Antonios Tsopanoglou, George S. Ypsilandis and Anna Mouti

Piloting a polychotomous partial-credit scoring procedure in a multiple-choice test

Abstract: Multiple-choice (MC) tests are frequently used to measure language competence because they are quick, economical and straightforward to score. While degrees of correctness have been investigated for partially correct responses in combined-response MC tests, degrees of incorrectness in distractors and the role they play in determining the test-taker’s final score remain comparatively unexplored. This pilot study examines degrees of incorrectness in MC test items and their potential impact on the overall reliability of scoring procedures, test micro impact, and scoring performance of individual test-takers. Investigating the above properties as quality dimensions of language testing is crucial since decisions made on the basis of such tests influence the careers and lives of test-takers. Three scoring procedures were applied to an authentic test: dichotomous, which is straightforward – correct/incorrect; polychotomous, which offers partial credit to very plausible and plausible options; and negative, which includes negative scoring when a wholly incorrect answer is selected. While these three procedures proved to be equally reliable, the polychotomous scoring method offered more sensitive and precise results, while negative scoring was particularly disadvantageous for test-takers. The polychotomous method of scoring moves away from the absolute right or wrong view and enhances sensitivity in the field of individualised language testing.

Keywords: multiple-choice tests, scoring procedure, language testing, dichotomous scoring, polychotomous scoring, negative scoring

DOI 10.1515/cercles-2014-0004

Antonios Tsopanoglou: Department of Italian Studies, Aristotle University of Thessaloniki, Greece. E-mail: tsop@itl.auth.gr

George S. Ypsilandis: Department of Italian Studies, Aristotle University of Thessaloniki, Greece. E-mail: ypsi@itl.auth.gr

Anna Mouti: Greek language teacher, Aristotle University of Thessaloniki and EAP-ESP Teacher, University of Thessaly, Greece. E-mail: mouti@lit.auth.gr
1 Introduction

One of the formats commonly used for language tests is multiple-choice (MC), which requires the test-taker to select one of several given responses. In 1996 Bachman and Palmer, specifying the criteria for correctness in MC tests, distinguished two types of possible response: best answer for the task types “in which the test-taker is expected to choose the best answer from among the choices given” (1996: 202), and correct answer, “which implies that there is only one correct answer in the world, and that this answer is among the choices provided” (1996: 202–203). In the best answer task type, the level of item difficulty is often defined by the degree of synonymy among the options as this is the first thing test-takers need to determine before making their final selection. In this article plausibility is the term used to describe these synonymous options. Bachman and Palmer (1996: 202) indicated that 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.

At this point a major problem appears for developers of MC language tests, because they face the following dilemma:

(a) If a MC test is constructed in which the correct answer stands out clearly from the other options (which are equally and totally wrong), the correct answer may be obvious to candidates. In other words, distractors may not function as they should and the item may lack discriminatory power, since most (if not all) test-takers may answer it correctly.

(b) If the options offered share a degree of synonymy, they can actually distract (i.e. they can be chosen by individuals who do not know the correct answer), and thereby contribute to the test’s discriminatory power, but in that case a more delicate scoring method may be required.

In MC tests that belong to category (a), a single-criterion method of scoring is traditionally applied and in some cases negative scoring as well; that is, responses are scored 1/0 or 1/−1. In MC tests that belong to category (b), however, this scoring pattern may be found rather insensitive, because those test-takers who select a more plausible response are penalized to the same extent as those who select a totally irrelevant one. Bachman and Palmer (1996: 204) argue that tests of this kind “lend themselves particularly well to partial credit scoring” (1996:204), despite the fact that traditionally they have been scored “according to a single criterion”.

The MC format is widely used in tests that determine initial or further academic careers (impact at a micro level; see Bachman and Palmer, 1996: 30). GMAT and TOEFL, for example, are frequently used for purposes of academic
“gatekeeping”, for example in the United States. Cronbach considered this use of tests “as an impartial way to perform a political function – that of determining who gets what” (Cronbach, 1984: 5, cited by Bachman, 1990: 280), while McNamara (2000: 72) used the term “impact” to describe “the effect of a test beyond the classroom . . . in the wider educational and social world”. This use of MC tests, which usually entails a single-criterion method of scoring, raises ethical questions which Briggs et al. (2006) suggest could be addressed either by formatting the possible answer choices in an orderly developmental manner or by providing a polychotomous partial-credit scoring procedure as proposed in this article.

Method of scoring and test impact are two areas in which the ethical intentions and predisposition of the tester or the test may be revealed more clearly, as well as underlying beliefs about language pedagogy. The question of test ethics has been widely discussed, but equal attention has not been paid to the ethics of scoring methods, which is (or should be) of concern to testers and test-takers alike. In section 2 we argue in greater detail for the implementation of polychotomous partial-credit scoring in MC tests as one means of responding to the ethical challenge. The article then describes a pilot project in which three scoring procedures were applied to tests taken by a small number of students in an academic environment. An experimental polychotomous partial-credit scoring procedure was implemented and its reliability and impact on final results were examined in comparison to traditional dichotomous and negative scoring procedures. We hypothesized the polychotomous partial-credit procedure could
(a) contribute to a better (more refined) understanding of the stage of interlanguage development individual test-takers had reached (differences in the final score were expected to result from the implementation of this scoring procedure);
(b) increase test reliability without jeopardizing test results in the sense of significantly altering final score ranking (no differences were expected in the final score ranking produced by the three methods of scoring).

Additionally, it is our objective to present an algorithm that could be implemented and may be further tested to increase precision and sensitivity in scoring MC tests and the decisions that arise from them.

2 Multiple-choice type tests

A typical MC test item consists of two basic parts, the stem (a question or a problem to be solved) and a list of possible answers, which usually contains one
correct (*expected*) answer and a number of incorrect options (*distractors*). Burton et al. (1991: 12–15) give five types of MC items:

(a) Items of the *single-correct-answer* variety, in which only one of the given alternatives is correct which is expected to be selected by the test-taker.

(b) Items of the *best-answer* variety, in which the alternatives differ in their degrees of correctness.

(c) Items of the *negative* variety, where the test-taker is directed to identify either the alternative that is an incorrect answer (*error*), the one that does not fit in the group (*odd one out*), or the alternative that is the *worst* answer.

(d) Items of the *multiple-response* variety, where the test-takers are expected to identify any number of correct alternatives (*more than one*), and

(e) Items of the *combined-response* variety, in which the respondent must choose the correct combination of accepted answers, i.e. *a and b, all of the above*, etc.

(f) Items of the *multiple true or false* variety, where there is a series of related true-false questions presented together as a group.

### 2.1 Guessing in MC tests

The structure of MC tests, one may assume, encourages guessing as the answers are already provided. If they do not know the correct answer or are not confident about it, test-takers need only to select the one that seems to them more likely to be correct than the others. Genesee and Upshur (1996: 58) provide support for this assumption when they explain the difference between the scores of two students in terms of chance, which they attribute to the instrument. They use the term “instrument-related reliability” “because the type of examination they took provided opportunity for guessing”. In testing theory, guessing is seen as unwelcome behaviour on the part of the test-taker. Increasing the number of choices is often the technique recommended to tackle this problem: “it is a good idea to have at least four alternative answers, so that the chance of a student guessing an answer is only 25%” (Alderson et al. 1995: 48). In addition, negative scoring is also used with some response types (particularly yes/no items) to neutralize the effects of guessing. In this case, the number of wrong answers divided by the number of alternatives minus one is subtracted from the number of correct answers (Alderson et al. 1995: 148).

Bachman and Palmer (1996: 205) distinguish two types of guessing: (a) one that is totally incidental (blind, chance or random) and is clearly to be discouraged, and (b) a second, in which partial knowledge is used to narrow down “the number of possible correct responses”, or careful thought and inference from the stem or the choices provided activates broad knowledge of the target language,
“topical knowledge and metacognitive strategies”. Although the existence of totally blind guessing is an issue that has been investigated to a certain extent (Currie and Chiramanee 2010; Rupp et al. 2006; Rodriguez 2005; Landrum et al. 1993; Rogers and Harley 1999), Alderson et al. (1995: 149) note that the negative scoring method is controversial and recommend the above formula only for cases “where blind guessing is known to occur”. Notice that blind or incidental guessing is seen as a genuine possibility. Nevo’s (1989) findings, reported by Cohen (1998: 101), on test-taking strategies in a multiple-choice test of reading comprehension are illuminating. The participants in Nevo’s study claimed to have benefited “greatly from the opportunity to become aware of how they took reading tests, having been largely unaware of their strategies prior to this study”. This suggests that learners are not likely to be fully aware of the test-taking strategies they are using, and implies that if the selection of MC responses is not consciously controlled by the individual, it is triggered almost automatically. It is thus possible to argue that totally blind guessing does not exist (the process of inferencing can be automatically activated in the human brain), unless the test-taker simply does not read the stem and the options provided. Nowadays inferencing is perceived as a valuable skill in language learning and a useful communication strategy, if we accept that it is genuinely impossible for language learners to achieve total command of a second language. Moreover, inferencing shows a high level of target language awareness (particularly in MC tests, though perhaps less in yes/no questions), so prompting and practising inferencing skills is often recommended. Indeed, Bachman and Palmer (1996: 205) recommend that test-takers should be encouraged to make informed guesses and that “this should be rewarded, preferably through partial credit scoring, and that correction for guessing would be inappropriate”.

Rather than providing partial credit scoring, Briggs et al. (2006) introduced an item format which they called Ordered Multiple Choice (OMC), with each item linked to developmental levels of student understanding, and further presented a model of student cognitive development for the construct being measured. They argued that this would enhance diagnostic interpretation of student responses and perhaps provide information about the development of students’ understanding. A similar approach was adopted by Lin et al. (2010) with a level-driven distractor analysis system for MC items: distractors were created at different comprehension levels to identify sources of misunderstanding. Lin et al. provided empirical data which indicated that their approach offered potential for useful feedback in formative evaluation and thus assisted teachers to adapt instructional and intervention strategies in their courses. Both these approaches are time-consuming, however, because they require a detailed analysis of the options provided.
2.2 Scoring and interpreting the results of tests

Test results are calculated to produce a final score for each test-taker. The most common method of scoring, provided there is no item or test component weighting, is to assign one point to correct responses and zero to wrong ones. Marks are then added to produce a total score for each subtest or for the entire test. A passing level is assigned, in many cases 50%, 60% or even 70% of all items, and a final score ranking is produced.

This type of dichotomous scoring has been an issue of debate not only in MC tests, where the test-taker selects from a list of options, but also in cloze tests, where the test-taker must think of a word that best fills the gap. Although the two test types are unrelated as regards both their structure and the mental processes required of test-takers, it is worth examining scoring methods used in cloze tests as they present an interesting analogy to the polychotomous scoring method for MC tests that we propose later. Notice that the focus is on scoring procedures and not test structure. In classical cloze procedure two methods of scoring may be used: (a) exact scoring, which gives credit only for the exact word that was deleted from the original text, and (b) acceptable scoring, which gives credit for words that are semantically and grammatically appropriate in the context. According to Ranalli (2002), classical cloze procedure is problematic. While exact scoring is quick and easy to carry out, it imposes an arbitrary standard of correctness that a test-taker cannot meet if gaps can be filled appropriately in more than one way. Ranalli argues that this explains why even educated adult native speakers may get a large number of items “wrong” when exact scoring is used. He concludes that cloze test scores may contain a substantial amount of variance that is unrelated to proficiency in the target language. It is practically impossible either to anticipate all possible acceptable answers in a cloze test, or to create a test with only one acceptable answer for each gap. If acceptable scoring is not used, test-takers may be unable to score full marks; but if it is used, the ease and speed of exact scoring will be lost and scorers may disagree on the correctness of responses (acceptability judgments are especially difficult to make when scorers are non-native speakers of the language in question). It is generally accepted that in some circumstances it may be appropriate to give partial credit for partially correct answers; and since acceptable scoring seems to be recognized as a constructive scoring method for cloze tests, it is worth experimenting with an analogous scoring concept in MC tests: more or less plausible options.

Ranalli (2002) found a strong linear relationship between exact-word (EW) and semantically-acceptable-word (SEMAC) scoring procedures in cloze tests, neither procedure significantly altering the final score. This may link directly to our hypothesis. If a SEMAC procedure does not significantly affect the final score,
and thus does not jeopardize test reliability, the same may be true of a polychotomous scoring procedure in MC tests.

Another study on scoring procedures is that of Henning et al. (1993), which compared different item formats and scoring techniques. Both MC and open-ended (OE) item formats were used, and open-ended responses were scored not only as correct/incorrect but also according to degrees of correctness and incorrectness. Henning et al. found that

when the open-ended items were scored as correct or incorrect in the same binary mode as the multiple-choice items, the multiple-choice version had a slight reliability advantage. However, as the open-ended scoring techniques increasingly took into account degrees of correctness, the reliability advantage turned in favour of the open-ended version of the test. (Henning et al. 1993: 127)

Henning et al. (1993) did not investigate the possibility of MC test items not being scored using a binary procedure. It is our view that rather than replacing MC items with OE questions, it is possible to modify multiple-choice scoring processes, following a similar arrangement to that of non-binary degrees of incorrectness.

In another study, Wongwiwatthananukit et al. (2000) utilized combined-response multiple-choice (CRMC), which involves selecting more than one response to give the correct answer, in order to accommodate and reward students with partial credit in MC questions that included responses/choices with varying degrees of completeness or correctness (i.e., most correct, somewhat correct). This scoring method aimed to optimize measurement of student knowledge without penalty for incorrect responses, by assigning weighting to the choices provided. The full point value, or assigned weight ($W$), for each item was two points, and the authors used two scoring methods: dichotomous scoring ($S_1$), where $S_1 = W(2)$ if the choice was correct and $S_1 = W(0)$ if the choice was incorrect; and partial-credit scoring ($S_2$), defined as the first best, second best, and third best answers to a question. The assigned weight for the first best answer was two points, for the second best one point, and for the third best 0.5 points ($S_2 = W_{1st}(2), W_{2nd}(1), W_{3rd}(0.5)$). Results from this study showed that the partial-credit scoring method clearly increased item difficulty as Bachman and Palmer (1996) also claim. This finding may be attributed to the fact that the distinction between most correct and somewhat correct was more difficult to identify because the possible answers were placed on a continuum of correctness. Wongwiwatthananukit et al. (2000) also found that partial-credit scoring provided a more accurate reflection of student knowledge than dichotomous scoring.

In this study, a similar pattern is followed (very plausible, plausible). However, the opposite procedure is also tested by assigning weight to incorrect though plausible choices. This is explored empirically by rewarding with partial
credit test-takers who avoid selecting wholly irrelevant options, choosing instead incorrect though plausible options. It is hoped that this will make final scores more refined because they will reflect not only what students know but also their ability to deal with the unknown.

3 Research method

3.1 Participants

Two researchers, both teachers of English, were employed to locate the items included in the test that was to be used as the main instrument for data collection. The items were then submitted to the judgement of an experienced examiner, a native speaker of English and local examiner for Cambridge Proficiency (his judgements are presented in the next section). Eighteen second- and third-year university students completed the test. They were all attending an English for Specific Purposes (ESP) course as part of their studies, ranged in age from 19 to 21, and had different levels of proficiency in English. Gender differences were not pursued as an independent variable in the study. The test was administered in paper format as part of the students’ language training in order to achieve genuine participation and an authentic response.

3.2 Materials and scoring procedures

Instead of constructing test items we collected them from an electronic TOEFL Computer Adaptive Test. The TOEFL test measures overall proficiency and does not concentrate on a single aspect, such as grammar or vocabulary. The selection of this test was deliberate, as it is a good example of an authentic “gatekeeping” test and suited the purposes of this study. It is thus treated as a control variable, i.e. it remains constant and unchanged in our experiment to test the impact on the independent variables without being measured itself. Our assistants extracted test items by running the test several times and in various modes (right/wrong) to uncover the items included in its data bank. The final protocol test (PT) for this experiment consisted of 80 MC test items, of which 72.5% followed a dichotomous and 27.5% a polychotomous pattern. This was the result of the expert’s ranking of the choices on a Likert scale – correct, very plausible, plausible, and totally irrelevant. By not informing the test takers about the impact of selecting a very plausible item we were able to obtain a genuine response and not one influenced
by our research agenda (see also further hypotheses in the discussion below). The
test was then corrected using three modes of scoring:
(a) a traditional dichotomous scoring method (DSM), where one point was
awarded for the correct answer and zero points for all other choices;
(b) a polychotomous scoring method, referred to here as the experimental scor-
ing method (ESM), where two points were awarded for the correct answer,
one point for the very plausible option, half a point for the plausible option,
and zero for the totally wrong answer (similarly to the study by Wongwiwat-
athanukit et al. 2000);
(c) a polychotomous negative scoring method (NSM), where two 2 points were
awarded for the correct answer, one for the very plausible option, zero for the
plausible option, and minus one (−1) for the totally wrong answer.

4 Results and analysis

Results are presented in two sections. In the first we compare the final scores
yielded by the three methods of scoring and compare their respective reliability.
Micro impact is examined only in the cases of students who scored close to 50%,
the pass/fail cut-off score. Clearly, the cut-off score set in the study is arbitrary
and only serves our research purposes. Central tendencies and dispersion analy-
sis follows to show frequencies and variance of test scores. In the second section
we examine the performance of the scoring methods by correlating the sum of
scores in twos for each individual participant separately for the polychotomous
items only. Finally, the performance of participants with particularly low scores
is considered and findings are presented in a bar chart.

4.1 Test scoring and correlations between the three modes of
scoring

The test was initially scored using the three methods described above. A compari-
on of the DSM and ESM scores showed that 16 out of 18 students achieved higher
ESM than DSM scores, while 2 achieved the same score in both modes. For 4 of
the participants who achieved higher ESM scores the difference was insignificant,
not more than two percentage points on the final score. The difference for 11 of the
remaining 12 participants was slightly higher, ranging from three to five percent-
age points; for 1 it was nine percentage points. A comparison of the ESM and NSM
scores showed that all participants achieved higher ESM than NSM scores. For 3
participants the difference was between three and nine percentage points on the
final score, for 8 it was between ten and nineteen percentage points, and for 7 it was between twenty and twenty-nine percentage points. Thus NSM was the least advantageous method of scoring and ESM was the most advantageous, closely followed by DSM. The Pearson r correlation coefficient was used to investigate the reliability of ESM by comparing the independent variables in twos (Bachman [2004] proposes this test to investigate relationships among different sets of test scores). It showed that all three scoring procedures are highly correlated. The correlation between DSM and ESM is .999, between DSM and NSM .959, and between ESM and NSM .956. All values are close to one (1), which is the absolute relationship, while correlation is significant at the 0.01 level (2-tailed) and the critical value is (n=18):0.590.

4.2 Micro impact

In order to investigate test impact at a micro level, where it would matter most, scores close to the cut-off point of 50% were examined. One (1) participant scored not less than 48% and not more than 49% using DSM and would have failed the test. When ESM was applied, he/she scored above 50% and thus would have passed. The 3 participants who scored 51% using DSM (and thus were at risk of scoring less than 50% and failing the test), scored more than 50% using ESM. The NSM scores of all participants were lower than their DSM and ESM scores, in 7 cases more than twenty percentage points lower. Five (5) of those 7 participants scored less than 55% using ESM. This indicates that negative scoring does not significantly affect the performance of those who score high in a MC test but has a serious impact on scores less than 55% and up to 63%.

The numbers of participants who achieved more than 50% in their final score were as follows: 14 using DSM, 15 using the polychotomous ESM, and 7 using NSM. A central tendency test for the three modes of scoring indicated that in the dichotomous and polychotomous scoring procedures the average score was above 50% of the top value, $\bar{X}=42.50/80$ and $\bar{X}=88.28/160$ respectively. This was not true, however, for the negative scoring procedure ($\bar{X}=67.61/160$). This finding is supported by the Median value (the score that splits the group in half, 50/50) for the three evaluation procedures, Mdn=40(80), 84(160), 55.5(160). This is partially explained by the Mode value (the score that was achieved by most participants or the score with the highest frequency), which is greater than 50% of the total score (respectively 41/80, 88/160) in the case of DSM and ESM, whereas in the case of NSM it is far smaller (16/160). Thus, it may be argued that a negative scoring procedure is particularly disadvantageous for students. The statistics are presented in Table 1.
Next, a dispersion analysis, which measures the spread or variation from the mean in the data sample, showed that the range for NSM shows greater dispersion (77.5%) when compared to the other two modes of scoring – the low and high scores for each scoring method were: DSM 15–73, ESM 37–146, NSM 16–140. This indicates that negative scoring does not prevent students from scoring high in a test but reduces the probability that average students will score above the mean ($\bar{X}=67.61$ for the NSM). The standard deviation values, which look at the average variability of all the scores around the mean, confirm this (DSM SD = 17.787; ESM SD = 34.376; NSM SD = 39.906), showing that the variance in DSM and ESM scores shows an equal dispersion, whereas it is more extreme in the case of NSM.

Looking at the quartile deviation (Table 2), four readily classified groups of scores were identified, in all three modes of scoring. It becomes evident that the deviation in the ESM group division is markedly higher than in the NSM group division (100% for both these groups is 160) and slightly higher in the DSM group division (where 100% is 80). This may be taken as an indication that ESM is more test-taker-friendly and fairer than the other two scoring modes, while at the same time it offers more insights about the test-taker’s interlanguage level.

<table>
<thead>
<tr>
<th>Percentiles</th>
<th>DSM</th>
<th>ESM</th>
<th>NSM</th>
</tr>
</thead>
<tbody>
<tr>
<td>25%</td>
<td>27.75</td>
<td>61.13</td>
<td>38.00</td>
</tr>
<tr>
<td>50%</td>
<td>40.00</td>
<td>85.00</td>
<td>55.50</td>
</tr>
<tr>
<td>75%</td>
<td>60.75</td>
<td>124.63</td>
<td>108.50</td>
</tr>
</tbody>
</table>

### 4.3 Further analysis

In order to explore the test-taking strategies of high and low scorers, it was decided to analyse their performance further. The Pearson r bivariate correlation coefficient was employed to look at the sums of scores for each participant separately in terms of their correct, very plausible, plausible, and totally irrelevant choices in
the polytomous items only. Statistically significant negative correlation was found between correct answers and totally irrelevant answers \((r=-.808, \text{DF}=16, p<0.01)\), whereas the correlation between correct answers and highly plausible options was low and insignificant \((r=-.332, \text{DF}=16, p<.179)\) in the case of participants who scored high in the test. This reveals that high-scoring test-takers select very few totally irrelevant options (as one would expect), while they show a tendency to be distracted by the very plausible ones. As the number of correct answers drops, the number of very plausible options selected increases, but not the number of totally irrelevant ones. The same negative correlation is found between the very plausible and plausible pair; i.e. test-takers who select very plausible options do not select plausible ones \((r=-.311, \text{DF}=16, p<.209)\). Finally, totally irrelevant and very plausible options do not correlate significantly \((r=-.131, \text{DF}=16, p<.605)\). This clearly indicates that when a high-scoring test-taker does not know the correct answer, he/she selects the one closest to it (very plausible), which further suggests that the selection is not totally coincidental (or pure luck) and may be seen as evidence of that test-taker’s more developed interlanguage.

The responses of the 7 participants who gave fewer than seven correct answers were further analysed (see Figure 1). Here it is possible to observe the same test-taking behaviour that was discussed above. When there are more correct responses (first bar) than totally irrelevant ones (fourth bar), the number of very plausible responses rises (participants 5, 6, 7). It is possible to argue that these participants are in a middling interlanguage stage, though they would not pass the test as their overall score is below the 50% cut-off point. ESM might benefit cases such as these by helping to build test-takers’ self-confidence while not changing the scores to the point of jeopardizing reliability. On the other hand, when there are more totally irrelevant answers (fourth bar) than correct ones (first bar), the number of very plausible answers drops (participants 1, 2, 3, 4). These participants are at a lower interlanguage stage and were not prepared adequately.

![Fig. 1: Bar chart showing the responses of seven participants with fewer than seven correct answers](image-url)
for the test. For cases such as these, ESM would again not significantly alter final scores.

5 Conclusion

Given that our study is small-scale and exploratory, conclusions can only be tentative; nevertheless our hypothesis has gained support in the data. In particular, (a) the partial-credit polychotomous scoring we implemented seems to have provided the expected refined understanding of the test-takers’ language knowledge, as Wongwiwatthanuikit et al. (2000) also found, and (b) test reliability was not affected, as all three scoring procedures were shown to be in a strong linear relationship to each other (cf. Ranalli 2002).

In addition, ESM was found to be more test-taker-friendly, having a favourable impact at micro level, particularly in the case of test-takers who scored very close to the set-off point; in this way it increased test fairness. Although Zieky (2002: 11) finds that “there is no magic bullet to guarantee fairness”, ESM increased sensitivity and precision in test scoring by providing more refined insights into students’ interlanguage level. It may thus be possible to conclude that if polychotomous items are used in a test, polychotomous scoring would increase the quality of results and enhance fairness. If test-takers are made aware of the scoring scheme, test ethics may also be enhanced.

On the other hand it is clear that polychotomous scoring is time-consuming if applied by a human tester. It would be easier for a computer to complete the task, provided there is a mathematical formula behind the surface structure of the test. One simple formula for partial-credit scoring could be 2–1–0–0 (2a+1b) or a more refined version (in Maple), as shown in Table 3.

<table>
<thead>
<tr>
<th>Table 3: Mathematical formula for partial-credit scoring in Maple (courtesy of Nikos Lygeros and Thomas Vougiouklis)</th>
</tr>
</thead>
<tbody>
<tr>
<td>test:= proc (A, B, G, D)</td>
</tr>
<tr>
<td>[a=A, b=B, g=G, d=D]</td>
</tr>
<tr>
<td>if a&lt;40 then 0</td>
</tr>
<tr>
<td>elif b&gt;20 then e:=b−20</td>
</tr>
<tr>
<td>d:=d−e:</td>
</tr>
<tr>
<td>fi:</td>
</tr>
<tr>
<td>2a+1b−d</td>
</tr>
<tr>
<td>end:</td>
</tr>
</tbody>
</table>

In this formula a test-taker would be required to select a minimum of 40% correct answers (to protect test reliability). 20% of very plausible and plausible selections would be rewarded. If he/she selects more than 20% of very plausible distractors then these would be used only to balance the totally irrelevant answers.
The type of negative scoring used in our study was found to be particularly
disadvantageous for test-takers. It is our view that it should not be used as it does
not contribute to the fairness of a test: it does not reward inferencing, a language
skill which could reveal the full linguistic potential of the test-taker; and it does
not recognize that choices made on the basis of pure luck are minimal, occurring
only when the test-taker does not read the stem and the options. Finally, it
should be mentioned here that a polychotomous method of scoring allows us to
take account of inferencing. The latest version of Hot Potatoes (free testing soft-
ware available on the internet) tends in this direction, since it allows polychoto-
mous scoring. Making test-takers aware that very plausible or plausible options
are given partial credit may alter their perception of the ethics of the test-taking
experience.

Among the shortcomings of this study are the low number of participants
(18) and the fact that test items were evaluated by only one judge rather than
by a group of experts. We have designed a larger-scale study to be undertaken
in the near future. Hypotheses that would need to be investigated further may
be related to measurement, the impact of a test containing more polychotomous
items, and participants’ attitude to the test once they have been made aware that
a polychotomous scoring method is being used. Among the questions to explore
are: “Would awareness of partial-credit scoring increase test-takers’ involvement
and sense of responsibility in answering the questions?” and “Would final results
change significantly?”

Acknowledgments: The authors wish to thank Phil Hubbard of the University of
Illinois, USA, for his valuable comments on an earlier version of this article, and
Nikos Lygeros of the University of Lyon, France, and Thomas Vougiouklis of the
Democritus University of Thrace, Greece, for preparing the mathematical formula
in Maple.

References

Alderson, J. Charles, Caroline Clapham & Dianne Wall. 1995. Language test construction and
evaluation. Cambridge: Cambridge University Press.
University Press.
University Press.
University Press.


Bionotes

*Antonios Tsopanoglou* is an emeritus professor of language didactics at the Aristotle University of Thessaloniki, Department of Italian Studies, and a designer of the Greek Foreign Language Exam System (KPG).

*George Ypsilantis* is an associate professor of didactics at the Aristotle University of Thessaloniki, Department of Italian Studies, with an MA from the University of North Wales, Bangor (UK) and a PhD from the Aristotle University.

*Anna Mouti* holds a BA in English and Italian Studies, an MA in Language Didactics, and a PhD in Language and Communication Sciences from the Aristotle University of Thessaloniki, Greece. She is a teacher of EAP/ESP and Greek.