"Quaderni di Ricerca in Didattica (Mathematics)", n. 21, 2011
G.R.I.M. (Department of Mathematics, University of Palermo, Italy)

Multiple-choice math tests: should we worry about guessing?

Chiara Andrà, and Guido Magnano
Dipartimento di Matematica, Università di Torino
Via Carlo Alberto, 10
10123 Torino (ITALY)

chiara.andra@gmail.com, guido.magnano@unito.it

Abstract. With reference to the recent diffusion of multiple-choice testing in Italy, we argue that – in the absence of an appropriate methodological background – the ubiquitous adoption of "formula scoring" may reveal an incorrect view of the functioning of multiple-choice items. We point out that common motivations for assigning a penalty to incorrect answers are based on a misconception of the effects on the answering strategy of the examinees. To support our viewpoint against indiscriminate use of formula scoring, we report the case study of math items used in entrance tests for undergraduate curricula at the University of Turin (Italy), where comparative IRT analysis shows no evidence of systematic guessing under right-only scoring.

MSC: 97C06

1. Introduction and background

In Italy, multiple-choice have been increasingly adopted after 2001 as entrance tests for undergraduate curricula and in the admissions to qualification (postgraduate) courses for perspective high-school teachers. Yet, such tests are still perceived as foreign to the Italian educational system, which traditionally relies on oral examinations and non-standardized written tests.

At present, most Italian education professionals hold for granted that in multiple-choice tests one should assign a negative score for incorrect answers, according to the well-known formula-scoring rule. The use of the formula scoring, as opposed to the right-only scoring, has been the subject of a lively debate in countries with a long-established practice of multiple-choice tests since in the first half of the XX Century (Thurstone, 1919; Holzinger, 1924). In Italy, this debate seems to be ignored.

Multiple-choice tests have specific features that distinguish them from open-ended ones. One of the most evident differences is that only the result of the student's solving process is observable. In our experience, we have often observed that math teachers, used to "traditional" types of examinations, tend to regard multiple-choice items as substitutes for open-ended exercises, and therefore to assume that a single multiple-choice question should provide a reliable evidence on the ability of the student to solve a specific problem. In view of that, they expect that each item should be answered by reading the question, identifying and performing the appropriate computations, and only afterwards comparing the found solution with the proposed answers (possibly retaking the previous steps, if none of the answers corresponds).

These assumptions entail almost naturally a deep concern about the possibility that a student may select the correct answer – among the k available choices – as a result of blind guessing (with probability 1/k, which is

usually far from being negligible). The better solution to get rid of guessing appears to be the use of formula scoring.

The simplest scoring rule – often referred to as number-right or right-only scoring – merely counts the number r of correct answers. This scoring rule is nowadays used, for example, by the American College Testing (ACT) and by the Graduate Record Examinations (GRE) general exams. Although under this rule it is never better to omit the response than to choose an answer at random, it has been repeatedly observed that some examinees do not answer all items. For example, in 1984, shortly after the introduction of number right scoring rule, only 44% of GRE examinees answered all questions and 5% of them (about 3000 examinees) omitted 20 items or more (Grandy, 1987).

The most known formula-scoring rule consists instead in assigning one full point for each right answer, zero points for each omitted answer and -1/(k–1) points for each wrong answer, so that an examinee answering in a totally random way to all questions would have an expected score of zero points.

An alternative rule (Traub & Hambleton, 1972) achieves a similar result by adding for each omission 1/k points and being neutral regarding incorrect responses. Although in the latter version the idea of "penalty" is not explicitly suggested, it is easy to see that the outcomes of the two methods are linearly related.

The typical motivations for the choice of a formula scoring rule tend to be either "technical" or "moral". In the first case, it is believed that formula scoring, just because it produces a zero average score in the case of pure blind guessing, has the effect of suppressing the distortion due to the guessing practice. According to the "moral" viewpoint, on the other hand, the practice of guessing is unfair, because a guesser would obtain a higher score with respect to an equally skilled, but more "honest", candidate (this consideration typically arises whenever the test is used in a competition or to produce an admission list: in Italy, such tests invariably adopt formula scoring). The formula scoring is believed, in this case, to effectively discourage an "unfair" practice through the introduction of an appropriate "penalty".

A large literature focused on development, evaluation, and comparison of different scoring rules, and on the effects of each scoring method on the answering strategy of the examinees. Diamond and Evans (1973) assumed that examinees either know the answer to a test item or else choose among all alternative responses at random. Lord (1975) suggested that examinees may be in an intermediate position, being able to rule out one or more of the alternatives with a certain level of assurance. From a psychometrical point of view, Patnaik and Traub (1973) found evidence of higher internal consistency for formula-weighted scores. Hambleton, Roberts, and Traub (1970), instead, found no significant increase in internal consistency for formula-obtained scores.

Kansup and Hakstian (1975; Hakstian & Kansup, 1975) observed that formula-scoring methods require special training for examinees and considerably more testing time. Their results, jointly with earlier results (Collet, 1971; Coombs, Milholland, & Womer, 1956; Dressel & Schmid, 1953), had shown that a significant increase of reliability and validity is not met when formula-scoring procedures are compared to number-right

ones, during tests requiring the same time. The stability coefficients were not increased (and, in fact, were lower) for the formula-weighted scores, and they obtained a significant decrease in validity with the Mathematical Reasoning test. They also measured what they called "an additional trait", largely unrelated to the ones of interest. A further investigation of personality traits affecting the reaction of examinees to formula scoring can be found in Cross and Frary (1977). More recently, Budescu and Bar-Hillel (1993) argued that

> Guessing is bad for test makers, not necessarily for test takers. Formula scoring was initially developed to discourage guessing. For the ideal test taker, however, formula scoring merely obviates guessing – and only random guessing at that. To really discourage guessing, the penalty for errors should exceed 1/(k-1).

In the next section of this paper, we shall address the motivations often leading test makers to formula scoring, and question their soundness. Furthermore, we argue that the actual incidence of guessing under right-only scoring cannot be assumed a priori, and should be experimentally assessed. Math items deserve a specific attention from this viewpoint: first, the diversity of available solving strategies (with respect to open-ended exercises) is more likely to be overlooked by test makers; second, the alternatives are often chosen to correspond to the possible outcomes of typical misconceptions or procedural errors, and it is therefore more likely that an examinee is led to a wrong answer instead of guessing at random. In the third section we report some preliminary results of a research project aimed at revealing the occurrence of guessing in math tests under right-only scoring, for first-year undergraduate students scarcely trained in multiple-choice tests (the standard present situation in Italy). We observe that encoding the assumption of systematic guessing in the IRT model used for the analysis of our test data produces a lower reliability for math items, while it increases the reliability for vocabulary and grammar items (on a different group of examinees).

2. Formula scoring: a critical perspective

We have already remarked that "naïve" motivations for the use of formula scoring rest on the paradigm that a multiple-choice questionnaire should work as a *collection of exercises*, each testing the knowledge of a specific method of solution. This paradigm obscures the fact that in most cases the set of answers proposed for each item provides additional information on the problem. Extracting and using this information is actually *a component of the ability one wishes to measure*; it is rather odd to pretend, for instance, that the examinees do not get the correct answer by exclusion, whenever this is the most efficient way to obtain it. One should accept the fact that *discriminating between the proposed answers* is the core of the response process in a multiple-choice test, and a "clever" guessing may be a perfectly legitimate strategy in some cases. It is true that the possibility of guessing prevents the test makers from inferring that an examinee choosing the right answer is actually able to solve the posed problem. Indeed, in a multiple-choice test it is the *full set of responses* to all questions that is relevant to measure the examinee's ability (Rasch, 1980).

Still, let us take for a moment the viewpoint of those who aim at wiping out any instance of guessing. Accordingly, we shall assume that the formula scoring has the ultimate power of inducing every examinee to omit the response if he/she is unable to single out the correct answer, whereby under the right-only score he/she would respond randomly. As stated by Lord (1975),

> The difference between an answer sheet obtained under formula-scoring directions [...] and the same answer sheet obtained under number-right scoring directions [...] is only that omitted responses, if any, on the former answer sheet are replaced by random guesses on the latter. (Lord, 1975, p. 8)

This conclusion is in accordance with the fact that whenever an examinee is at least able to exclude one or more answers, it is statistically advantageous for him/her to guess among the remaining answers, rather than omitting the response. However, if Lord's statement is true, then formula scoring is pointless. Let us suppose that in a test of n multiple-choice items (with k possible answers to each question), a given examinee is confident about the answers to s items ($s \leq n$), and uncertain on the remaining $n - s$ items. Let c be the number of right answers for the s "non-guessed" items ($c \leq s$: the examinee may have been misled in a number of cases), which is not affected by the scoring rule. In the right-only scoring scenario, the (average) expected score of the examinee would be $r_{RO} = c + \frac{n-s}{k}$. In the formula-scoring scenario, the score would be $r_{FS} = c - \frac{s-c}{k-1}$. Then, $r_{FS} = \frac{k}{k-1} r_{RO} - \frac{n}{k-1}$: the two scores are linearly related. If examinees' behavior changes with the scoring rules exactly as described by Lord, then *the average effect of the formula scoring directions only amounts to an overall rescaling of the expected scores*.

This conclusion, however, rests on the assumption that all examinees adopt the optimal answering strategy in either situation. As a matter of fact, this representation is not realistic. With respect to each item in a multiple-choice test, an examinee can be in one of three (subjective) states: absolute certainty, total uncertainty, or partial uncertainty. These states correspond, respectively, to being 100% sure of an answer, assigning a probability of $1/k$ to each answer, or assigning some non-uniform subjective probability distribution over the possible answers (Budescu *et al.*, 1993). Let us stress that the subjective state of an examinee with respect to an item does not reflect his/her *ability*, but rather his/her *beliefs*. The examinee's actual behavior, in the case of partial uncertainty, is in turn determined by the "additional trait" observed by Hakstian and Kansup (1975). Under formula-scoring directions, an examinee with a low level of self-confidence may omit responses even in cases where he/she had singled out the correct answer, but is not "completely sure" of it. To avoid such distortion, which may heavily affect the ability measure, all examinees should be *trained* to apply the optimal response strategy, which does *not* consist in "honestly" omitting the response whenever one is unable to solve correctly the problem, but instead in guessing systematically the answer whenever one or more alternatives can be excluded. The effect is paradoxical: "honest" examinees will eventually get a lower score with respect to "clever guessers". Furthermore, the strategy of judging the likelihood of each answer,

rather that solving the problem on the basis of the question alone, is promoted rather than discouraged. These effects are exactly opposite to the above-reported motivations for the adoption of formula scoring.

3. Guessing under right-only scoring: a case study

We have compared the results of three item pools: two of them (M1 and M2) were employed to test undergraduate students enrolling at the Faculty of Mathematical, Physical and Natural Sciences of the Turin University. M1 consists in a set of 86 math items administered (via computer) to 500 students in September 2004. The second set (M2) includes 58 math items, partly in common with the previous set, administered to 409 students in September 2005. The third set (Ph), considered for comparison purposes, contains 150 questions on Italian vocabulary and grammar, administered to 500 students enrolling at the Philosophy undergraduate course of the Turin University from 2004 to 2007.

Each student taking test M1 responded to a form containing 40 items: each item was extracted among two or more variants of comparable content and difficulty. In analogous way, the test forms extracted from M2 included 36 items, while those extracted from the Ph pool included 20 items. Each question had four alternative answers, only one being correct. All students were informed that zero points were assigned to either incorrect or omitted answers; examinees were neither advised to always respond, nor discouraged from guessing. Taking the test was compulsory for all enrolling students; the only formal consequence of failing the test was a more or less stringent advice to attend additional tutorial activities. Further details of the design of the three tests can be found in Andrà (2009).

Both sets of math items require calculation or problem solving to be answered correctly, while the third set of items (Ph) request mainly mnemonic knowledge. Our hypothesis is that, under right-only scoring, guessing is more likely to occur for the latter type of items.

The data have been analysed according to four different response models (Birnbaum, 1968): 1-PL (one-parameter logistic) model, 2-PL model, (unconstrained) 3-PL model and 3-PL with the constraint that the lower asymptote of all item characteristic curves (usually referred to as "guessing" or "pseudo-guessing" parameter) be fixed at c = 0.25.

Birnbaum's 1-PL model assumes that the probability of answering correctly to item *i* for the examinee *j (item characteristic curve)* is given by the expression

$$P_1(\vartheta_j, \beta_i) = \frac{1}{1+e^{\alpha(\beta_i - \vartheta_j)}}$$

where ϑ_j is the *ability* of the examinee, β_i is the *difficulty* of the item and α is an overall slope parameter, equal for all items. This model coincides with the Rasch model up to rescaling of the difficulty and ability scale by a factor α.

In the 2-PL model one has instead

$$P_2(\vartheta_j, \alpha_i, \beta_i) = \frac{1}{1+e^{\alpha_i(\beta_i - \vartheta_j)}}$$

(the slope parameter α_i is now different for each item). These models assume that the probability of correct answer tends to vanish when $\beta_i - \vartheta_j$ is large. This assumption is contradictory with the hypothesis that examinees being unable to solve the posed problem choose the answer at random, for in this case the probability should be equal to 1/*k* when $\beta_i >> \vartheta_j$. The 3-PL item characteristic curve is

$$P_3(\vartheta_j, \alpha_i, \beta_i, c_i) = c_i + \frac{1-c_i}{1+e^{\alpha_i(\beta_i - \vartheta_j)}}$$

whereby a different lower asymptote can be assigned to each item. For a set of *n* items, 3*n* free parameters have to be estimated under 3-PL model, in contrast to 2-PL (*2n* parameters), 1-PL (*n* + 1) or Rasch model (*n*) (Rasch, 1980). A larger number of parameters is expected to improve the goodness-of-fit and the overall reliability of the test estimates. If we constrain the parameter *c* in the 3-PL model to be equal to 1/*k*, the number of free parameters is the same as for the 2-PL model, but the latter should in principle fit better if guessing is negligible, while the former (constrained 3-PL) should have higher reliability if guessing occurs systematically.

The MMLE (maximum marginal likelihood) parameter estimates have been done using MULTILOG 7.0.3, which provides for each set of items and for each response model the resulting marginal reliability (Thissen, 1991). Table 1 shows the observed values for each test under the four models.

Test label	Marginal reliability			
	1-PL	2-PL	3-PL, unconstrained	3-PL, c = 0.25
M1	0.954	0.954	0.956	0.944
M2	0.925	0.930	0.937	0.915
Ph	0.915	0.957	0.982	0.979

Table 1: Observed MMLE marginal reliability.

As expected, the marginal reliability increases with the number of free parameters; yet for both math item pools, as soon as the lower asymptote of the item response curve is uniformly set to be 1/*k*, the reliability decreases not only with respect to the 2-PL model, but even in comparison to the 1-PL estimates. For the Ph pool, in contrast, the assumption that examinees guess at random when they don't know the correct answer still produces a much better fit with respect to 2-PL.

The software MULTILOG computes the G^2 statistics (-2 times the logarithm of the likelihood), which could in principle be used for a comparative test of fit between the different models (Maydeu-Olivares & Cai, 2006). The difference between constrained 3-PL model and the unconstrained 3-PL shows, as one should expect, a significant improvement of fit in the absence of constraints, for all the three item pools. The G^2 difference between constrained 3-PL and 2-PL model, in turn, is largely positive for the two math pools and negative for the (Ph) pool – quite in line with what has been observed for marginal reliability – but a proper

statistical test for the difference in goodness-of-fit would require one model to be "nested" into the other one, which is not the case.

4. Conclusion

Our observations support the feeling that guessing may be not a primary concern in mathematical multiple-choice testing under right-only scoring directions, at least in situations comparable to the studied one, and suggest that guessing is reduced when items require reasoning rather than mnemonic knowledge.

These results, to be complemented by a careful study of the applicability of the likelihood ratio test to compare the goodness-of-fit of the different models, should be regarded as preliminary to further studies based on large-scale tests at national level (started in September 2008). Such studies could be significant in view of a follow-up of the overall impact of the diffusion of multiple-choice assessment on mathematical teaching in Italy.

References

Andrà, C. (2009). Assessment of prerequisites for undergraduate studies through multiple-choice tests: the case of the University of Turin (Italy). *Ph. D. Dissertation*, Università degli Studi di Torino, Torino, IT.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In: F. M. Lord, & M. R. Novick (Eds.), *Statistical theories of mental scores*, (pp. 397-472), Reading, MA: Addison-Welsley.

Budescu, D. & Bar-Hillel, M. (1993). To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring. *Journal of Educational Measurement, 30*, 277-291.

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

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

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

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

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

Grandy, J. (1987). *Characteristics of examinees who leave questions unanswered on the GRE general test under right-only scoring.* Princeton, NJ: Educational Testing Service.

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

Holzinger, K. J. (1924). On scoring multiple-response tests. *Journal of Educational Measurement, 15*, 445-447.

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

Hakstian, A. R. & Kansup, W. (1975). A comparison of several methods of assessing partial knowledge in multiple-choice tests: II. Testing procedures. *Journal of Educational Measurement, 12*, 231-239.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement, 12*, 7-11.

Maydeu-Olivares, A. & Cai, L. (2006). A Cautionary Note on Using G^2(dif) to Assess Relative Model Fit in Categorical Data Analysis. *Multivariate Behavioral Research, 41*, 55-64

Patnaik, D. & Traub, R. E. (1973). Differential weighting by judged degree of correctness. *Journal of Educational Measurement, 10*, 281-286.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment test.* Chicago, IL: University of Chicago Press.

Thissen, D. (1991). *MULTILOG: multiple category item analysis and test scoring using item response theory*. Chicago, IL: Scientific Software International, Inc.

Thurstone, L. L. (1919). A method for scoring tests. *Psychological Bulletin, 16,* 235-240.

Traub, R. E. & Hambleton, R. K. (1972). The effects of scoring instructions and degree of speededness on the validity and reliability of multiple-choice tests. *Educational and Psychological Measurement, 32,* 737-758.