ORIGINAL ARTICLE

Impact of different scoring algorithms applied to multiple-mark survey items on outcome assessment: an in-field study on health-related knowledge

A. DOMNICH, D. PANATTO, L. ARATA, I. BEVILACQUA, L. APPRATO, R. GASPARINI, D. AMICIZIA
Department of Health Sciences, University of Genoa, Italy

Key words

Multiple-mark items • Multiple answer items • Pick-N items

Summary

Introduction. *Health-related knowledge is often assessed through multiple-choice tests. Among the different types of formats, researchers may opt to use multiple-mark items, i.e. with more than one correct answer. Although multiple-mark items have long been used in the academic setting – sometimes with scant or inconclusive results – little is known about the implementation of this format in research on in-field health education and promotion.*
Methods. *A study population of secondary school students completed a survey on nutrition-related knowledge, followed by a single-lecture intervention. Answers were scored by means of eight different scoring algorithms and analyzed from the perspective of classical test theory. The same survey was re-administered to a sample of the students in order to evaluate the short-term change in their knowledge.*
Results. *In all, 286 questionnaires were analyzed. Partial scoring algorithms displayed better psychometric characteristics than the dichotomous rule. In particular, the algorithm proposed by Ripkey and the balanced rule showed greater internal consistency and relative efficiency in scoring multiple-mark items. A penalizing algorithm in which the proportion of marked distracters was subtracted from that of marked correct answers was the only one that highlighted a significant difference in performance between natives and immigrants, probably owing to its slightly better discriminatory ability. This algorithm was also associated with the largest effect size in the pre-/post-intervention score change.*
Discussion. *The choice of an appropriate rule for scoring multiple-mark items in research on health education and promotion should consider not only the psychometric properties of single algorithms but also the study aims and outcomes, since scoring rules differ in terms of biasness, reliability, difficulty, sensitivity to guessing and discrimination.*

Introduction

Knowledge of topics related to lifestyle, health and healthcare may guide people's health-related choices and determine their health status. Assessment of these issues is growing, as any inadequacies identified can be targeted by specifically designed health promotion interventions [1]. Health-related knowledge may be seen as a part of health literacy [2], which is a broader concept defined as "the constellation of skills, including the ability to perform basic reading and numerical tasks required to function in the healthcare environment" [3]. Health literacy is considered a priority public health goal [4], as its low level is a well-known predictor of poor health-related outcomes (reviewed in [5]).

Factual knowledge on health-related topics is usually assessed by means of questionnaires and, in particular, multiple-choice tests or quizzes. Of the different types of question formats, multiple-choice (type-A or one-out-of-N items, i.e. with the single best option) and true-false items are among the most widely used to assess health-related knowledge. The widespread use of these two formats is probably due to such characteristics as their objectivity, rapidity in testing numerous subjects and receiving respondents' feedback, and the possibility of automatic scoring [6]. Alongside these strengths, however, multiple-choice questions also have weaknesses [6], such as reduced validity due to the possibility (or even encouragement) of guessing [7] and their failure to distinguish between partial knowledge and absence of knowledge [8, 9]. Alternative item formats may partially solve the shortcomings of type-A tests. Among these, multiple true-false (MTF, also known also as type-X) and multiple-mark (MM) items with several correct answers (also dubbed multiple choice multiple answer or pick-N items) have been extensively studied [10-19]. MTF tests, in which the respondent classifies each option as a separate true-false statement [14], are somewhat similar to MM items, in which the correct options chosen are regarded as true options while unmarked distracters are classed as false options. Indeed, Cronbach [10] has established that MTF and MM formats are very similar in terms of reliability, validity and respondents' performance. In certain situations, the MM format may be advantageous in terms of item construction, in that it allows more natural wording of both questions and response options and may need fewer distracters [15]. Pomplun and Omar [13] have demonstrated that MM questions are a feasible objective format and display acceptable reliability and validity, while Berk [20] has underlined that the MM

format preserves the main qualities of the type-A format while at the same time quantifying complex cognitive outcomes by assessing respondents' lines of reasoning in selecting answers. Moreover, MM items are useful in evaluating people with average and above-average knowledge of a topic [12].

One of the main issues regarding the MM format is the choice of an appropriate scoring rule. The most computationally simple scoring algorithm (SA) is the so-called dichotomous rule, whereby the respondent gets the full score for all correctly marked options, but nothing otherwise. An important drawback of the dichotomous SA, however, is its inability to give credit for partial knowledge; a respondent who gets all but one answer correct obtains the same score as one who is unable to provide any correct answer or even selects all wrong answers [15, 21, 22]. Indeed, in research on health education and promotion, laypeople's knowledge of health-related topics is often dubbed as partial knowledge.

In recent years, several SAs that are able to award partial credit, with or without penalties for guessing, have been developed and studied [12, 15, 16, 18, 19]. However, the results of these studies have often been inconsistent. Thus, Hsu et al. [12] established that none of the six SAs used in their study was significantly better than the others, while a partial-credit SA developed by Ripkey et al. [16] proved to be superior to the dichotomous SA in terms of item difficulty and discrimination parameters. These latter findings were later replicated by Bauer et al. [19], who documented the superiority of two different partial SAs to the dichotomous SA. More recently, the balanced SA, specifically designed for MM items, has been proposed as an improvement on Ripkey's algorithm [18].

Most of the above-mentioned studies were carried out in the academic setting in order to evaluate students' performances in exams and find an optimal item format. However, little is known about how different scoring rules applied to MM survey items would affect the evaluation of health-promotion outcomes. The present study therefore aimed to evaluate whether the choice of a scoring rule could impact on the evaluation of findings. Specifically, we posed two research questions: (1) do the psychometric properties of different SAs applied to the evaluation of factual health-related knowledge differ? and (2) do different SAs applied to the evaluation of factual health-related knowledge impact on the outcome?

Methods

Study design and setting

The nutrition-related knowledge of students from seven secondary schools in the Genoa metropolitan area was assessed in 2012/2013 by means of a self-administered paper-and-pencil survey. Participation was voluntary and anonymity was assured. No time limit was placed on compilation of the questionnaire, though students took less than 20 min.; survey administration was strictly supervised in order to prevent cheating. The study and the test were approved by the boards of each school.

This initial assessment of nutrition-related knowledge was followed by a single interactive lecture given by appropriately trained medical staff accompanied by teachers. The lecture lasted approximately 45 minutes and covered both general food- and nutrition-related topics (e.g. healthy diet, dietary recommendations, notions of macro- and micronutrients) and questions frequently asked by the students during the pre-intervention survey administration.

To evaluate changes in knowledge scores, the same survey was re-administered to a sample of students 2 weeks after the lecture.

Survey instrument for assessing nutrition-related knowledge

The factual nutrition-related knowledge part of the survey consisted of 14 items. Two knowledge items were excluded from the analysis, as formal flaws (poor specification of questions) were detected after survey administration; a total of 12 items were therefore analyzed. The survey also contained 7 perceived knowledge items (such as, *Do you know what carbohydrates are?*) and 2 open-ended items (such as, *What would you like to know about nutrition?*). These items were introduced after agreement with teachers, in order to plan the content of the upcoming lecture and of future school-based health-promotion interventions, and were not analyzed in the present study. Conceptually, the survey consisted of two nutrition-related topics, namely the understanding of food terms and the main sources of nutrients. Two formats were adopted: 9 items were MM, while the remaining 3 were type-A. The items did not conform to a single pattern; among the MM items, the number of options ranged from 4 to 8, the number of correct options from 2 to 5, and the number of distracters from 1 to 5. The type-A items had 2 or 3 distracters. To discourage guessing [23], a "don't know" option was also provided. All questionnaires were checked for quality control and responses were entered into an *ad hoc* database.

Scoring algorithms

The type-A items were scored by the conventional method: one point if the respondent marked only the keyed correct option and zero otherwise. To score the MM items, a total of eight SAs were implemented (Tab. I). The first was the dichotomous algorithm, which does not allow partial knowledge to be quantified ("all or nothing"); this SA has been widely used as a comparator versus partial SAs [12, 16, 18, 19]. The partial SAs 2-5 were adapted from the paper by Hsu et al. [12]; SAs 2, 4 and 5 involve some penalty for incorrectly chosen options, while SA3 does not. The formula of SA2 is similar to that of SA3, except for the fact that it penalizes incorrect answers; SA2 has been judged rather "severe" regardless of the number of marked distracters and unmarked correct answers provided by a respondent [12]. SA4 and its modifications are among the first methods of partial scoring described in the literature [24, 25]; SA4 consists of subtracting the

Tab. I. Description of scoring algorithms used in the study.

Scoring algorithm	Definition	Reference
SA1	S = 1 if IC = 0, otherwise S = 0	12, 16, 18, 19
SA2	S = (CC – IC)/TO	12
SA3	S = CC/TO	12
SA4	S = MCO/CO – (MIO/(TO – CO))	12
SA5	S = CC/TO – ((TO!/IC!·(TO – IC)!)/2^TO)	12
SA6	p = MCO/CO, if p > 0 ⇒ x = MO/TO – CO/TO, otherwise p = S; if x > 0 ⇒ S = p – (x/(1 – CO/TO)), otherwise p = S	18
SA7	S = MCO/CO if MCO ≤ CO, otherwise S = 0	16
SA8	S = 1 if IC = 0, S = 0.5 if 0.5·CO ≤ MCO < CO, otherwise S = 0	19

S: Respondent's score on a multiple-mark item (max = 1); CO: Number of keyed correct options; CC: Correct choices made by a respondent (both marked correct answers and unmarked distracters); IC: Incorrect choices made by a respondent (both marked distracters and unmarked correct answers); TO: Total number of item options; MCO: Correct options marked by a respondent; MIO: Incorrect options marked by a respondent; MO: Options marked by a respondent; p: Points for MCO; x: Penalty.

proportion of marked distracters from that of marked correct answers. SA5 involves a binomial coefficient and assumes that the incorrect choices made by a respondent are the result of guessing [12]. SAs 2, 3 and 5 treat MM items as MTF ones. SA6, known as balanced SA, has recently been described by Tarasowa and Auer [18]; it includes some logical operators and a penalty is applied only when the number of marked options exceeds that of keyed correct options. The SA7 proposed by Ripkey [16] yields a proportion-of-possible-points score only if the number of marked options does not exceed the number of keyed correct options. SA8, dubbed PS_{50} by Bauer et al. [19], awards the full score if all correct options are marked (no distracters must be marked), half the score if least 50% of correct options are marked, and zero points otherwise.

Items to which no response was given or the "don't know" option was selected were awarded zero points. The "don't know" option was not included in the count of the total number of options used for scoring and data analysis.

Scores of individual items were summed to produce a total score. By agreement between the research team and teachers, for scoring purposes all 12 items were assumed to have the same level of difficulty of 1; the highest possible score was therefore 12.

Independent variables

Demographic variables of age, sex and immigrant background were recorded from each participant. Body mass index (BMI) was calculated from self-reported height and weight, mapped to the BMI-for-age growth charts and classified in underweight (< 5th percentile), normal weight (5th-85th percentile), overweight (85th-95th percentile) and obese (≥ 95th percentile) categories.

Statistical analysis

For purposes of analysis, the factual nutrition-related knowledge part of the survey was divided (by item type format) into two subsets: the MM subset and the whole survey, which also included 3 type-A items.

Students' scores calculated according to the different SAs were compared by means of repeated-measures analysis of variance (rANOVA); the Greenhouse-Geisser correction for sphericity was applied by applying a significant Mauchly's test statistic. Post-hoc *t* tests for paired data, with *p*-values corrected by means of Bonferroni's method, were subsequently performed. Tarasowa and Auer [18] have suggested that the dichotomous SA1 should be used as a reference rule for scoring MM items (as it virtually excludes the probability of guessing) and that respondents' rankings should then be compared among different SAs; we therefore calculated Spearman's ρ coefficients with 95% confidence intervals (CIs) in order to compare students' rankings yielded by SA1 and the other seven SAs.

The psychometric properties of each SA were evaluated from the perspective of classical test theory. To measure internal consistency, Cronbach's α coefficients with 95% CIs were computed. The eight dependent α coefficients and subsequent pairwise comparisons with adjusted *p*-values were compared by means of Feldt's formulas [26, 27] implemented in the cocron R package [28]. The standard errors (SEs) of students' scores were determined as $SD\sqrt{1-\alpha}$, where SD is the standard deviation of the scores [29]. The efficiency of an SA was evaluated by means of the coefficient of effective length; two SAs with a coefficient of effective length of 1 were considered equally efficient (relative efficiency) [12]. The Spearman-Brown prophecy formula was applied in order to estimate the number of items needed to reach a desirable α of 0.7 and to compare the reliability coefficients of type-A and MM items, considering their different numbers. The item-difficulty index *p*, calculated as the mean score of an item, was categorized as "difficult" ($p < 0.2$), "acceptable" ($0.2 < p < 0.8$) and "easy" ($p \geq 0.8$) [29, 30]. The mean difficulty indexes of the eight SAs were compared by means of rANOVA. The item-discrimination index *D* was computed for each SA; items with $D > 0.2$ were considered acceptable [31].

The differences in the total scores according to the independent variables of interest (gender, immigration background and BMI categories) were quantified by applying standardized mean differences (SMDs) with 95% CIs; SMD was interpreted as large (0.8), medium (0.5) and small (0.2) [32]. Any association between the total score and the independent variables was checked by means of analysis of variance (ANOVA), while that

between the score and the participants' age was checked by means of Pearson's correlation coefficient *r*. These tests were performed separately for each SA.
Assuming an SMD of 0.5 between pre- and post-lecture scores when two-sided α is 0.05 and β is 0.9, we calculated that at least 44 subjects were needed. Cochran's *Q* test was performed to evaluate whether the different SAs had identical effects on the pre- to post-lecture change in individual scores (improved vs. not improved). The pre/post score changes were quantified by means of SMDs. The statistical significance level was conventionally set to two-sided $p < 0.05$. All data were analyzed by means of the R stats package, version 3.1.2 [33] and GPower, version 3.1.9.2 [34].

Results

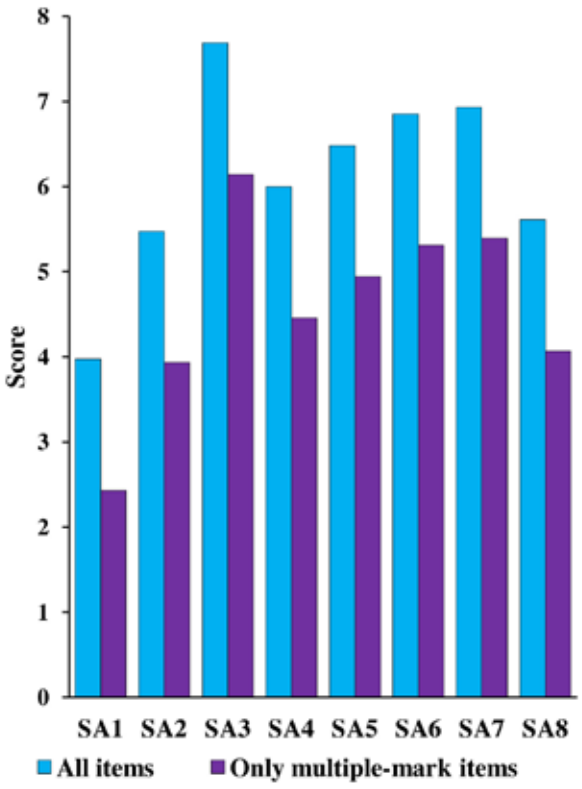
Sample characteristics

Students took an active part in the survey, completed questionnaires (total 298) being received from all participants. However, 12 questionnaires did not pass the quality check: 9 students had not attempted to answer any question, including demographic ones, while 3 questionnaires contained unlikely answers (such as improbable weight or height). These 12 were discarded and a total of 286 questionnaires were analyzed. Male and female students participated in approximately equal proportions (males: 51.0%) and their mean age was 16.1 (SD 1.1, range 14-20) years. Approximately a quarter of subjects [22.7% (95% CI: 18.0-28.0%)] were from an immigrant background. As calculated from self-reported height and weight, more than four fifths [82.2% (95% CI: 77.2-86.4%)] of students were of normal weight for their age and sex, while 2.1% (95% CI: 0.8-4.5%), 12.2% (95% CI: 8.7-16.6%) and 3.5% (95% CI: 1.7-6.3%) were classified as underweight, overweight and obese, respectively.

Difference in students' performance, by algorithm

As shown in Figure 1, the summary scores of the seven partial SAs were higher than those yielded by the dichotomous algorithm (Δ means: 1.50, 3.71, 2.03, 2.51, 2.88, 2.96 and 1.64 for SA2-8, respectively); as expected, the partial, non-penalizing SA3 yielded the highest scores. The mean scores of SA6 and SA7 were very close to each other; the mean scores yielded by SA7 were 1.1% and 1.5% higher than those of SA6 in the whole survey and the MM subset, respectively. rANOVA with corrected for sphericity (ε = 0.41) degrees of freedom showed a significant ($p < 0.001$) within-subject effect of SA on students' performances. All pairwise comparisons proved statistically significant. As shown by rank correlation coefficients (Tab. II), students' scores calculated according to SA8 were the most highly correlated with those of SA1; the lower limit of 95% CIs of ρ between SA1 and SA8 did not overlap with the upper limits of most of the other pairwise coefficients.

Fig. 1. Students' scores, by scoring algorithm and survey subset.



Psychometric properties of the scoring algorithms

There was a perceptible difference in the reliability measures of the SA: the dichotomous SA displayed a lower α coefficient (0.48) than any of the partial algorithms (Tab. III). Notably, considering only MM items, SA3, SA6 and SA7 increased their reliability coefficients, but only SA7 reached an $\alpha > 0.7$. The SE of measurement was lowest for SA3 (0.68), while SA1, SA2 and SA4 showed substantially higher SEs (1.08, 1.03 and 1.06, respectively). Analogously, the Spearman-Brown prophecy formula showed that, in order to achieve an α of 0.7, the number of items would need to be more than doubled for the dichotomous SA1, while for the balanced SA6 only three items would need to be added (Tab. III). The reliability coefficient of the three type-A

Tab. II. Spearman's ρ correlation coefficients between the dichotomous scoring algorithm 1 (SA1) and the other seven partial scoring rules applied to the multiple-mark survey subset (all $p < .001$).

Scoring algorithm	ρ	95% CI
SA2	0.79	0.74-0.83
SA3	0.73	0.67-0.78
SA4	0.74	0.68-0.79
SA5	0.82	0.78-0.85
SA6	0.81	0.77-0.85
SA7	0.78	0.73-0.82
SA8	0.88	0.85-0.90

Tab. III. Reliability measures of the scoring algorithms (SAs), by survey subset.

Scoring algorithm	α (95% CI)		N of items needed to reach α = 0.7	
	All (N = 12)	MM (N = 9)	All (N = 12)	MM (N = 9)
SA1	0.48 (0.38-0.56)	0.42 (0.32-0.52)	31	29
SA2	0.60 (0.53-0.67)	0.57 (0.49-0.64)	19	16
SA3	0.65 (0.59-0.71)	0.66 (0.60-0.72)	15	11
SA4	0.59 (0.51-0.65)	0.53 (0.45-0.61)	20	19
SA5	0.65 (0.58-0.71)	0.65 (0.59-0.71)	16	12
SA6	0.66 (0.60-0.72)	0.68 (0.63-0.74)	15	10
SA7	0.67 (0.62-0.73)	0.71 (0.65-0.76)	14	–
SA8	0.60 (0.53-0.68)	0.59 (0.52-0.66)	19	15

items was 0.32 (95% CI: 0.17-0.44). The projected coefficient for N = 9 type-A items was estimated to be 0.58, which was lower than the α coefficients of the 9 MM items scored according to SAs 3, 5-8 (Tab. III).

As demonstrated by the coefficients of effective length (Tab. IV), SA1 was the least efficient algorithm, while SA7 was the most efficient. More generally, SAs 3, 5-7 were at least twice as efficient as SA1. The eight reliability coefficients of both survey subsets differed significantly ($p < 0.001$). Several pairwise comparisons of α coefficients proved statistically significant in both survey subsets (Tab. V). In the MM survey subset, the α of SA7 was significantly higher than those of the other seven SAs, while, considering all items, the α of SA7 did not differ significantly from those of SA3 and SA6.

The mean difficulty index (Tab. VI) was the lowest when SA1 was applied, while the quiz was the "easiest" when SA3 was used. The differences among mean difficulty coefficients adjusted for sphericity violations proved to be highly significant in both subsets ($p < 0.001$). All type-A items had difficulty indexes p between 0.2 and 0.8; thus the numbers of easy and difficult items in both survey subsets matched. The highest number (N = 4) of difficult items ($p < 0.2$) was observed when SA1 was used, while according to SAs 3, 5-7, no difficult questions were present in the survey. Conversely, according to SA3, three items were classified as easy ($p > 0.8$), while none were when the dichotomous SA1 was applied.

The item discrimination analysis reported in Table VII did not reveal any negative total item correlation coefficient, while the number of items with $D > 0.2$ varied by SA. SA2 and SA4 showed slightly higher mean discrimination indexes; notably, all MM items scored by SA4 had desirable point-biserial coefficients.

Tab. IV. Relative efficiency of the scoring algorithms, as measured by the coefficient of effective length of all items (upper right triangle) and only multiple-mark items (lower left triangle).

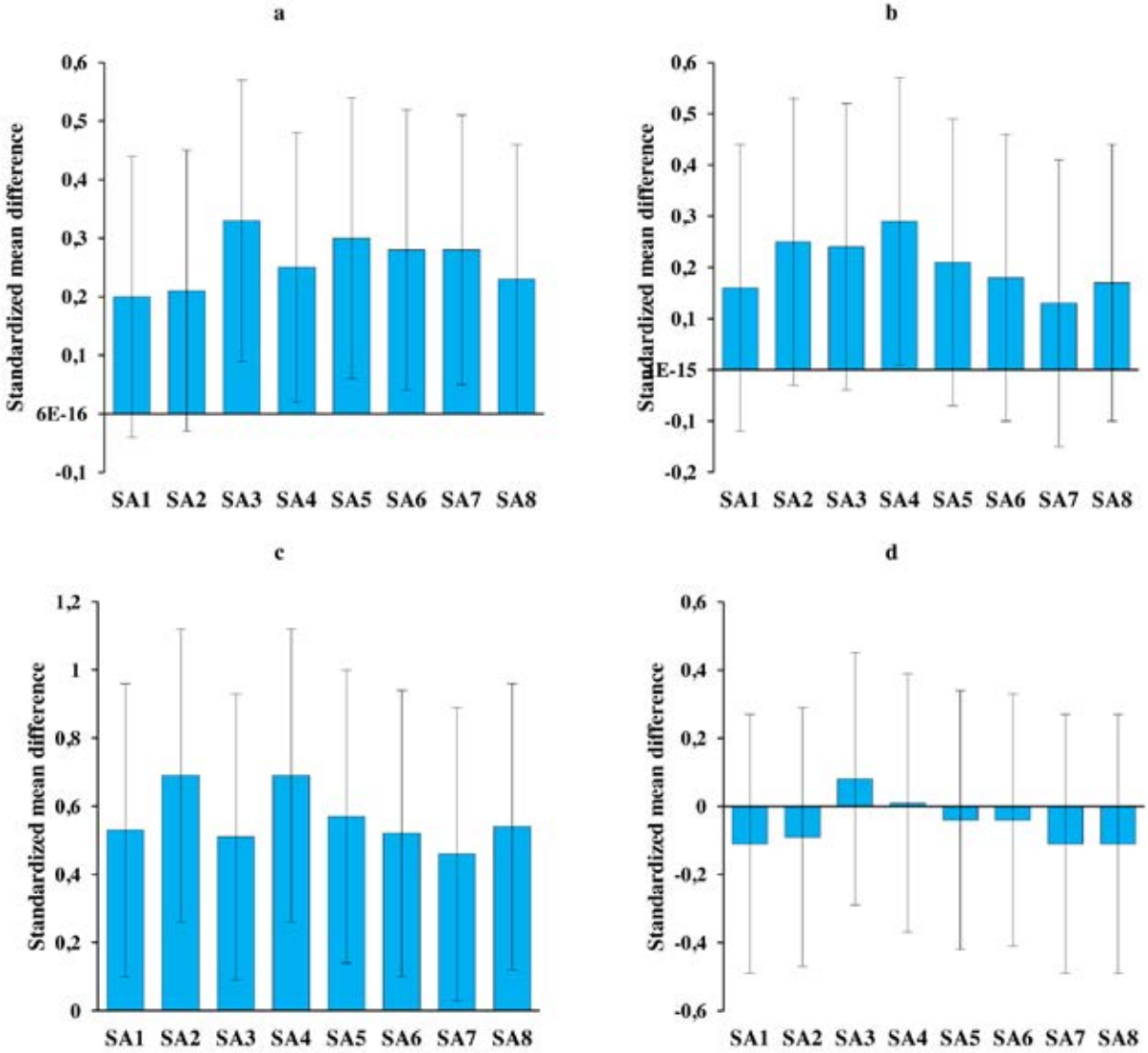
Scoring algorithm	SA1	SA2	SA3	SA4	SA5	SA6	SA7	SA8
SA1	–	1.63	2.01	1.56	2.01	2.10	2.20	1.63
SA2	1.83	–	1.24	0.96	1.24	1.29	1.35	1.00
SA3	2.68	1.46	–	0.77	1.00	1.05	1.09	0.81
SA4	1.56	0.85	0.58	–	1.29	1.35	1.41	1.04
SA5	2.56	1.40	0.96	1.65	–	1.05	1.09	0.81
SA6	2.93	1.60	1.09	1.88	1.14	–	1.05	0.77
SA7	3.38	1.85	1.26	2.17	1.32	1.15	–	0.74
SA8	1.99	1.09	0.74	1.28	0.77	0.68	0.59	–

Tab. V. Pairwise comparisons* of Cronbach's α coefficients of all items (upper right triangle) and only multiple-mark items (lower left triangle)..

Scoring algorithm	SA1	SA2	SA3	SA4	SA5	SA6	SA7	SA8
SA1	–	< 0.001	< 0.001	0.011	< 0.001	< 0.001	< 0.001	< 0.001
SA2	0.003	–	0.036	0.99	< 0.001	< 0.001	< 0.001	0.99
SA3	< 0.001	< 0.001	–	< 0.001	0.99	0.99	0.94	0.046
SA4	0.31	0.23	< 0.001	–	< 0.001	< 0.001	< 0.001	0.99
SA5	< 0.001	< 0.001	0.99	< 0.001	–	0.99	0.048	0.001
SA6	< 0.001	< 0.001	0.99	< 0.001	0.12	–	0.51	< 0.001
SA7	< 0.001	< 0.001	0.037	< 0.001	< 0.001	0.004	–	< 0.001
SA8	< 0.001	0.99	0.018	0.99	0.001	< 0.001	< 0.001	–

*: Results are reported as p-values corrected by means of Bonferroni's method.

Fig. 2. Standardized mean differences in total scores between females and males (a), native and immigrant students (b), native male and immigrant male students (c) and native female and immigrant female students (d), by scoring algorithm.



In-field assessment: impact of the scoring algorithm on outcome assessment

Neither BMI category nor age was associated with the total score yielded by any SA. Female students scored significantly higher than males on 5 of the 7 partial SAs. The effect size was, however, judged small. By contrast, SA1, SA2 and SA8 were unable to highlight the effect of gender on the respondents' nutrition knowledge (Fig. 2a). Foreign students tended to score lower than Italians, though the difference reached the significance level (low effect size of 0.29) only when SA4 was applied (Fig. 2b). However, the total score yielded by most algorithms was probably determined by a combined effect of gender and immigration background; foreign-born males scored much lower than native male students (Fig. 2c), while no obvious pattern emerged regarding differences in scores between immigrant and Italian females (Fig. 2d). The effect size in scores between foreign and native male students was medium for all but one (SA7) rule. Final ANOVA models of the main effects of gender and nationality and their interaction confirmed the results of univariable statistics, although patterns of main and interaction effects differed by SA (Tab.VIII).

A total of 42 students completed the post-lecture survey. Most students improved their pre-lecture scores, though the proportions differed significantly ($p = 0.006$) by SA (54.8%, 76.2%, 73.8%, 66.7%, 71.4%, 61.9%, 66.7% and 64.3% on using SA1–SA8, respectively). Figure 3 reports effect sizes for pre- and post-lecture scores. The highest effect sizes were observed for SA4 (0.60) and SA2 (0.53), and were judged medium, while the other SAs displayed low effect sizes.

Discussion

The present study investigated the application of eight different scoring rules for MM items and demonstrated

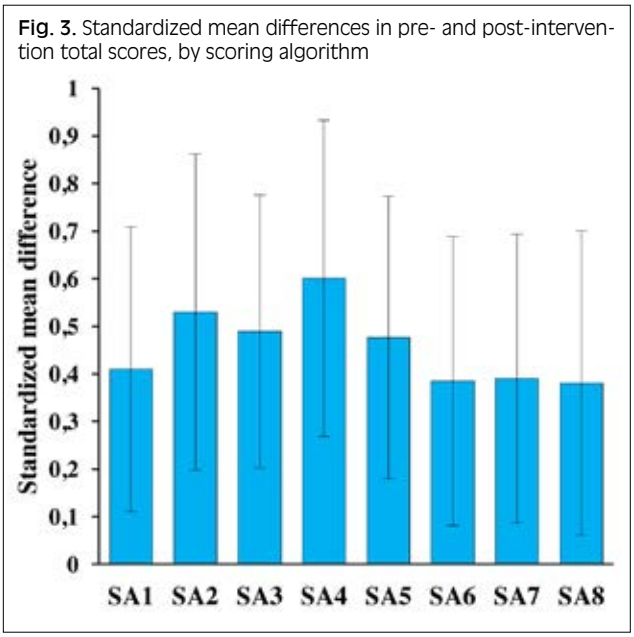


Fig. 3. Standardized mean differences in pre- and post-intervention total scores, by scoring algorithm

how they may affect the evaluation of health education outcomes. In line with previous findings [16, 18, 19], we found greater internal consistency and relative efficiency of Ripkey's rule (SA7) and its modifications, such as the balanced algorithm (SA6), in scoring MM items. The SA proposed by Ripkey also performed comparatively well with regard to item difficulty. We found that the choice of SA may have a great influence on student performance; application of the non-penalizing SA3 approximately doubled the total score yielded by the dichotomous SA1. This finding is consistent with previous research [12, 16, 18, 19], which has indicated that MM items scored dichotomously are relatively difficult. In addition, our results support those of Tarasowa and Auer [18], in that SA6 and SA7 penalized respondents more than SA3, which is an MTF-like algorithm, but less than the other rules. This may have important implications for the comparison of knowledge scores obtained from different studies. For instance, previous European studies on nutrition knowledge among adolescents [35-38] have found about 60% of correct responses in multiple choice tests, which roughly corresponds to our estimate (mean percent scores of 54%-64%) obtained by applying partial SAs 3,5-7. By contrast, the dichotomous algorithm produced a substantially lower score of 33%.

Despite the somewhat superior psychometric properties of the Ripkey and the balanced scoring rules, our analysis revealed that SA4 was the only one that identified the negative impact of an immigrant background on the total score. This observation was probably due to a slightly higher discriminatory ability of SA4. The relationship between immigrant status and knowledge scores seems to be plausible; indeed, a large European study [35] conducted in nine countries found a 10% difference in nutrition knowledge scores between native and immigrant adolescents. This coincides with our estimate of a 10.6% mean score difference between Italian and migrant teenagers. On the other hand, the association between immigrant background and knowledge scores probably depends on sex, as shown by the fact that foreign-born male students displayed the poorest performance, regardless of the scoring rule used. Similarly, in the quasi-experimental part of the study, all algorithms were able

Tab. VI. Difficulty parameters of survey items, as measured by the different scoring algorithms, by survey subset.

Scoring algorithm	Mean difficulty, *p* (SD)		N of easy items	N of difficult items
	All (N = 12)	MM (N = 9)		
SA1	0.33 (0.27)	0.27 (0.25)	0	4
SA2	0.46 (0.29)	0.44 (0.32)	1	2
SA3	0.64 (0.20)	0.68 (0.17)	3	0
SA4	0.50 (0.26)	0.50 (0.28)	1	2
SA5	0.54 (0.22)	0.55 (0.22)	1	0
SA6	0.57 (0.20)	0.59 (0.20)	1	0
SA7	0.58 (0.21)	0.60 (0.20)	1	0
SA8	0.47 (0.25)	0.45 (0.26)	1	2

Tab. VII. Discrimination parameters of survey items, as measured by the different scoring algorithms, by survey subset.

Scoring algorithm	Mean discrimination index, *D* (SD)		N of items with $D > 0.2$	
	All (N = 12)	MM (N = 9)	All (N = 12)	MM (N = 9)
SA1	0.33 (0.21)	0.35 (0.28)	8	5
SA2	0.36 (0.12)	0.38 (0.16)	11	8
SA3	0.30 (0.16)	0.30 (0.10)	8	6
SA4	0.36 (0.12)	0.37 (0.15)	11	9
SA5	0.32 (0.15)	0.30 (0.14)	9	7
SA6	0.32 (0.14)	0.30 (0.11)	9	8
SA7	0.32 (0.13)	0.31 (0.12)	9	8
SA8	0.32 (0.16)	0.32 (0.16)	10	6

Tab. VIII. Results of ANOVA models for sex and nationality, by scoring algorithm.

Scoring algorithm	Sex		Nation		Sex : Nation	
	F	p	F	p	F	p
SA1	2.88	0.091	1.87	0.17	5.43	0.021
SA2	3.42	0.066	3.90	0.049	8.05	0.005
SA3	8.07	0.005	4.13	0.043	3.07	0.081
SA4	4.66	0.032	5.53	0.019	6.47	0.012
SA5	6.64	0.011	3.28	0.071	5.35	0.021
SA6	5.79	0.017	2.53	0.11	4.41	0.037
SA7	5.62	0.018	1.44	0.23	4.40	0.037
SA8	3.84	0.051	1.98	0.16	5.49	0.020

to highlight the efficacy of the single-lecture intervention in improving students' scores, although SA4 yielded the highest effect size for the pre-/post-intervention difference. Given the above-mentioned patterns, we believe that the choice of an SA for MM items should take into account not only the psychometric properties of single SAs but also the study aims, study population and research topic. This supports the principal conclusions of Muijtjens et al. [39], who suggested that the choice between less biased number-right (e.g. one point for each correct response) and more reliable formula-based scoring rules should be balanced by considering several education factors. For instance, it is acknowledged that females know more about nutrition than males do [35, 40, 41]; it has also been established that females are less likely to guess in multiple-choice tests [42, 43]. It could therefore be speculated that SAs with a correction for guessing would, to some extent, adjust for gender difference in scores. Furthermore, the choice of scoring rule for MM items may affect the statistical power of the analysis, and thus somehow alter outcome assessment. An appropriate SA should therefore be chosen during the design and planning (e.g. sample size calculation) of surveys on health-related knowledge containing this type of item format.

More generally, our results support the principal findings and conclusions of earlier studies [10, 13, 18, 19] on the feasibility of the MM format, since the MM items scored by most of the partial algorithms displayed at least equal internal consistency of the type-A items. MTF and MM items are not rare in health-related knowledge surveys [44, 45], including food- and nutrition-related ones [46], and these items have usually been scored by means of the conventional number-right method. Nevertheless, the guidelines for assessing nutrition-related knowledge, attitudes and practices issued by the *Food and Agriculture Organization of the United Nations* [47] discourage the use of multiple-choice and true-false formats because of the probability of "lucky guessing", and thus overestimation of knowledge scores. However, a correctly guessed answer may be the result of either a blind guess (i.e. a random response given by a fully uninformed subject) or an educated guess (i.e. a response given by a partially informed subject) [48]. Despite its main disadvantage of giving no credit for partial knowledge, use of the dichotomous rule in scoring MM items almost excludes the measurement error due to blind guesses [18]. SA8 showed the highest rank correlation with the dichotomous reference rule (SA1); this confirms the findings of Bauer et al. [19], which indicated that partially scored MM items with a 50% threshold of correct answers may separate the two types of guessing. Scoring MM items as MTF items did not yield any advantage; SAs 2, 3 and 5 neither displayed better psychometric characteristics nor were superior to the others in the on-field outcome evaluation. Despite some similarities between MM-item and MTF-item structures, Cronbach [10, 49] noted a significant difference in questions marked as true, and dubbed this an "acquiescence bias"; poor respondents tend to perform better on items to which the correct answer is "true" rather than "false". In turn, this bias contributes to the skewness of responses [18]. An added advantage of algorithms, especially the balanced SA6, that do not treat MM items as MTF items, is that they allow both MM and type-A questions to be scored. In other words, MM items scored in accordance with SA6 and similar rules make these items a "subspecies" of the type-A items widely used and recognized in health education/promotion research [18].

Overall, our sample may be considered as representative of the adolescent population of Genoa. Furthermore, the distribution of BMI was very close to the estimates obtained by the *Health Behavior in School-Aged Children* (HBSC) study [50] in the Liguria region (underweight: 2.3%, normal weight: 83.1%, overweight: 13.2%, and obese: 1.5%). A very high participation rate enabled us to minimize the response bias. Alongside its strengths, the present study had some limitations. First of all, we used a survey instrument that had not been fully validated, although it was highly comprehensible (as shown by a Gulpease readability index of 78.4, i.e. easy for subjects with a middle-school education) and sensitive to changes. Secondly, relatively low reliability coefficients of the knowledge part of the questionnaire were observed; this was probably due to the small number of survey items. However, Cronbach's α of > 0.6 is still acceptable [51] and the coefficients yielded by some partials SAs were comparable to those of well-established literacy instruments (e.g. the Spanish version of the New Vital Sign has an α of 0.69 [52]).

In conclusion, the past few years have seen a revival of the use of MM items to assess factual knowledge [22],

including health-related knowledge [44-46]. In research on health education and promotion, the choice between number-right and formula-based scoring rules, and between formulas that penalize guessing and those that do not, should balance the psychometric properties of single scoring rules and the outcomes of interest. The dichotomous "all or nothing" algorithm should be applied with caution to MM items, especially in cross-sectional study designs, owing to its poorer reliability, item difficulty and discrimination properties. Considering its high sensitivity to blind guessing, we believe that implementation of the dichotomous scoring rule should be limited to highly standardized survey instruments with excellent content validity. However, since school-based health-promotion interventions often require close collaboration with teachers in preparing knowledge-evaluation surveys, the validity of these questionnaires may be far from optimal. In the present study, the scoring rule proposed by Ripkey [16] and the balanced algorithm described by Tarasowa and Auer [18] showed greater internal consistency and relative efficiency in scoring MM items, while the penalizing SA4 was associated with largest effect sizes in the in-field evaluation.

References

[1] Domnich A, Panatto D, Signori A, et al. *Uncontrolled web-based administration of surveys on factual health-related knowledge: a randomized study of untimed versus timed quizzing*. J Med Internet Res 2015;13:17:e94.

[2] Baker DW. *The meaning and the measure of health literacy*. J Gen Intern Med 2006;21:878-83.

[3] Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs, American Medical Association. *Health literacy: report of the Council on Scientific Affairs*. JAMA 1999;281:552-7.

[4] Nutbeam D. *Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century*. Health Promot Int 2000;15:259-67.

[5] Dewalt DA, Berkman ND, Sheridan S, et al. *Literacy and health outcomes: a systematic review of the literature*. J Gen Intern Med 2004;19:1228-39.

[6] Roberts TS. *The use of multiple choice tests for formative and summative assessment*. In: *Proceedings of the 8th Australasian Conference on Computing Education*. Australian Computer Society 2006;52:175-80.

[7] Newble DI, Baxter A, Elmslie RG. *A comparison of multiple-choice tests and free-response tests in examinations of clinical competence*. Med Educ 1979;13:263-8.

[8] Chang SH, Lin PC, Lin ZC. *Measures of partial knowledge and unexpected responses in multiple-choice tests*. Educ Technol Soc 2007;10:95-109.

[9] Lau PNK, Lau SH, Hong KS, et al. *Guessing, partial knowledge, and misconceptions in multiple-choice tests*. Educ Technol Soc 2011;14:99-110.

[10] Cronbach LJ. *An experimental comparison of the multiple true-false and multiple multiple-choice tests*. J Educ Psychol 1941;32:533.

[11] Albanese MA, Kent TH, Whitney DR. *Cluing in multiple-choice test items with combinations of correct responses*. Acad Med 1979;54:948-50.

[12] Hsu TC, Moss PA, Khampalikit C. *The merits of multiple-answer items as evaluated by using six scoring formulas*. J Exp Educ 1984;52:152-8.

[13] Pomplun M, Omar MH. *Multiple-mark items: An alternative objective item format?* Educ Psychol Meas1997;57:949-62.

[14] Frisbie DA, Druva CA. *Estimating the reliability of multiple true-false tests.* J Educ Meas 1986;23:99-105.

[15] Duncan GT, Milton EO. *Multiple-answer multiple-choice test items: responding and scoring through Bayes and minimax strategies*. Psychometrika 1978;43:43-57.

[16] Ripkey DR, Case SM, Swanson DB. *A 'new' item format for assessing aspects of clinical competence*. Acad Med 1996;71(Suppl. 10):S34-6.

[17] Bandaranayake R, Payne J, White S. *Using multiple response true-false multiple choice questions*. Aust N Z J Surg 1999;69:311-5.

[18] Tarasowa D, Auer S. *Balanced scoring method for multiple-mark questions.* CSEDU 2013 – Proceedings of the 5th International Conference on Computer Supported Education 2013, pp. 411-6.

[19] Bauer D, Holzer M, Kopp V, et al. *Pick-N multiple choice-exams: a comparison of scoring algorithms.* Adv Health Sci Educ Theory Pract 2011;16:211-21.

[20] Berk RA. *A consumer's guide to multiple-choice item formats that measure complex cognitive outcomes*. Available at: http://images.pearsonassessments.com/images/NES_Publications/1996_12Berk_368_1.pdf

[21] Frary RB. *Partial-credit scoring methods for multiple-choice tests.* Appl Meas Educ 1989;2:79-96.

[22] Verbić S. *Information value of multiple response questions*. Psihologija 2013;45:467-85.

[23] Pennington HR, Pachana NA, Coyle SL. *Use of the facts on aging quiz in New Zealand: validation of questions, performance of a student sample, and effects of a don't know option*. Educ Gerontol 2001;27:409-16.

[24] Dressel PL, Schmid J. *Some modifications of the multiple-choice item*. Educ Psychol Meas1953;13:574-95.

[25] Morgan MRJ. *MCQ: An interactive computer program for multiple-choice self-testing*. Biochem Educ 1979;7:67-9.

[26] Feldt LS, Woodruff DJ, Salih FA. *Statistical inference for coefficient alpha*. Appl Psychol Meas1987;11:93-103.

[27] Feldt LS. *A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample*. Psychometrika 1980;45:99-105.

[28] Package "cocron". Available at: http://cran.r-project.org/web/packages/cocron/cocron.pdf

[29] Kline P. *The handbook of psychological testing. 2nd edition*. Abingdon: Routledge 1993.

[30] Mackison D, Wrieden WL, Anderson AS. *Validity and reliability testing of a short questionnaire developed to assess consumers' use, understanding and perception of food labels*. Eur J Clin Nutr 2010;64:210-7.

[31] Crocker L, Algina J. *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston 1986.

[32] Cohen J. *Statistical power analysis for the behavioral sciences. 2nd edition*. Hillsdale, NJ: Lawrence Erlbaum Associates 1988.

[33] R Core Team R. *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing 2014. Available at: http://www.R-project.org/.

[34] Faul F, Erdfelder E, Buchner A, et al. *Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses*. Behav Res Methods 2009;41:1149-60.

[35] Sichert-Hellert W, Beghin L, De Henauw S, et al. *Nutritional knowledge in European adolescents: results from the HELENA (Healthy Lifestyle in Europe by Nutrition in Adolescence) study*. Public Health Nutr 2011;14:2083-91.

[36] Reinehr T, Kersting M, Chahda C, et al. *Nutritional knowledge of obese compared to non obese children*. Nutr Res 2003;23:645-9.

[37] Osler M, Hansen ET. *Dietary knowledge and behaviour among*

schoolchildren in Copenhagen, Denmark. Scand J Soc Med 1993;21:135-40.

[38] Cunningham-Sabo LD, Davis SM, Koehler KM, et al. *Food preferences, practices, and cancer-related food and nutrition knowledge of southwestern American Indian youth*. Cancer 1996;78(Suppl. 7):1617-22.

[39] Muijtjens AM, Mameren HV, Hoogenboom RJ, et al. *The effect of a 'don't know' option on test scores: number-right and formula scoring compare*d. Med Educ 1999;33:267-75.

[40] Shepherd R, Towler G. *Nutrition knowledge, attitudes and fat intake: application of the theory of reasoned action*. J Hum Nutr Diet 2007;20:159-69.

[41] Shah P, Misra A, Gupta N, et al. *Improvement in nutrition-related knowledge and behaviour of urban Asian Indian school children: findings from the 'Medical education for children/ Adolescents for Realistic prevention of obesity and diabetes and for healthy aGeing' (MARG) intervention study*. Br J Nutr 2010;104:427-36.

[42] Slakter MJ, Koehler RA, Hampton SH, et al. *Sex, grade level, and risk taking on objective examinations*. J Exp Educ1971;39:65-8.

[43] Ben-Shakhar G, Sinai Y. *Gender differences in multiple-choice tests: the role of differential guessing tendencies*. J Educ Meas 1991;28:23-35.

[44] Suda AL, Jennings F, Bueno VC, et al. *Development and validation of Fibromyalgia Knowledge Questionnaire: FKQ*. Rheumatol Int 2012;32:655-62.

[45] Maciel SC, Jennings F, Jones A, et al. *The development and validation of a Low Back Pain Knowledge Questionnaire – LKQ*. Clinics (Sao Paulo) 2009;64:1167-75.

[46] Byrd-Bredbenner C, Wheatley V, Schaffner D, et al. *Development and implementation of a food safety knowledge instrument*. J Food Sci Education 2007;6:46-55.

[47] Macías YF, Glasauer P. *Guidelines for assessing nutrition-related knowledge, attitudes and practices.* Rome: Food and Agriculture Organization of the United Nations 2014.

[48] Mondak JJ, Davis BC. *Asked and answered: Knowledge levels when we will not take "don't know" for an answer*. Polit Behav 2001;23:199-224.

[49] Cronbach LJ. *Studies of acquiescence as a factor in the true-false test*. J Educ Psychol 1942;33:401-15.

[50] Health Behavior in the School-Aged Children study: rapporto sui dati 2010. Rapporto ISTISAN 13/5, 2013. Available at: http://www.hbsc.unito.it/it/images/pdf/hbsc/report_nazionale_2010.pdf

[51] Nunnally J. *Psychometric theory. Second edition*. New York: McGraw-Hill 1978.

[52] Weiss BD, Mays MZ, Martz W, et al. *Quick assessment of literacy in primary care: the newest vital sign*. Ann Fam Med 2005;3:514-22.

■ Received on June 10, 2015. Accepted on October 30, 2015.

■ Correspondence: Alexander Domnich, Department of Health Sciences, University of Genoa, via Pastore 1, 16132 Genoa, Italy - Tel. +39 010 3538524 - Fax +39 010 3538541 - E-mail: alexander.domnich@gmail.com