Wilkinson, T. (2009). A randomized controlled trial comparing instructions regarding unsafe response options in a MCQ examination. *Medical Teacher, 31 (1),* 51-65.

Weir, C. W. (1990). *Communicative language testing*. New York: Prentice Hall.