Ypsilandis, G.S., & Mouti, A. (2019). Fairness and Ethics in Multiple Choice (MC) Scoring: An Empirical Study. *Journal of Language and Education, 5*(1), xx-xx. doi: 10.17323/2411-7390-2019-5-1-xx-xx

Fairness and Ethics in Multiple Choice (MC) Scoring: An Empirical Study

George S. Ypsilandis
Aristotle University of Thessaloniki

Correspondence concerning this article should be addressed to George S. Ypsilandis, Department of Italian Language and Literature, Aristotle University of Thessaloniki, University Campus, 54124, Thessaloniki, Greece. E-mail: ypsi@itl.auth.gr

Anna Mouti
University of Thessaly
Aristotle University of Thessaloniki

Correspondence concerning this article should be addressed to Anna Mouti, Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia Campus, Papasiopoulou 2-4 Street, 35131 Galaneika - Lamia, Greece. E-mail: mouti@uth.gr

One among the main concerns of language testers in the design and implementation of tests is selecting the method of scoring for the tool used to perform the evaluation. This attribute indirectly reveals the tester's ethical beliefs and personal stance on testing pedagogy. This is another study challenging the typical 1-0 method of scoring in Multiple Choice Tests (MCT) and implements, for experimental purposes, a simple polychotomous partial-credit scoring system on official tests administered for the National Foreign Language Exam System in Greece (NFLES-Gr). The study comes in support of earlier findings on the subject by the same authors in analogous smaller-scale studies. The MCT items chosen were completed by a total of 1,922 subjects in different levels of the NFLES-Gr test for Italian as an L2 in Greece. Results clearly indicate that the tested scoring procedure provides refined insights into students' interlanguage levels, enhances sensitivity in scoring procedures, and may provide significant differences for testees found to be close to the pass/non-pass borderline without jeopardizing test reliability.

Keywords: fairness, ethics, multiple-choice items, scoring procedure, language testing

In language, or any other type of testing situation, it is always better to be on the side of the tester. Almost everyone has experienced that bitter aftertaste that comes with assessment in some form. The adjective *fair* or the noun *fairness* are then used to describe this familiar feeling, and in many cases, the relationship between fairness and ethics of assessment is discussed, as there seems to be a common ground in the minds of the testees. The two notions are receiving substantial attention in recent testing scholarly discourse and are being considered profoundly. As a result, the related bibliography is increasing considerably: a) from a theoretical consideration of the issue (Hamp-Lyons, 1997; Farhady, 1999; McCoubrie, 2004; Kunnan, 2014), b) to suggestions for fairness integration in practice (Xi, 2010; Davis, 2010), or c) to suggestions for testing fairness statistically (Kunnan, 2010). This development is welcomed by both testers and testees alike.

The problem couldn't have been better phrased than it was in the words of Spolsky (1981), who expressed his reservations on 'whether language testers have enough evidence to be sure of the decisions made based on test scores' (as referred to in Bachman, 1990, p. 280). In this statement, Spolsky not only questions the ethics and fairness in testing but also focusses on the adequacy of the evidence collected, and thereupon on the scoring methods implemented as possible sources of the problem. Evidence refers to the quantity and quality of information collected and evaluated (scored) to determine the level and depth of language knowledge of the testee and thereupon reach political decisions or offer advice for further learning, depending on the type of test (e.g. if the test is diagnostic it gives information for the testee to use and improve his/her learning, but if it is an entrance test then decisions are made as to who gets what). There seem to be three (not exhaustively) major

problem areas in language testing:

a) The first relates to the construct *language,* which is defined to be very complex since language ability has not been clearly defined (Farhady, 1999). Language ability has been depicted as growing in a continuum from the L1 towards the L2 with the learner being found at an intermedium stage at any point in this development. The term InterLanguage (IL) was suggested by Selinker in 1972 to describe this process and in agreement, quite recently, Lau, Lau, Hong, and Usop (2011, p. 101) further stated 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.". Accepting that there is an IL stage in the language continuum, one would need to question *why this partial knowledge has not been commonly awarded any credit in language testing?*
b) The next issue lies with the *scoring method* adopted to calculate the results. This issue has been the subject of a great deal of research and discussion in the related literature, a selection of which is presented in more detail in the following chapter.
c) The third problem may be related with other *intervening independent variables* involved, which may affect final test scores, testee attitudes to a test or procedures of test implementation.

Two scoring systems (independent variables) with two values, were examined: a) a traditional dichotomous, which is the standard scoring method, and b) an experimental polychotomous, which is the proposed system under investigation.

There are three hypotheses under investigation: a) H_1 there are differences in the final scores between the two scoring procedures, the *partial credit polychotomous option weighting* scoring method (experimental) and its dichotomous (traditional) scoring counterpart, b) H_0 the experimental scoring method does not affect test reliability, c) H_1 the experimental scoring method provides a more comprehensive outcome as to the interlanguage level of the learner (alternative hypothesis). Clearly the alternative hypotheses in (a) and (c) are aimed for in the study as well as the null hypothesis in (b). Previous empirical studies challenging the 1-0 scoring procedure in MCT have been conducted by the same authors (Mouti, Tsopanoglou, & Ypsilandis, 2012; Tsopanoglou, Ypsilandis & Mouti, 2014; Ypsilandis & Mouti, 2017), while it should be noted that a preliminary small-scale study (Tsopanoglou, Ypsilandis and Mouti, 2014), following the same scoring procedure, supported the alternative hypothesis in (a) and (c) as well as the null hypothesis for (b). These previous studies were preliminary and small scale in addressing the same topic more qualitatively while this study is a large-scale study involving almost 2,000 participants. Furthermore, it approaches such a process in a non-language-specific manner as earlier studies were conducted with English as an L2 language tests while this study involves language tests with Italian as an L2. Lastly, the experimental negative scoring procedure used in these preliminary studies was not included in this research design as it was found to be disadvantageous for the participants.

MC weighted scoring

The introduction of psychometrics in testing has brought about several mathematical equations and statistical tests to transform the linguistic, often nominal, variables measured with ratio score calculations. Naturally the focus was on *scoring systems*. Various simple and more complicated types of assessment and thereon scoring procedures and statistical tests have been suggested to measure the main dependent variable (language in our case). Some of these related to MC scoring are presented in short below.

While the MCT is probably the most common tool widely used to perform the act of language testing, it presents a major problem for fair scoring. Whereas the correct answer is clearly awarded a given prescribed mark, it is also clear that distractors do not equally vary in relation to the correct/expected answer. Weighted scoring is a proposed answer to the problem and two possible scenarios have been discussed: a) item weighting, and b) option weighting. The first rewards different grades for correct answers to each test item (Frary, 1989, p.80), while the second grants different credit for choosing each MC option/answer/distractor. Weights on the option-weighting procedure may be determined by judgments of: a) a panel of experts, or b) a single individual or examinee judgement methods (e.g. examinees evaluate their confidence as to the correctness of the options, known as *confidence testing/weighting*), or even c) an empirical option weighting (Frary,1989) based on item analysis; selected distractors "by a large number of students could be given higher weights than distractors that are selected by a small number of students" (Parkens & Zimmaro (2016, p. 65). The two scholars (Parkens

& Zimmaro, 2016, p.65) suggest that weights could be determined according to an item difficulty index and more specifically according to distractors analysis; while more recently, Hameed (2016) proposed a method to automatically evaluate MCQs considering the importance and complexity of each question and providing a fairer way to discriminating between students with equal total scores. It should be added here that Parkens & Zimmaro's (2016) suggestion may be accepted provided that the most popular option is the one closer to the expected answer and that the testees who select it are the ones who score higher and pass the test.

Several other scholars investigated this issue, such as Claudy (1978 and the references within) who examined test reliability with option weighting scoring. In this procedure every option was awarded different points (including negative points for totally irrelevant selections). Haladyna (1990) also presented several reviews on option weighting with clear advantages for improved reliability and higher precision gains on the lower third of score distribution. More recently, Lau, Lau, Hong, and Usop (2011) tried a *Number Right Elimination Testing* (NRET) scoring method that worked through the elimination of incorrect options in MCTs, and the selection of the correct one. Each correct elimination was offered one point while three points were deducted if the correct option was eliminated (a kind of negative marking). The authors claimed that scores were more reliable, and that the scoring method was able to diagnose misconceptions. In Tsopanoglou, Ypsilandis and Mouti (2014), a polychotomous pattern of correct/very-plausible/plausible/totally irrelevant was identified in 27.5% of 80 items extracted from a well-known TOEFL computer adoptive test. In their article, the authors argued that the selection of a highly relevant/suitable to the correct answer distractor is a knowledge marker that shows partial knowledge and thus a person's higher interlanguage stage from those who selected the totally irrelevant option. Their hypothesis was supported by evidence from statistically significant correlations in that those who scored higher on a test also selected the very relevant/suitable option when they did not find the correct answer, while those who selected the totally irrelevant option were those who scored lower. In this light, it was suggested that the selection of very plausible options needed to be awarded partial credit. Negative marking, which was also tested in the above-mentioned study, proved to be particularly disadvantageous for the student-testee and thus it was recommended to be abandoned in language testing. Following the same reasoning, Sočan (2015) maintained that, applying empirically determined weights to the chosen options of a multiple-choice test should produce more valid test scores by providing more information on examinees' knowledge and further concluded that the relative performance of scoring methods strongly depends on the instructions provided and on distractors' properties, and only to a lesser extent on sample size and test length. In favour of polychotomous scoring is also the study of Diederhofen & Musch (2017) who found that in comparison to dichotomous scoring, empirical option weighting employing polychotomous scoring increases both the reliability and validity of multiple-choice knowledge tests. In addition, the authors suggest the computation of empirical option weights for existing multiple-choice knowledge tests that have previously been scored dichotomously as a potential application of their proposal. Yoo et al (2018), addressed validity and fairness by exploring other independent variables of test-taker characteristics such as gender, age, educational background, language exposure, and previous experience with assessment. The authors explored and provided evidence in support of the use of score equity assessment to promote the validity and fairness of reported comparable scores for various groups of test takers across the various subgroups examined.

The solution of offering different weights to distractors in MC testing is not totally new. It has been discussed in related literature and several scoring procedures have been offered to tackle the problem. Weighting suggestions could be divided into two major categories: a) those which require a very basic knowledge of arithmetic and no statistics, frequently used by language teachers and b) those involving a considerable knowledge of statistics. A list of five scoring procedures examined by Claudy (1978) under these two broad categories are offered below very briefly.

a) Commonly used scoring procedures with no option weighting.

1. The most known dichotomous scoring system is the one in which the correct option receives one (1) point while all the other options receive no (0) credit. Claudy (1978) refers to this as *number right scoring* system, while it is commonly called the 1-0 scoring procedure.
2. Another widely spread (at least for a period) trichotomous scoring practice is Claudy's (1978) *correction for guessing scoring* (also known as negative scoring), in which the correct answer receives one (1) point while all the other options receive minus one (-1) point. This scoring procedure was suggested and implemented to address lucky selections by the testees. The term "guessing" that was used, revealed the

belief of many language teachers that learners select options by chance when they do not know the correct answer. Consequently, this attitude was unwelcome and therefore incorrect guesses were to be punished with a negative mark. The NRET by Lau, Lau, Hong, and Usop, (2011) was also in this direction. Sočan (2009) refers to several articles addressing this issue with an equal number of solutions. Tsopanoglou, Ypsilandis and Mouti (2014) have argued for a different understanding over the guessing or inferencing problem. They claimed that MC inferencing in option selection tests is activated automatically in the human brain when the individual does not possess the correct answer and it is made by actuating relevant knowledge of the target language or a general knowledge of languages. By that respect the authors claimed that this brain procedure cannot be totally blind (unless the testee does not read at all the stem and the options) and by that respect a very plausible selection by inferencing would also need to be rewarded 'through partial credit scoring' (Bachman and Palmer, 1996, p. 205).

b) Option weighting practices

Claudy (1978, pp. 26-27), presents the following three weighted scoring methods:

1. Guttman Weights scoring, "in which the weight assigned to each option, including 'omit,' the z-score corresponding to the mean score on the other test items for examinees who select the option. Theoretically, the weights can take any value a z-score can assume".
2. Biserial Weights scoring, "in which the weight assigned to each option, including 'omit,' is the Brogden Biserial (Brogden, 1944) or Clemans Lambda (Clemans, 1958) correlation coefficient between selecting the option (scored 1 or 0) and the score on the other items in the test. In effect, this is a discrimination index for the option. Theoretically, the weights can take any value between minus one and plus one inclusive".
3. Proportion Weights scoring, "in which the weight assigned to each option including 'omit,' is the proportion of examinees in an upper score subgroup of examinees who select the option. Theoretically, the weights can take any value between zero and one inclusive, and sum to one for each item. Two overlapping subgroups were used in this study: the upper 25% and the upper 50% of examinees on the basis of total score".

These option-weighting practices have not been widely implemented as they are complex and thus costly and non-time-effective while they require a computer and a solid knowledge of statistics from the part of the teacher-tester and the testee alike. In this respect these marking systems violate a significant recommendation for ethical testing implementation, i.e. for the testees to be aware of how the marking is implemented.

Sočan (2009, p. 79), also presents several different weighting options, which can be based on the following criteria: 1. subjective judgment of the test constructor on the degree of "falseness" of each alternative, 2. correlation with an external criterion, also by other 'judges', or 3. relation to an internal criterion. Other scoring methods to credit partial knowledge proposed by Lau, Lau, Hong, and Usop (2011) are the following:

> "confidence weighting (CW), elimination testing (ET), subset selection testing (SST), probability measurement (PM), answer-until-correct (AUC), option weighting, item weighting, rank ordering the option, and partial ordering. All these scoring methods aim to extract information from the examinees that can provide better estimates of their abilities."

Weight in some of the methods is decided before test implementation while in other cases it is decided after a test is implemented. One more after-test-implementation approach is Haladyna's (2004, p.224) *choice mean* described as a "differential distractor functioning". Haladyna (2004) proposes that:

> "for any option, we can calculate the mean of all examinees who chose that option. For the right answer, this mean will typically be higher than the means of any wrong answer. We can analyze the relationship of the choice mean to total score or the choice of the option to total score".

Based on the above analysis, Haladyna (2004) recommends two different statistical strata: a) the product-moment correlation between the choice mean and total score, where the choice mean is substituted for the distractor choice, and b) the eta coefficient; the independent variable is option choice and the dependent

variable is total score.

Following this line of argumentation, the present paper supports that the correct answers and various distractors could be differentially weighted according to their approximate correctness and further, implement a polychotomous *weighting* scoring method without the use of negative marking for totally irrelevant selections. In agreement with Haladyna (2004, p.253), it is believed that "the use of distractor information for test scoring" may "increase the reliability of test scores, which in turn should lead to more accurate decisions in high-stakes pass-fail testing". Arguably, this decision would need to be made before test implementation and further be made known to the testees to increase ethics and fairness in testing. The problem here then becomes finding the right distractors to deliver the task, a process that is expected to be demanding particularly before a test is implemented and an item analysis could be employed.

In order to tackle this problem, Hsu et al. (2018) proposed an estimation model for item difficulty as an alternative for pretesting procedures. Their findings suggest that the semantic similarity between a stem and the options have the strongest impact on item difficulty, which they claim is the most challenging part of the multiple-choice item-writing process, i.e. to develop plausible distractors. Susanti et al. (2018) also offered their proposal to the problem through a novel method to automatically generate distractors for multiple-choice English vocabulary questions. Their suggestion unfolded in two phases: a) by utilizing initially semantic similarity and collocation information, and b) by consulting an expert to judge the quality of the distractors generated (which is the traditional method). The results in their study showed that their proposed method ensured the validity of the items and produced fewer problematic distractors than the baseline method, being also comparable with that of human-made distractors. Their findings coincided to a similar earlier study, where Susanti et al. (2015) attempted to construct English vocabulary automatic tests in which human evaluation indicated that 45% of the responses from English teachers mistakenly judged the automatically generated questions to be human-generated questions. Finally, Ha and Yaneva (2018) proposed their fully automatic method for generating distractor suggestions for multiple-choice questions used in high-stakes medical exams by using a question stem and the correct answer as an input. Last, El Masri et al (2016, p. 62) distinguished item difficulty and item demands by adopting Pollitt et al.'s (2007) measurement of phenomena for both terms, and stated, "Item difficulty is the empirical location of an item on a scale. Item demands designate the set of knowledge, comprehension, skills and processes required for a student to respond fully or partially correctly to an item (Ferrara & Duncan, 2011; Ferrara et al., 2011)". Both Ferrara et al. (2011) and Pollitt et al. (2007) agree that the relationship between the notions of *demands* and *difficulty* is not as straightforward since more demanding questions present by definition greater difficulty, while the reverse relationship is not necessarily applicable (Pollitt et al., 2007) as cited by El Masri et al (2016, p. 62).

Ethics in language testing

Method of scoring and test impact are two possible areas where the ethical intentions and predisposition of the tester or the test can be revealed more clearly, in addition to his/her beliefs and knowledge on language pedagogy. Thus, the term 'ethics' may be considered to describe a stance or rather the testers' predisposition towards the notion of assessment while attempts are made for ethics to be applied in practice. This is (or should be) the concern of testers and test-takers alike and it is the responsibility of both. While the former is being discussed, the latter has not received equal attention. More than 30 years ago the Educational Testing Service (ETS) started to review its materials to ensure fairness for all participants and thereupon issued a first report of fairness guidelines. However, it was Spolsky in 1981 (as referred in Bachman 1990 & Hughes 1989) who was the only language testing researcher up to that point to address the ethical considerations of testing impact by questioning 'whether language testers have enough evidence to be sure of the decisions made on the basis of test scores' (Bachman 1990, p. 280). This ethical problem becomes crucial in situations where language proficiency tests such as TOEFL or GMAT are used as a mechanism 'for making college admission decisions' (Bachman 1990, p. 282). Since then, 'ethical issues' in language testing practices have been raised (Canale, 1988, p. 77) and the topic has received considerable attention among test developers and test users, while researchers have begun to focus directly and indirectly on this subject. Further, awareness has been raised as to the responsibility of the individual language testing practitioner (Hamp Lyons, 2000) for the professional standards by which language testing as professional practice are to be held and '… the responsibility of test users to ensure that language tests are valuable experiences and yield positive consequences for all involved' (Douglas & Chapelle, 1993, p. 4). The issue of responsibility was also discussed by Bachman & Palmer (2010, p. 191) who argued that 'identifying

who is responsible for making sure that the consequences are beneficial is surely a complicated issue without easy, all-purpose answers'.

Hamp-Lyons (1997) raised ethical questions about washback impact (washback is a term used to describe the effect on teaching by the application of a test) and validity, while Shohamy (1998) examined possible benefits from language testing to second language acquisition and vice versa. In 1997, the *Language Testing* special issue (Vol. 14 Num. 3), was dedicated to ethics in language testing and contained a collection of papers discussing ethical matters relevant to the field, while Kunnan (1999) focused on the most significant theoretical developments in the field of language testing in the 1990s and paid considerable attention to the role of ethics among testers as well as the role of fairness in test validation. Continuing the discussion on ethics and validity, Davies (2003, p. 361) specified, 'If a valid test is by definition ethical, that frees up the tester to concentrate on validity. If, however, ethics is a separate add-on, then the demands on the tester may be too great.' Despite the fact that the discussion on ethics continues, the issue has not been defined nor has it been very clear as to how this applies to language testing.

Fairness in language testing. Ethics is a term used to describe the field of moral philosophy which 'involves systematizing, defending, and recommending concepts of right and wrong behaviour' (Internet Encyclopedia of Philosophy). Three areas of study are registered metaethics, normative ethics, and applied ethics. Applied ethics is a fairly recent and quite straightforward development as it attempts to apply ethical theory of specific moral issues to different situations, e.g. business ethics. It is thus possible to add ethics in testing as one possible branch of applied ethics that may contribute to the analysis and discussion of certain moral issues in the field. In modern ethics one may also find the term consequential ethics in which the end (the outcome or the consequence) of an action justifies the means (moral right or wrong) to achieve it. If consequences are good then the action is morally ethical. The exact opposite is deontological ethics, which argues that it is the approach to ethics that determines the goodness or rightness of acts or intentions, e.g. if a person seeks to do what is right (has good intentions) then it is the act that determines it and not its consequences. Ideally, both are required in testing practice since the good will of test developers needs to be witnessed in the consequences to the test-takers, test developers, and testing philosophy as well. The above discussion also implies the existence of ethical and non-ethical acts which may relate as follows: when the consequences are dear, good intentions are considered ethical (although this may have little to offer for consolidation to those who have to suffer from the effects - impact), but when the intentions are bad, any good consequences resulting from that may still be considered unethical (and coincidental).

Metaethics involves the study of ethics from a theoretical point of view, that is, creating awareness on the topic similar to other areas of study where the prefix meta- is used. In particular, the focus is on the origin of ethics (how this is understood), the meaning of ethical concepts such as right and wrong, e.g. discussing what could be considered right and wrong behaviour in this specific field of study and how this would affect the notion of assessment in general.

Normative ethics may be seen as the ultimate target in ethics and it involves finding moral standards that could determine a moral 'course of action' with practical value, in our case for language testing. These moral principles of right and wrong practice (actions) in language testing could be set and used as a guide for moral decisions in the field; a code of ethics for testing. The CoE argued for '… ethical behaviour by all language testers' (ILTA 2000, CoE p. 1). The Code of Practice which came along later identified 'the minimum requirements for practice in the profession…' and focused '… on the clarification of professional misconduct and unprofessional conduct' (ibid.). Following the above discussion, one may conclude that ethics in language testing is more of an abstract notion and should not be treated as a separate value per se but rather as a construct value composed of certain qualities which may find application in different stages of assessment, at a preparation stage more deontological and at the impact stage more consequential.

Neither fairness nor ethics can be applied to a test by a simple yes/no answer. It is not an absolute situation and it could be better understood in the form of a value on a scale that increases or decreases depending on certain qualities and is different in each assessment stage from test preparation and administration to test scoring method and impact. It is clear, however, that despite the considerable attention on fairness, as Zieky (2002, p.2) argued, not much has been accomplished as 'there is no statistic that you can use to prove that the items in a test are fair, and there is no statistic that you can use to prove that the test as a whole is fair. The best way to

ensure test fairness is to build fairness into the development, administration, and scoring processes'.

In this light, Figure 1 attempts to depict schematically the relation between the two (ethics and fairness) and possible areas of operation, in three major stages of assessment. Clearly, the qualities presented in the diagram are merely suggestive and not exhaustive, but it is our view that these could be used to continue the discussion on this topic. In particular: (a) Test preparation stage. During planning, it may be possible to discuss matters related to the audience, purpose, and rational (deontological ethics) of a test, while during development, the focus should be on the selection of an appropriate instrument, the content, the item type, and the scoring method. Bachman (2010, p. 2) suggests the term 'assessment use argument' which can 'guide the design and development of assessments and can also lead to a focused, efficient program for collecting the most critical evidence in support of the interpretations and uses for which the assessment is intended.' It should be noted here that it is the purpose of a test to create the most appropriate environment in which test-takers could bring out the best of their linguistic abilities. (b) Test administration stage. Here, there are topics that should be discussed *before the test*: access of test-takers to the test location; *during the test*: the class environment, the test process, and vigilance; and finally *after the test*: correction method and method of presenting results or even advice for further preparation for those who failed the test. (c) Test impact stage. This is the final stage in which discussions are related to the decisions taken at a macro or a micro level (terms suggested by Bachman & Palmer, 1996) upon the results of a test (consequential ethics). Needless to say, the test-taker is to be made aware of the above.

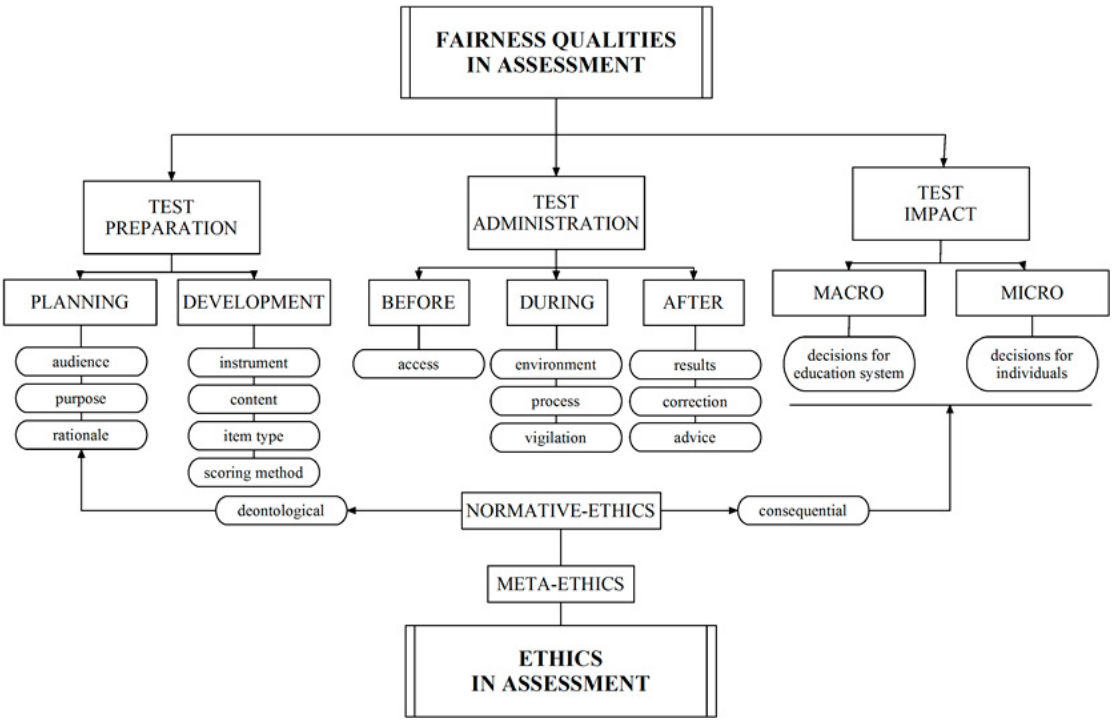

Figure 1. Fairness typology in relation to ethics in assessment.

Materials and Methods

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 are 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.

Participants

Two groups of participants were engaged. The first group consisted of four native speakers/teachers and two proficient and experienced teachers of Italian (judges, from here on) who classified the options according to the four stances Likert scale (correct/very-plausible/plausible/totally irrelevant). 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 the A1-A2, 1,294 at the B1-B2, and 228 at the C1 levels). The L1 of the test-takers was Greek.

Design, Procedure, and Scoring System

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 four 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 three possible answers (one correct and two 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.

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 a very plausible/plausible alternative and zero (0) points for the selection of a totally irrelevant/wrong answer. An example of the polychotomous pattern examined and scored is as follows:

> *Oltre mille persone sono arrivate da tutta la Sicilia (e non solo) per_____ (A. provare B. tentare C. cercare) a entrare nella casa piu famosa della TV.*

In this example the first alternative was presented by the Scoring Key as the correct answer while B was identified by the judges as a very plausible option.

Data Analysis - Scoring Procedures

The longer completed tests were administered in official settings and completed on paper by the participants in an authentic environment. The three Protocol Tests (PTs, the study sample), resulting from extracted items, consisted of 53 language awareness items in a multiple choice format. The analysis of the data was deployed in two basic stages. During the first stage, an initial investigation was conducted to ensure that the sample contained items could be evaluated according to the aforementioned polychotomous Likert pattern. In the 53 items that were examined and divided into five 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. Table 1 shows the exact figures of dichotomous and polychotomous items at each level.

Table 1
Dichotomous and Polychotomous Patterns Investigated

	Dichotomous Pattern	Polychotomous Pattern
A/ 10 items	9	1
B1/7 items	7	0
B2/9 items	4	5
C1/12 items	6	6
C1/15 items	10	5

Table 2
Distractor Analysis of Polychotomous Items

Item	Correct Answer	Very Plausible answer	Option/highest %	Judgments & Item Analysis
A/1	B (0.47)	A (0.48)	***A***	+
B2/2	A (0.59)	B (0.25)	A	+
B2/4	A (0.17)	B (0.31)	C (0.52)	?
B2/6	A (0.33)	B (0.57)	***B***	+
B2/7	A (0.17)	C (0.67)	***C***	+
B2/8	B (0.4)	A (0.43)	***A***	+
C1.1/3	C (0.5)	A (0.44)	C	+
C1.1/4	C (0.21)	B (0.55)	***B***	+
C1.1/7	A (0.54)	B (0.08)	A	+
C1.1/8	B (0.97)	B (0.01)	B	+
C1.1/11	C (0.18)	B (0.45)	***B***	+
C1.1/12	B (0.71)	A (0.26)	B	+
C1.2/5	B (0.39)	C (0.59)	***C***	+
C1.2/9	B (0.31)	A (0.45)	***A***	+
C1.2/10	A (0.36)	B (0.61)	***B***	+
C1.2/12	C (0.38)	B (0.39)	***B***	+
C1.2/13	A (0.39)	C (0.43)	***C***	+
Total 53 items		**36 (68%)**		**17 (32%)**

Decisions as to the polychotomous pattern of these 17 items were based on the conscious judgment and unconscious judgment of the judges. For 11 items (25%), the polychotomous pattern was confirmed by the judges while for six items not all the expert judges were able to trace the correct answer, being distracted themselves, and therefore these items were also examined and included in this polychotomous category.

These judgments were examined and verified empirically in relation to an *item* and *distractor analysis*. For 11 out of the 17 polychotomous items (marked with bold and italics in the following table), the *very plausible* answer/option was the one with a slightly higher choice mean/percentage compared to the *correct answer* (as presented in the second and third columns). In only five cases, the mean of the *correct answer* was higher than the mean of the *very plausible* option. Note also that in one (1) case (line 3 of Table 2), indicated with a question mark in the fifth column, the results showed that the highest mean was found in a third option. This demonstrates that the distractors in most cases performed their job and that the correct and the very plausible options were indeed very close. This indicates that the subjects who selected it were on route to the learning of the phenomenon tested and clearly at a higher level than those who selected a totally irrelevant option.

At a second stage the polychotomous pattern items were scored with both the DSM and the ESM scoring procedures, while the ones where no polychotomous pattern was identified by the experts were only scored in the traditional DSM. Results from the different scoring procedures are initially presented as mean, standard deviation and alpha level (α), which is the probability of error, rejecting the null hypothesis when it is true. Means are then statistically compared through the SPSS statistical package (paired samples t-test) in relation to the first hypothesis [H(a)] stated at the introduction to investigate whether the differences are statistically significant. Finally, the relationship between the two scoring methods (ESM and DSM) was examined at each level separately through the correlation coefficient Pearson r (second hypothesis) to offer insights regarding possible associations between the scoring procedures (testing the reliability of the ESM) [H(b)]. The third hypothesis [H(c)] was examined by the variation of the observed differences (random or linear). The three hypotheses were tested in three levels of language proficiency, namely A, B2, and C1.

Results at A and B levels were not examined with the ESM as items were single-correct answer in agreement with the judges, who did not trace polychotomous patterns in them. Therefore, they are not presented. It may

be argued, however, that at lower levels degrees of incorrectness and polychotomous patterns cannot be easily applied as this would increase test difficulty significantly.

B Level: 2nd set of items-9 MC questions

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

Experts recognized a polychtomous pattern in five 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> Polychotmous FI=0.32). The statistical analysis revealed that in four items the plausible distractor was chosen by more test-takers instead of the correct answer. In three of those items the selection coincided with the one provided by the expert judges as correct. This was the case in which the selection distracted the judges as well.

The scores were altered when the ESM was implemented. In particular, the *facility indexes* were increased and the differences were statistically significant: Mean DSM=3.38, Mean ESM=4.5. Further, a paired samples t-test showed statistically significant differences between the two scores t=72.941, df=1293 (.000) which supported the alternative first hypothesis [H(a)]. In order to investigate reliability of the ESM [H(b)] the Pearson r correlation coefficient was employed (examines the relationship among variables) to compare the independent variables in pairs. 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 the DSM and the ESM is r=.937 (.000) and correlation is significant at the 0.01 level (2-tailed). Thus, test results are not jeopardized [H(b)]. Observable differences were not explained by random variation and thus it may be suggested that the ESM offers more sensitive scoring, statistically different from the DSM [H(c)].

C1 Level: 1st set of items-12 MC questions

(Mean=7.36, SD=1.84, Alpha=0.41)

A polychotomous pattern for six items was identified by the expert judges at this level. 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 three items the distractor was chosen by most test-takers instead of the expected/ correct answer. In two of these three cases the subjects' erroneous selection again coincided with the one selected by the experts' judgments. Apparently, the distractors were too good; good enough to mislead the native judges as well. Implementing the ESM exhibited alterations in the scores, the Facility Indexes were increased: Mean DSM=7.36, Mean ESM=8.20. A paired sample t-test [H(a)] showed that these differences were statistically significant: t=26.215, df=228 (.000). A Pearson r correlation analysis that followed [H(b)], between the DSM and the ESM, was at r= 0.966 (.000) with a 0.01 level of significance (two-tailed), which supports that reliability was not jeopardized. This confirmed again that the two scoring procedures indeed exist in a strong linear relationship to each other. Once more, observable differences were not explained by random variation and thus it is confirmed at this level as well that the ESM offers a more sensitive scoring statistically different from the DSM [H(c)].

C1 Level: 2nd set of items-15 MC questions

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

Experts recognized a polychtomous pattern in five items, which again proved to be more difficult to complete than the dichotomous items, as indicated by the *facility index* (Dichotomous FI =0.73> Polychotmous FI=0.40). The statistical analysis revealed that for all five items, the plausible distractor was chosen by most test-takers than the correct answer and once again their selection coincided with the experts' judgments. Implementing the ESM once again altered the scores, the *facility indexes* were increased and the differences were statistically significant: Mean DSM=8.22, Mean ESM=9.46. Similarly to the B level results above, a paired samples t-test showed statistically significant differences: t-test: t=29.509, df=228 (.000) which again supported the alternative hypothesis. Once again the two scoring procedures proved to exist in a strong linear relationship to each other,

as the Pearson r value between the DSM and the ESM was r=0.967 (000) and correlation was significant at the 0.01 level (two-tailed). Again, observable differences were not explained by random variation in support of the third hypothesis.

Discussion

The three initial hypotheses under investigation have been adequately supported by the evidence for the three test levels used for the study. More specifically, it was shown that: a) there was a statistically significant difference between the two tested scoring procedures [H(a)] with the ESM presenting significantly higher results, b) the two final scores proved to exist in a strong linear relationship to each other, which demonstrates that test reliability was not affected [H(b)], similar to the findings in Claudy (1978 and the citations within) and Haladyna (1990), and c) any observable differences in the scores were not explained by a random variation, which can be interpreted as an indication that the polychotomous scoring method tested provides more refined information, a more comprehensive outcome, and thus a more sensitive final scoring to mirror the interlanguage stage of the learner [H(c)], similarly to claims in Haladyna (1990) concerning weighted scoring. These findings from an authentic language test in an authentic testing setup concur with earlier results found in previous small-scale experimental studies by Mouti, Tsopanoglou and Ypsilandis (2012), and Tsopanoglou, Ypsilandis, & Mouti (2014).

An additional finding is that dichotomous items proved to be easier to answer as the correct option stands out and becomes transparent. On the other hand, polychotomous items can be traced to a greater degree at the higher B2-C1 levels and not at the lower A1-B1 (the higher the level the more polychotomous items identified). These findings come in agreement to earlier claims by Bachman and Palmer (1996: 202) who 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”. In support, Andrich & Marais (2018), declared that the more difficult the item, the greater the degree of guessing and that persons with greater proficiency tended to correctly answer the more difficult items at a greater rate than the less proficient.

Subsequently, it is possible to claim that polychotomous items form a paradox. On the one hand they cannot be avoided in language testing, particularly at the higher levels of language testing as they are more demanding, while on the other they increase the mean in the final score when a polychotomous scoring method is adopted, despite the registered increase in difficulty. The evident and subsequent increase of score levels from the adoption of the ESM would not affect norm-referenced testing situations as the test-takers’ ranking remains the same (without affecting test reliability), although 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 precisely here that polychotomous scoring would have a significant impact and would need to be implemented. It is thus possible to argue that the ESM adopted in this study may indeed provide a more complete view of the interlanguage stage of an individual by offering a more detailed, in-depth, and precise analysis of results as well as enhance sensitivity and thus contribute 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 successful 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”. Subsequently, it is possible to argue that hypothesis [H(c)] has also been supported by the evidence. Although Zieky (2002, p. 11) finds that “there is no magic bullet to guarantee fairness” this ESM increased sensitivity and precision in test scoring by providing more refined insights of student interlanguage levels and therefore becomes more reflective of student knowledge. It may be possible to conclude that if polychotomous items are used in a test, polychotomous scoring would increase the quality of results and *fairness* in scoring would be served. The construct of *ethics* in language testing may also be increased by the increased sensitivity of the scoring method and by notifying test-takers about it. Thereon, testees would be expected to become more engaged in answering a MC item, even when the expected answer is not known to them at a first glance.

Limitations in this work lay predominantly with the levels examined to explore the hypotheses and thus more research of the same type would need to follow in this direction in authentic testing situations. Also, a second

control-scoring-method to examine student interlanguage level would support claims of sensitivity of the ESM tested here. Other hypotheses that would need to be investigated further may compare testee attitudes, engagement, responsibilities, and performance in answering polychotomous items when the testees know that polychotomous scoring has been implemented with their attitudes and performance when this knowledge is lacking.

Conclusion

This study developed naturally as a comprehensive test of experimental findings from earlier studies of the same authors, mentioned above, to address ethics and fairness in selected response types of tests, which were explored initially in an experimental setup. It is possible to conclude, by the exploration of the three hypotheses in a combined interpretative fashion, that selection items which use polychotomous patterns would also need to implement a polychotomous scoring method, such as the one suggested in this study.

The ESM method of weighted scoring in the suggested polychotomous fashion tested here performed equally well with the same statistical results at all three language levels examined in the experimental and authentic studies, which clearly demonstrates that it can be safely used with all selected response items irrespective of level, and thus serve the notions of ethics and fairness in language testing, particularly when political decisions are reached from test results. The suggested option-weighted scoring (1-0.5-0) can be easily implemented in language tests as it is easy and straightforward to apply, it is not time consuming and does not need the involvement of a computer, nor does it require high expertise in mathematics or item analysis, even in the case of manual correction.

Although this option weighting does not settle ethics and fairness in language testing by itself, it provides clear advantages on improved reliability and higher precision gains to other scoring methods; it is more testee-friendly and the findings in this study, from a large authentic sample, coincide to the above claims and add to the issue.

References

Andrich, D., & Marais, I. (2018). Controlling Bias in Both Constructed Response and Multiple-Choice Items When Analyzed With the Dichotomous Rasch Model. *Journal of Educational Measurement*, *55*(2), 281-307. doi:10.1111/jedm.12176

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

Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge, UK Cambridge University Press.

Bachman, L. F., & Palmer S. A. (2010). *Language assessment in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Palmer, S. A. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Bachman, L. F., & Purpura, J. E. (2008). Language assessments: Gate-keepers or gate- openers? In B. Spolsky & F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 456-468). Oxford, UK: Blackwell Publishing.

Boyd, K., & Davies, A. (2002). Doctors' order for language testers: the origin and purpose of ethical codes. *Language Testing*, 19(3), 296-322.

Claudy, J. G. (1978). Biserial Weights: A New Approach to Test Item Option Weighting. *Applied Psychological Measurement,* 2(1), 25–30. doi: 10.1177/014662167800200102

Davies, A. (2010). Test fairness: a response. *Language Testing*, 27(2), 171-176. doi: 10.1177/0265532209349466

Davies, A. (2003). Three heresies of language testing research. *Language Testing,* 20(4), 355-368.

Diedenhofen, B., & Musch, J. (2017). Empirical option weights improve the validity of a multiple-choice knowledge test. *European Journal of Psychological Assessment, 33*(5), 336-344.

El Masri, Y. H., Ferrara, S., Foltz, P. W., & Baird, J. (2016). Predicting item difficulty of science national curriculum tests: the case of key stage 2 assessments. *The Curriculum Journal*, *28*(1), 59-82. doi:10.1080/09585176.2016.1232201

Farhady H. (1996). Varieties of cloze procedure in EFL Education. *Roshd Foreign Language Teaching Journal, 12*, 217-229.

Farhady, H. (1999). Ethics in language testing. *Moddaress Journal*, *3*(11), pp. 447-464.
Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*, 79-96.
Ha, L. A., & Yaneva, V. (2018). Automatic Distractor Suggestion for Multiple-Choice Tests Using Concept Embeddings and Information Retrieval. *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. doi:10.18653/v1/w18-054
Haladyna, T. M. (2004). Developing and validating multiple-choice test items. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
Hameed, I. A. (2016). A Fuzzy System to Automatically Evaluate and Improve Fairness of Multiple-Choice Questions (MCQs) based Exams. *Proceedings of the 8th International Conference on Computer Supported Education*. doi:10.5220/0005897204760481
Hamp-Lyons, L. (1997). Ethics in Language Testing. In C. Clapham and D. Corson (eds.), *Encyclopedia of Language and Education. (Volume 7, Language Testing and Assessment),* 323–333. Dordrecht: Kluwer Academic.
Hsu, F., Lee, H., Chang, T., & Sung, Y. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. Information Processing & Management, 54(6), 969-984. doi:10.1016/j.ipm.2018.06.007
Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European Language Testing in a Global Context* (pp. 27-48). Cambridge, UK: Cambridge University Press.
Kunnan, A. J. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.) *The Companion to Language Assessment* (pp. 1098-1114). Boston, MA: Wiley-Blackwell.
Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, 27, 183–189.
Lau, P. N. K., Lau, S. H., Hong, K. S., & Usop, H. (2011). Guessing, partial knowledge, and misconceptions in multiple-choice tests. *Educational Technology and Society, 14*(4), 99-110.
McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: a literature review. *Medical Teacher*, *26*(8), 709-712.
Mouti A., Ypsilandis G., & Tsopanoglou A. (2012). Investigating fairness in multiple choice tests. In Z. Gavriilidou, A. Efthymiou, E. Thomadaki & P. Kambakis-Vougiouklis (eds), *,Selected papers of the 10th ICGL*, pp. 965-972. Komotini/Greece: Democritus University of Thrace.
Parkes, J., & Zimmaro, D. (2016). *Learning and assessing with multiple-choice questions in college classrooms.* New York, NY: Routledge.
Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*, 209–231.
Sočan, G. (2009). Scoring of multiple choice items by means of internal linear weighting. *Review of Psychology*, *16*(2), 77-85.
Sočan, G. (2015). Empirical option weights for multiple-choice items: Interactions with item properties and testing design. Metodološki zvezki, 12(1), 25-43.
Susanti, Y., Iida, R., Tokunaga, T. (2015). Automatic generation of English vocabulary tests. In *Proceedings of the 7th International Conference on Computer Supported Education*. INSTICC, Setubal, (pp. 77–87).
Susanti, Y., Tokunaga, T., Nishikawa, H., & Obari, H. (2018). Automatic distractor generation for multiple-choice English vocabulary questions. Research and Practice in Technology Enhanced Learning, 13(1). doi:10.1186/s41039-018-0082
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2), 147-170.
Yoo, H. J., Manna, V. F., Monfils, L. F., & Oh, H. (2018). Measuring English language proficiency across subgroups: Using score equity assessment to evaluate test fairness. Language Testing, 026553221877604. doi:10.1177/0265532218776040
Ypsilandis S.G. & Mouti A. (2017). 'Investigating Scoring Procedures in Language Testing'. In *the Proceedings of the 6th Association of Language Testers in Europe (ALTE) conference on Learning and Assessment: Making the Connections*, May 3-7, 2017 in Bologna, Italy, pp. 154-159.
Zieky, M. (2002). Ensuring the fairness of licensing tests, educational testing service CLEAR exam review, *12*(1), 20-26.
Tsopanoglou A., Ypsilandis G., & Mouti A. (2014). Piloting a polychotomous partial-credit scoring procedure in MC tests. *Journal of the European Confederation of Language Centres in Higher Education, 4*(1). Retrieved from http://www.degruyter.com/view/j/cercles.2014.4.issue1/cercles-2014-0004/cercles-2014-0004.xml