Investigating Scoring Procedures in Language Testing

George S. Ypsilandis, Aristotle University of Thessaloniki, Greece
Anna Mouti, University of Thessaly and Aristotle University of Thessaloniki, Greece

Abstract: One among the main concerns of language testers in the design and implementation of tests is to select the method of scoring for the tool used to perform the evaluation. This attribute indirectly reveals the tester’s ethical beliefs and personal stance in testing pedagogy. This study challenges the typical 1-0 method of scoring in Multiple Choice Tests (MCT) and implements a polychotomous partial-credit scoring system in official tests administered for the Greek State Certificate of Language Proficiency (GSCLP). The MCT items chosen were completed by a total of 1,922 subjects in different levels of the GSCLP test. Results clearly indicate that this scoring procedure provides refined insights to students’ interlanguage level and enhances sensitivity in scoring procedures without jeopardising test reliability.

1 Introduction

Fairness is considered a “fundamental concern” in language testing, although “describing this has proven elusive” (Bachman & Palmer, 2010, p. 127). Among the characteristics of fairness in language testing that have been discussed are those reported by Kunnan (2004): a) absence of bias, b) equity of access, c) validity of test scores, d) administration and e) impact. Bachman & Palmer (2010, p. 128) also suggested: f) equitable treatment of test-takers in the testing process, g) equality of testing outcomes for different groups, and h) equity in opportunities to learn the content that is measured in an achievement test. These tactics, mostly of a political aspect about testing, show that fairness covers a very large field of conceptualisation and the researchers/testers may deal with different characteristics, relevant qualities and perhaps measureable attributes. Some of those are related to test impact (e.g. g), planning (e.g. a+b), administration (e.g. d), or to the method of scoring (e.g. a), while one with general policies in education (e.g. h). It becomes apparent that fairness is a complicated issue and cannot be attributed to a test by a yes/no answer and thus “the best way to ensure test fairness is to build fairness into the development, administration, and scoring processes” (Zieky, 2002, p. 2). Despite its complexity, fairness is a fundamental characteristic of language testing because “irrespective of whether language assessments are used appropriately or inappropriately, they serve as both door-openers and gatekeepers” (Bachman and Purpura, 2008, p. 456).

One of the item formats commonly used in language tests, by which decisions are made, is the Multiple Choice (MC) which typically requires a selected response from among the choices provided. Two types are typically identified in MC selected response tests (Bachman and Palmer, 1996, pp. 202-203); best answer for the task types “in which the test-taker is expected to choose the best answer from among the choices given” (p. 202), and correct answers “which implies that there is only one correct answer in the world, and that this answer is among the choices provided”. In the selected response best answer type, the level of item difficulty is defined by the quality of the distractors and the plausibility of the synonymous options. Arguably, a MC item could be constructed either with the correct answer standing out of the other options (which are equally and totally wrong) or with a less transparent correct answer and more plausible alternative options to distract the test-takers. In MC tests of single correct answer type, the scoring pattern 1-0 or 1-(-1) is usually followed. In MC tests of best answer type however, the
above pattern may be found rather insensitive and probably unfair, as those test-takers who select a closer synonym against those who select a totally irrelevant answer are not rewarded.

This study lies within the area of test scoring procedures and applies to all test-takers alike. An experimental polychotomous partial credit scoring system is implemented and compared with the traditional dichotomous scoring procedure. It is hypothesised that this scoring system may provide a more refined score of the test-takers’ performance and thus mirror his/her interlanguage stage. This sensitive scoring approach is expected to increase test reliability without jeopardising test results. By that respect fairness may be served.

2 Scoring and interpreting the test scores

Test results are calculated to produce some form of final score for each test-taker. A common method of scoring, provided there is no item weighting, is to assign one point to correct responses and zero points to the wrong ones. In particular, in selected response MC items this is considered the norm, while in constructed response items (e.g. gap-filling) partial credit scoring is also considered as an option. Lau et al (2011, p.101) state that “the recognition of partial knowledge leads to the belief that a student’s level of knowledge falls on a continuum ranging from full knowledge to full misconception.” The authors review various scoring methods to credit partial knowledge: confidence weighting, elimination testing, subset selection testing, probability measurement, answer-until-correct, option weighting, item weighting, rank ordering the option, and partial ordering. Typically, distractors are equally weighted with 0, in 0-1 scoring, although there is differential information in them (Haladyna’s, 2004, p.219 “differential distractor functioning”). Consequently, differentially attractive distractors could provide the basis for improving scoring of item responses as they could be differentially weighted according to their approximate correctness (see also Method below).

3 Method

The study’s research design follows Tsopanoglou, Ypsilandis & Mouti’s (2014), and Mouti, Ypsilandis & Tsopanoglou’s (2013) studies where “option weighting” was used by awarding scalable points for choosing each MC option/answer/distractor. The option weighting approach may be implemented where MCQs contain distractors that are somewhat correct but not the best choice. This “weighting approach” is examined empirically by rewarding with partial credit scoring the test-takers who avoid selecting the totally irrelevant options in (polychotomous) MC items and choose a wrong although plausible option.

3.1 Participants

Two types of participants were engaged. The first group consisted of 4 native speakers/teachers and 2 proficient and experienced teachers of Italian (judges, from here on). Results from these judgements are presented below. The second group involved 1,922 test-takers who completed three Italian language tests in official settings (400 test-takers at A1–A2, 1,294 at the B1-B2 and 228 at the C1 levels). The L1 of the test-takers was Greek.

3.2 Materials
Data were collected from the Greek State Language Examinations for the Italian language (official tests in official authentic settings). The entire official test for each level included 4 papers (one for each macro skill): Speaking, Writing, Listening and Reading and Language Awareness. The study examined tests from Reading Comprehension and Language Awareness papers, from where a total of 53 dichotomously scored MC items (study sample) were extracted with 3 possible answers (1 correct and 2 wrong): 10 at the A1–A2, 15 at the B1–B2 and 27 at the C1 levels. The SPSS statistical package was used for test analysis.

3.3 Design, procedure and scoring system

In the study sample, polychotomous patterns and option weights were determined by the judges who ranked choices in a Likert scale, i.e. correct, very plausible/plausible and totally irrelevant/wrong. The polychotomous items were corrected with two modes of scoring: a) a traditional Dichotomous Scoring Method (DSM) where one (1) point is assigned for the selection of the correct answer and zero (0) points for all other choices, and b) a polychotomous scoring proposal (herewith Experimental Scoring Method, ESM) where one (1) point is provided for the correct answer, half a point (0.5) for the selection of the very plausible/plausible alternative and zero (0) points for the selection of the totally irrelevant/wrong answer.

4 Data analysis-scoring procedures

In the 53 items that were examined, divided in 5 testlets (sets of items), 67% (36 items/Facility Index = 0.65) followed a dichotomous pattern and 32% (17 items/Facility Index = 0.37) a polychotomous one. It should be pointed out at this stage that judgments were not unanimous in all cases. In 11 items (25%), the polychotomous pattern was confirmed by the judges while in 6 items the expert judges were not able to trace the correct answer (being distracted themselves) and therefore these items were also examined and included in this category. These judgments were examined and verified empirically in relation to the item analysis results and the distractor analysis: in 11 out of 17 polychotomous items, the very plausible answer/option was the one with the highest choice mean/percentage compared to the correct answer. In addition, the average choice means, for both the correct and the plausible answers, were almost the same, although it would have been expected for the correct answer choice mean to be higher (correct answer choice mean: 0.42, plausible answer choice mean: 0.44).

The items found to have a polychotomous pattern were scored with both the DSM and ESM, while the ones where no polychotomous pattern was identified by the experts were only scored in the traditional way. Results from the different scoring procedures were compared and statistical analyses with SPSS followed to offer insights in terms of correlations and differences between the scoring procedures.

A Level: 1st set of items: 10 MC ITEMS

(Mean = 5.92, SD = 2.14, Alpha = 0.56)
Option weighting and item analysis verification was attempted following the expert’s judgments. All the experts declared that all the above items were single correct answer, although the correct answer for the first item was not selected by all the experts. The specific set of items was not further examined with a polychotomous scoring method (there was only one item traced that presented traces of differentiality). It may be argued that in lower levels, degrees of incorrectness and polychotomous patterns cannot be easily applied as this would increase significantly test difficulty.

B Level: 1st set of items: 7 MC ITEMS

(Mean = 4.12, SD = 1.60, Alpha = 0.46)

All judges found that all the above items were single-correct answer items. The specific set of items was not further examined and the polychotomous scoring was not implemented, similarly to the above.

B Level: 2nd set of items: 9 MC ITEMS

(Mean = 3.39, SD = 1.54, Alpha = 0.22)

Experts recognised a polychotomous pattern in 5 items, which included a semi-correct/plausible answer. These polychotomous items proved to be more difficult to answer than the dichotomous items as indicated by the Facility Index (Dichotomous FI = 0.47> Polychotomous FI = 0.32). The statistical analysis revealed that in 4 items the plausible distractor was chosen by more test-takers instead of the correct answer. In 3 of those items the selection coincided with the one provided by the expert judges as correct! (The selection distracted the judges as well).

The scores were altered when the Experimental Polychotomous Scoring (EPS) was implemented. In particular, the Facility Indexes were increased and the differences were statistically significant: Mean TDS = 3.38, Mean EPS = 4.5. In order to investigate reliability of the ESM, the Pearson r correlation coefficient was employed (examines the relationship among variables) to compare the independent variables in twos. Bachman (2004) proposes this test to investigate relationships among different sets of test scores. This revealed that the two scoring procedures do indeed exist in a strong linear relationship to each other. In more detail the value between TDS and EPS is \( r = .937 \) (\( p \leq .001 \)) and correlation is significant at the 0.01 level (2-tailed). Thus, test results are not jeopardised. Furthermore, a paired-sample T-test showed significant differences between the two scores \( t = 72.941, df = 1.293 \) (\( p \leq .001 \)) which supports the alternative hypothesis. Observable differences are not explained by random variation and thus the EPS offers a more sensitive scoring statistically different from the TDS.

C1 Level: 1st set of items: 12 MC ITEMS

(Mean = 7.36, SD = 1.84, Alpha = 0.41)
A polychotomous pattern at 6 items was identified by the expert judges. These also proved to be more difficult to answer than the dichotomous items as indicated by the Facility Index (Dichotomous FI = 0.67 > Polychotmous FI = 0.39). In 3 items the distractor was chosen by most test-takers instead of the expected/correct answer. In 2 of these 3 cases the subjects’ erroneous selection again coincided with the one selected by the expert’s judgments! Apparently, the distractors were good, enough to mislead the native judges as well. Implementing the EPS, the scores were altered, the Facility Indexes were increased and the differences were statistically significant: Mean TDS = 7.36, Mean EPS = 8.20. This confirms again that the two scoring procedures do indeed exist in a strong linear relationship to each other. Pearson value between TDS and EPS is $r = .966 \ (p \leq .001)$, and correlation was found significant at the 0.01 level (2-tailed). A paired-sample T-test again showed significant differences: $t = 26.215, \ df = 228 \ (p \leq .001)$.

C1 Level: 2nd set of items: 15 MC ITEMS

(Mean = 8.41, SD = 2.41, Alpha = 0.49)

Experts recognised a polychotomous pattern at 5 items, which again proved to be more difficult than the Dichotomous items as indicated by the Facility Index (Dichotomous FI = 0.73 > Polychotmous FI = 0.40). The statistical analysis revealed that in all 5 items, the plausible distractor was chosen by most test-takers than the correct answer and once again their selection coincided with the expert’s judgments. Implementing the EPS system again altered the scores, the Facility Indexes were increased and the differences were statistically significant: Mean TDS = 8.22, Mean EPS = 9.46. Once again the two scoring procedures proved to exist in a strong linear relationship to each other, as the Pearson value between TDS and EPS was $r = .967 \ (p \leq .001)$ and correlation was significant at the 0.01 level (2-tailed). Similarly to the B level results above, a paired-sample T-test showed significant differences: T-test: $t=29.509, \ df=228 \ (p \leq .001)$ which again supported the alternative hypothesis.

5 Conclusion

Our hypothesis has been adequately supported by the evidence. In particular: a) the partial credit polychotomous scoring implemented has provided the expected refined understanding of the test-taker’s language knowledge and b) test reliability was not affected, as the two scoring procedures were found to be in a strong linear relationship to each other in all cases. These findings support the results of our preliminary study (Tsopanoglou et al., 2014).

In norm-referenced situations the increase at the level of scores may not have had significant impact as the test-takers’ ranking remained the same (high Pearson correlations). However, in criterion-referenced situations “where there exists a predetermined criterion for the students to meet, low scores would hurt those at the borderline” (Farhady, 1996, p. 222). It is here that our EPS would have a significant impact (supported by statistically significant T-tests).
Dichotomous items were easier than the polychotomous to answer, as the correct option in the former becomes transparent. Bachman and Palmer (1996, p. 202) indicated: “an item would be significantly more difficult if the options were closer in meaning because that would make identifying the correct answer more demanding for the test-taker”. Polychotomous patterns have been traced at the higher B2–C1 and not at the lower A1–B1 levels (the higher the more) as these often are analogous to level difficulty.

Finally, it is our belief that the EPS method, adopted in our study, may provide a more complete view of the interlanguage stage of an individual and thus it contributes to fairness and score accuracy, particularly for those test-takers who show high level of target language awareness (by choosing a plausible answer and not a totally irrelevant option, through inferencing). In support of this claim, Bachman & Palmer (1996, p. 205) recommended that test-takers should be encouraged to make informed guesses and that “this should be rewarded, preferably through partial credit scoring”.

References


Further Reading