DOCUMENT RESUME

ED 298 171 TM 012 320

AUTHOR Plake, Barbara S.; Melican, Gerald J.
TITLE Prediction of Item Performance by Expert Judges: A Methodology for Examining the Impact of Correction-for-Guessing Instructions on Test Taking Behavior.
PUB DATE [85]
NOTE 14p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS College Students; *Guessing (Tests); Higher Education; Mathematics Tests; Minimum Competency Testing; *Multiple Choice Tests; *Prediction; Professors; *Scoring Formulas; *Test Construction; Testing Problems; Test Items
IDENTIFIERS *Correction for Guessing; Nedelsky Method; Professional Judgment

ABSTRACT

A methodology for investigating the influence of correction-for-guessing directions and formula scoring on test performance was studied. Experts in the test content field used a judgmental item appraisal system to estimate the knowledge of the minimally competent candidate (MCC) and to predict those items that the MCC would omit on the test under correction-for-guessing and formula scoring directions. The ability of these experts to anticipate test behavior correctly was examined using a 28-item five option multiple choice mathematics achievement test requiring a clear statement when an item was omitted. A total of 156 students at a large midwestern university in the fall of 1985 took the examination. Ten professors using the Nedelsky method served as judges. The judges did a reasonable, although imperfect, job of picking items to be omitted. Results suggest that expert judgment may be used to assist researchers in formula scoring studies. (SLD)

Prediction of Item Performance by Expert Judges: A Methodology for Examining the Impact of Correction-for-Guessing Instructions on Test Taking Behavior

Barbara S. Plake

University of Nebraska

Gerald J. Melican

Educational Testing Service

Running Head: EXPERT PREDICTION

Prediction of Item Performance by Expert Judges: A Methodology for Examining the Impact of Correction-for-Guessing Instructions on Test Taking Behavior

The effect of guessing on test performance for multiple-choice examinations has been a source of concern for measurement and testing experts since the inception of multiple-choice testing. Directions warning against guessing (correction-for-guessing instructions) and scoring algorithms (e.g., formula scoring) have been developed to correct for guessing on multiple-choice examinations. Research and discussions concerning correction-for-guessing instructions and formula scoring have centered on two related areas: one, the effect of formula scoring on the psychometric properties of examinations and two, the effects of formula scoring instructions on examinee test taking strategies. Lord (1963) indicated that formula scoring should result in increased validity and Mattson (1965) suggested that formula scoring should result in increased reliability. Lord (1975) indicated that the sampling error of the corrected score associated with formula scoring will be smaller than the sampling error of the raw score associated with number right scoring, if the assumption underlying formula scoring is met.

Most empirical studies, however, have failed to substantiate the claims of superior psychometric properties for formula scoring (Michael, Stewart, Douglass and Rainwater, 1963; Sabers and Feldt, 1968; Traub, Hambleton, and Singh, 1969). The lack of

substantiation for the assertions of superior psychometric properties of formula scoring may be due to the inadequacy of the assumptions underlying formula scoring.

The original assumption for formula scoring was that examinees either knew the answer or randomly guessed among the choices. Lord (1975) suggested a different assumption, one he and others have found to be more tolerable. This assumption advanced by Lord was that the only difference between a formula scoring answer sheet and a number right answer sheet is that the omitted items of a formula scoring answer sheet are replaced by random guesses on a number right answer sheet. Lord's assumption was necessary because the original assumption did not describe examinee behavior accurately. Examinees often have partial knowledge and their responses were shown to deviate from random marking.

Translating Lord's assumption for formula scoring into test directions for the examinee requires instructing examinees to guess only on items when they have no knowledge of the correct answer. That is, examinees are instructed to guess if they can identify even one of the alternatives as incorrect. Lord's assumption does not hold, therefore, for examinees to refuse to guess even if they have partial knowledge and, for such examinees, formula scoring acts as a penalty (or over-correction). This penalty is not one associated with their knowledge but with their test taking strategy. As noted by Rowley and Traub (1977) utilization of formula scoring based on Lord's assumption depends

on that assumption being appropriate for all examinees. Studies by Sherriffs and Boomer (1954), Cross and Frary (1977), and Bliss (1980) suggest that there are examinees who fail to answer items on tests with no guessing instructions for which they have a better than chance probability of answering correctly. The results of these studies are not supportive of Lord's assumptions concerning test taking (guessing) behavior.

With the less than positive results of both the psychometric and the test taking strategy studies one wonders about the advisability of formula scoring. In fact, Cross and Frary (1977, 1985) and Bliss (1980) have questioned the ethics of formula scoring because of the negative effects the procedure has on some examinees, effects not associated with the variable supposedly being measured by the examination. It should be mentioned that the examinations reviewed tended to be power, not speeded, examinations; under speeded conditions Lord's assumption is more likely to hold for the not reached items and that formula scoring may be appropriate. The question of advisability of formula scoring power tests is still not answered.

Diamond and Evans (1973) commented that the public may be unreceptive to test results where it is perceived that scores may be increased by wild guessing on items that are unknown to the examinee. With the increased emphasis on testing for licensure and certification in professional fields this comment is very germane. The public needs to be assured that an insurance agent

or a respiratory therapist is not in practice due to the unlikely event of a particularly good day guessing on the licensure examination and formula scoring attempts to satisfy this concern. Lord (1975) further suggests that, if it is shown that examinee behavior is inconsistent with his test taking assumption, an effort should be undertaken to teach appropriate test taking skills to students.

Research on formula scoring needs to continue, with emphasis on identifying the types of examinees who are affected negatively by formula scoring, on the kinds of response sets (tendencies to respond differently under one set of instructions vs. another set) that are active for different content areas, on the types of instructions that minimize or eliminate incorrect response sets, and on educating examinees on appropriate test taking strategies. These research efforts will need to be performed for examinees of achievement and aptitude tests at each educational level as well as for tests used for purposes of licensure and certification.

Several studies have attempted to investigate the influence of correction-for-guessing directions and formula scoring on test performance. One approach is to have examinees initially respond to the test questions under no-guess directions. Subsequently, the examinees are administered items they previously omitted, under no-correction-for-guessing directions. If, upon readministration, the examinee has a higher than chance probability of getting the item correct, the conclusion made is

that the examinee had partial knowledge at the time of initial presentation and, therefore, should not have omitted the item. Plake and Wise (1985) provided a critical analyses of such "within group methodologies" for examining test taking behavior and concluded that competing hypotheses (such as learning or insight) could not be ruled out as potential explanations in addition to partial knowledge. Therefore, this methodology in its current state, is flawed (see also Angoff, 1985).

One possible research strategy which has the potential of providing information on test taking strategy under correction-for-guessing instructions is to utilize experts in the test content field by adapting a judgmental item appraisal system traditionally used for establishing passing scores on licensure and certification examinations (e.g., Nedelsky (1954)) expert prediction of test performance by a particular type of examinee ("minimally competent candidate"). Performance items that the judges predicted the MCC to have partial knowledge could be examined to investigate whether the MCC actually responded to or omitted the item. Further, performance by MCC's on items that were predicted by the judges to be omitted (therefore, ones the judges anticipated the minimally competent candidate had no partial knowledge) should actually be omitted by the minimally competent candidates.

In order for this methodology to be useful, however, it must first be established that experts can in fact predict the omit

behavior of minimally competent candidates. Melican and Plake (1985) reported results that suggest that judges using the Nedelsky (1954) standard setting method were able to do a reasonable job of identifying items that would be omitted by minimally competent candidates.

The purpose of this study was to extend the Melican and Plake (1985) research to provide a clearer picture of the ability of experts to anticipate the omit behavior and partial knowledge of minimally competent candidates.

Procedure

A 28-item five option multiple-choice Mathematics Achievement Test (MAT) was developed to identify students in Introductory Statistics classes at a large midwestern university who would benefit from a mathematics remediation lab. To focus the examinees attention on the decision to omit, so that omission was a conscious act, an additional option was added to each five-alternatives set which stated "I OMITTED THIS ITEM." This instrument was administered with correction-for-guessing instructions to a total of 156 students in the fall, 1985. The 28-item test had a mean of 12.92, a standard deviation of 6.29, and KR (20) reliability of 0.88.

A cut-score study was performed using the Nedelsky method with ten professors serving as expert judges. These professors regularly taught the introductory statistics course for which the MAT was developed and were familiar with the level of mathematics

required of a minimally competent student. The judges independently rated each item, indicating which of its incorrect options would be identified as incorrect by the minimally competent student. The suggested cut-score to identify subjects in need of remediation was 11 (11.18 rounded to nearest whole number, the average of the 10 judges individual estimates). The individual judges' estimates ranged from 8.00 to 19.40 with a standard deviation of 3.84. The intraclass correlation estimate of the reliability of this average was .55. Items were arbitrarily identified as "omit" items if four or more of the ten judges indicated that the minimally competent student would be unable to eliminate any incorrect option as obviously incorrect. Eight of the 28 items were identified as likely to be omitted by the minimally competent student.

The empirically determined minimally competent candidate was defined as an examinee who fell within one standard error of measurement (approximately three raw score units) of the cut-score. An item analysis was performed for these 58 subjects and any item for which there were more than 14 omits (25%) was identified as a high omit item. The cross tabulation of predicted omit/non-omit and observed omit/non-omit items is presented in Table 1.

Of the 8 items identified by the judges to be omitted by the MCC, 6 (75%) of them were actually omitted by more than 25% of the minimally competent examinees. Further, of the 20 items for which

the judges predicted the minimally competent candidates would have partial knowledge, 12 (60%) of them were omitted by less than 25% of the MCC's. However, 8 of these items (40%) were omitted by more than 25% of the MCC's even though the judges predicted they would have partial knowledge.

Discussion

Utilizing prior ratings of expert judges allows for an external criteria for identifying items particular examinees should or should not omit on examinations utilizing correction-for-guessing instructions and formula scoring. This approach avoids the confounding present in within group methodologies which have been previously used to examine the impact of formula scoring on test performance. These results are consistent with the previous findings in that they suggest that judges can do a reasonable, if imperfect, job of identifying items to be omitted. These results suggest that expert judgments about which items may be omitted under formula scoring conditions may be used to assist researchers in performing formula scoring research. Expert judgments may be used to establish which items would be omitted on the basis of knowledge only and, the items which do not fit the experts predictions, may be studied in detail to ascertain the presence of response sets. Personality inventories, as used in previous research, may be administered to experts and examinees alike in an effort to explain the inconsistencies between the experts prediction and the examinees performance. These research areas

need to be reviewed for aptitude and achievement tests as well as licensure and certification to ascertain whether there are differences in motivation and response sets over these situations.

The research on these topics will also have an effect on the ways in which cut-score studies are performed. The answers to questions regarding the lack of agreement between experts and examinees may lead to improved methods of performing cut-score studies of the Nedelsky type. It may be that the ability to define a minimally competent candidate may be enhanced by studying the responses of examinees who have taken similar tests and failed to fit the model.

Other methodologies for studying guessing require large samples of examinees (see Angoff & Shrader, 1985) and, possibly, additional responses from the examinees. With the expert judgments as an initial indication of which items should not be omitted based on examinee knowledge, it may be possible to identify the types of examinees who exhibit a different test taking strategy. Similarly, the types of items which do not fit the anticipated pattern may be identified using fewer examinees than required by other methodologies and without the inconvenience of having examinees artificially respond to items previously omitted. This methodology may be used alone or in conjunction with others to investigate most questions concerning guessing behavior.

References

Bliss, L. B. A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. Journal of Educational Measurement, 17(2), 147-153.

Cross, L. H. & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests. Journal of Educational Measurement, 14(4), 313-321.

Diamond, J. & Evans, W. (1973). The correction for guessing. Review of Educational Research, 43(2), 181-191.

Lord, F. M. (1963). Formula-scoring and validity. Educational and Psychological Measurement, 23, 663-672.

Lord, F. M. (1975). Formula-scoring and number-right scoring. Journal of Educational Measurement, 12(1), 7-12.

Mattson, D. (1965). The effects of guessing on the standard error of measurement and the reliability of test scores. Educational and Psychological Measurement, 25, 727-730.

Melican, G. J. & Plake, B. S. (1984). Are correction for guessing and Nedelsky's standard setting method compatible? Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Michael, W. B., Stewart, R., Douglass, B., & Rainwater, J. H. (1963). An experimental determination of the optimal scoring formula for a highly speeded test under different instructions

regarding scoring penalties. Educational and Psychological Measurement, 23(1), 83-99.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Rowley, G. L. & Traub, R. E. (1977). Formula scoring, number-right scoring and test-taking strategy. Journal of Educational Measurement, 14(1), 15-21.

Sabers, D. L., & Feldt, L. S. (1968). An empirical study of the effect of the correction for chance success on the reliability and validity of an aptitude test. Journal of Educational Measurement, 5(3), 251-258.

Sherriffs, A. C. & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? Journal of Educational Psychology, 45(1), 81-90.

Traub, R. E., Hambleton, R. K., & Singh, B. (1969). Effects of promised reward and threatened penalty on performance on a multiple-choice vocabulary test. Educational and Psychological Measurement, 29, 847-861.

Table 1

Cross tabulation of Omit/Non-omit Status by Prediction by Judges and Observation

	Judges Ratings	
Percent Omitted by Examinees	Omit	Not-omit
Less than 25%	2	12
More than 25%	6	8