TPM®

DO MIXED ITEM FORMATS THREATEN TEST UNIDIMENSIONALITY? RESULTS FROM A STANDARDIZED MATH ACHIEVEMENT TEST

DAVIDE MARENGO
UNIVERSITY OF TORINO

RENATO MICELI
UNIVERSITY OF TORINO
UNIVERSITY OF VALLE D'AOSTA

MICHELE SETTANNI
UNIVERSITY OF TORINO

Numerous studies suggest that while multiple-choice (MC) and constructed-response (CR) items for the most part measure the same construct, a residual local dependence among CR items producing multidimensionality can be observed in mixed-format tests. The main aim of this study was to investigate the existence of format-related multidimensionality in a standardized math test administered as part of the Italian statewide educational assessment program implemented by INVALSI in 2012. Working within the framework of Item Response Theory (IRT), we tested the hypothesis that a bidimensional compensatory IRT model letting MC and CR items load on a common latent dimension, and the CR items additionally load on an a secondary dimension would be more appropriate than a unidimensional model for the examined test. The existence of a secondary CR ability dimension was not fully supported by the results. Specific considerations for practitioners and future research are presented.

Key words: Item Response Theory; Achievement Test; Large-Scale Assessment; Test Dimensionality; Item format.

Correspondence concerning this article should be addressed to Michele Settanni, Department of Psychology, University of Torino, Via Verdi 10, 10124 Torino, Italy. Email: michele.settanni@unito.it

The concomitant inclusion of multiple-choice (MC) items and constructed-response (CR) items in standardized tests is common practice in large-scale educational assessment programs. Many researchers and practitioners claim that this testing strategy offers several advantages over assessments based on MC items only. Within the context of Bloom's well-known Taxonomy of Learning Goals (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956), criticism concerning the use of MC-based assessment has been mostly related to the supposed inability of such format to elicit cognitive skills beyond the knowledge level (Boodoo, 1993). Recent findings however suggest that MC items can also be designed to tap higher level processes (Hancock, 1994; Simkin & Kuechler, 2005). Still, it is a widely shared belief that due to their generally higher complexity and by emphasizing divergent production abilities, CR items test a deeper understanding of the subject material (Martinez, 1999). Recently, Kuechler and Simkin (2010) reported findings sup-

porting this belief: when comparing examinees' performances on MC and CR items designed to assess proficiency on the subject at the same cognitive level, the authors found that the procedural knowledge recalled by the CR questions was generally more sophisticated than the simple facts recalled to answer the MC questions. The use of CR questions is also expected to reduce the impact of item-guessing strategies, which in turn represent one of the major practical drawbacks related to the use of MC items, in spite of their relatively shorter administration time and easiness of scoring (Ercikan et al., 1998; Zimmerman & Williams, 2003).

When both MC and CR items are included in the same test, however, specific measurement issues may emerge. Concerning measurement invariance, significant differences in item functioning across genders have been reported when comparing MC and CR items, although varying in size and direction across age groups (Wilson & Zhang, 1998). Findings indicate the existence of a gender-guessing interaction effect favoring male examinees on MC items (Ben-Shakhar & Sinai, 1991; Walstad & Robinson, 1997), while CR tend to favor female examinees (Arthur & Everaert, 2012; DeMars, 2000; Ghorpade & Lackritz, 1998). When CR items are designed as extended open-ended questions (as in contrast to short-answer CR items), differential item functioning has also been shown to emerge when comparing groups characterized by different levels of writing proficiency (Walker & Beretvas, 2001). Differential performances across item formats emerged even when comparing students with different levels of proficiency on the assessed domain (Wang, 2002).

Findings concerning the dimensionality characteristics of composite tests are mixed. Early studies are consistent in describing the constructs underlying MC and CR items as nearly overlapping (Bennett, Rock, & Wang, 1991; Bridgeman, 1991; Wainer & Thissen, 1993). Findings from later studies, on the other side, suggest that while MC and CR items for the most part measure the same construct, a residual local dependence among CR items producing multidimensionality can be present in mixed-format tests (Ercikan et al. 1998; Lissitz, Hou, & Slater, 2012; Manhart, 1996; Perkhounkova & Dunbar, 1999; Thissen, Wainer, & Wang, 1994; Walker & Beretvas, 2003). Moreover, construct equivalence between the MC and CR format specific dimensions have been shown to increase when MC and CR items are designed to be stem and content equivalent, and to decrease when CR items are presented as essay questions as opposed to short-answer CR (Rodriguez, 2003). In standard testing practice, however, MC and CR items are often purposely designed to assess examinees' knowledge on different content domains and cognitive levels, rendering those strict indications difficult, if not impossible, to comply with.

Still, mixed-format achievement tests are commonly intended by test designers to measure single unidimensional constructs; accordingly, examinees' responses on the tests are then scored as to provide a single measure of proficiency. When the unidimensionality assumption is not fully met by the test, however, such measure does not represent a reliable index of student's proficiency on the targeted construct. Still, public institutions usually analyze aggregate distributions of these measures to inform policymakers' decision process. Hence, it is paramount to ensure that the measurement process is psychometrically sound.

In this view, the main aim of the present study was to investigate the presence of multidimensionality related to item format in a standardized test for the assessment of mathematical proficiency. The test was administered in the context of the Italian statewide assessment program designed and implemented by the Italian National Institute for the Evaluation of the Education System (INVALSI), as part of the compulsory final examination, which marks Italian 8th graders' passage from first- to second-grade secondary education. The test includes non-stem-equivalent MC and CR items designed to assess students' knowledge on multiple math contents domains. However, IN-

VALSI does not provide any documentation concerning the dimensionality of the test. Analyses conducted on previous examinations (2008-2011) indicated the presence of minor violations of Rasch unidimensionality in the INVALSI math tests due to the inclusion of mixed-item response formats (Miceli, Marengo, Molinengo, & Settanni, 2015). The present study further explores this subject by operating within the framework of Item response Theory (IRT). Our hypothesis is that a bidimensional compensatory IRT model letting MC and CR items load on a common latent dimension while the CR items additionally load on a second auxiliary dimension, may be more appropriate for the examined test than a unidimensional IRT model.

As a secondary aim, we investigated the relationship between, respectively, the primary math ability and secondary CR ability as modeled with the bidimensional model and a set of relevant grouping variables — that is, gender, citizenship status, and regularity in studies. The presence of relevant discrepancies in the relationship between the grouping variables and students' ability estimates across the two dimensions would represent a significant indication of the distinctiveness of the secondary CR dimension compared to the primary math ability, further supporting the appropriateness for the examined test of the proposed bidimensional model.

As a third and final aim, we investigated the impact of format-related multidimensionality on the adequacy of the unidimensional model in the categorization of students in different proficiency levels. More specifically, we examined the degree of concordance between two proficiency classifications as based respectively on the ability estimates distributions for the unidimensional model and the main (MC + CR) dimension of the bidimensional model. The existence of a low degree of concordance across classifications would suggest the need to revise the test scoring procedure in order to account for local dependencies on the responses to CR items.

METHOD

Participants and Procedure

Data used in this study concern a sample of the 8th grade student population which undertook the INVALSI standardized tests in the year 2012 in their basic unedited versions and with no extra time added for test completion (N = 519003). The INVALSI tests are norm-referenced assessments designed to assess Italian 8th grade students' proficiency in mathematics and reading. We obtained source data by filling in an online request through the INVALSI institutional website. Due to sample-size limitation of the employed version of the ConQuest software, analyses were performed on a random sample of 3000 examinees obtained by stratifying the student population on gender (49.6% female), citizenship status (90.4% Italian), regularity in studies (89.2% regular students), and macro-area of residence (44.1% Northern Italy area; 18.9% Central Italy area; 37.0% Southern Italy area). For the purpose of this study, responses from students requiring special accommodations for the test (e.g., visually impaired students, student with intellectual or specific learning disabilities) were not included in the analyses.

Instruments

In this study, we analyzed the mathematical component of the INVALSI 2012 test, which is a mixed-format test assessing students' proficiency on four different mathematical content do-

mains: (a) numbers, (b) relationships and functions, (c) space and figures, (d) measurement and predictions (INVALSI, 2012). Each item assesses student knowledge on a single math domain. The test consists of 38 items[1] (21 MC items and 17 CR items). The proportion of the four math content domains is balanced across the MC and CR subsets of the test. Two types of multiple-choice items are included: 20 simple MC items (one correct option, three distracters) and one multiple true-false (MTF) MC item (organized as bundles of four true-false items sharing a common stimulus). Two types of CR items are included: 15 short-answer CR items (items that require a short response, usually in numerical or graphical form) and two brief CR items (items that require the student to provide both the solution to a problem and to explain/illustrate the logic behind the provided solution). The INVALSI technical report about the test documents the use of the Rasch model for the analysis of responses. However, an in-depth examination of validity is missing. In particular, INVALSI does not provide fit statistics for the obtained Rasch measures. In addition, INVALSI does not document analyses examining the dimensionality of the test. An estimate of test reliability was provided by computing Cronbach's alpha ($\alpha = .84$).

For the purpose of this study, all items allowing partial credit scoring (i.e., MTF-MC and brief CR items) were dichotomized prior to the analyses. As done by INVALSI, in order to control for potential violations of local independence of items sharing a common stimulus, the MTF-MC item is scored 1 for a number of correct responses ≥ 3, otherwise score is 0.

Data Analysis

Dimensionality Analyses

In this study, two alternative models of students' responses on the test were implemented and compared. Analyses were performed using the ConQuest 3.0 software (Adams, Wu, Haldane, & Sun, 2012). By letting MC and CR items load on the same dimension, the unidimensional Rasch model was firstly implemented on examinees' response data to the test. As a second step, a multidimensional random coefficient multinomial logit model (MRCMLM; Adams, Wilson, & Wang, 1997) characterized by within-item dimensionality for the CR items was implemented.

The MRCMLM is a Rasch model belonging to the broader family of multidimensional IRT (MIRT) models. In MIRT models characterized by within-item dimensionality (as opposed to between-item dimensionality; Adams et al., 1997), the probability of a correct response to an item is not conditional on a single ability variable θ, but on a vector θ of k latent ability dimensions. In MIRT modeling, the relationship between multiple abilities in predicting the probability of correct response to an item can be either additive or multiplicative (Hartig & Höhler, 2009). When the abilities necessary to answer an item are additive, MIRT models are called compensatory: this means that a low ability in one dimension can be compensated by a high ability in a second dimension, and vice versa. When the relationship is multiplicative the models are usually defined as non-compensatory, meaning that the probability of success will only approach one if all abilities required for a particular item are high. Given a dichotomous response to item i, then, the following formulation represents the item response function (IRF) for a MIRT compensatory model:

$$P\left(X_i = 1 | \theta\right) = \frac{1}{e^{-\left(\lambda_1' \theta - \delta_i\right)}}$$

where λ_i is a $k \times 1$ vector of factor loadings and δ_i is the location parameter for item i (i.e., the item difficulty). If λ_i contains more than one nonzero loading, a low ability in one dimension required for an item i can be compensated by a high value in a second dimension, and vice versa. In multidimensional Rasch models, such as the model implemented in the present study, all nonzero elements of vector λ_i are fixed to one.

In this study, we implemented a bidimensional compensatory model by letting MC and CR items load on a common main dimension and by letting CR items additionally load on an auxiliary dimension nonorthogonal to the main dimension. Figure 1 shows the diagram for the bidimensional model. The fit to the data of the unidimensional and bidimensional models was compared by using the likelihood-ratio G^2 statistics test (Briggs & Wilson, 2003) and by examining both the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) information-based fit indices.

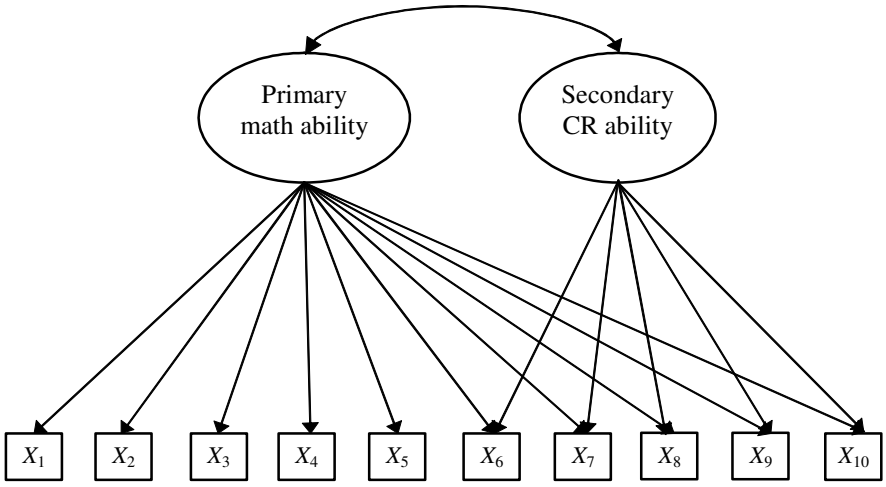

FIGURE 1
Diagram for the bidimensional within-item MIRT model for math ability.

In the estimation of both models, the mean of the item difficulty parameters was fixed to 0 for the purpose of model identification. Descriptive statistics for the item difficulty parameters and fit statistics — that is, unweigthed (outfit) and weighted (infit) mean-square residuals statistics (Wright & Masters, 1982; for cut-off values, see also Wright, Linacre, Gustafson, & Martin-Lof, 1994) — and the person ability expected-a-posteriori (EAP; Bock & Mislevy, 1982) estimates for both the models are documented. The latent structure underlying the bidimensional model was further investigated by examining the correlation between the primary math ability dimension and the secondary CR dimension. Finally, in order to evaluate the degree of consistency of the estimated person and item parameters across the unidimensional model and the bidimensional model, correlations between the obtained measures were also examined.

As a preliminary step to the dimensionality analyses, we performed two separate unidimensional Rasch analyses on the MC and CR sections of the test. Reliability for the MC and CR sections was respectively .68 and .73. The correlation between the EAP person ability estimates

for the two sections was moderate (ρ = .62), suggesting the appropriateness of a more in-depth examination of test dimensionality, in spite of the item fit statistics indicating substantial unidimensionality when calibrating the full test to the Rasch model (see Table 2).

Group Ability Differences

Two multiple regression models were implemented to test the role of the examinees' gender, citizenship status, and regularity in studies as predictors of the ability on the main (MC + CR) and secondary (CR) ability EAP estimates as obtained by fitting the bidimensional model. The variables were tested in the models by using the following dummy coding structure: gender, citizenship status, and regularity in studies were coded as 0/1 respectively for male/female, repeating/regular student, and non-Italian/Italian. Consistently with the literature indicating differential performances on CR items when controlling for gender (DeMars, 1998, 2000; Taylor & Lee, 2012), native/non-native speaker status (Ilich, 2013), and different levels of proficiency (Wang, 2002), our specific hypothesis is that different patterns of ability differences could be found across the two dimensions when controlling for the selected grouping variables.

Classification Analyses

In order to examine the agreement between the unidimensional and the bidimensional models in discriminating between different levels of math proficiency, a tentative procedure for the classification of students in proficiency levels is implemented in this study. More specifically, two 4-level classifications of students were obtained on the standardized EAP ability scores for the unidimensional model and the main dimension of the bidimensional model. The following ability cut-points were used: (1) Poor: ability < -1 *SD*; (2) Low achieving: -1 *SD* $\leq$ ability < 0; (3) Proficient: $0 \leq$ ability $< +1.00$ *SD*; (4) Highly proficient: $+1$ *SD* $\leq$ ability. The degree of agreement between the two classifications was then evaluated using Cohen's kappa coefficient.

Even though no explicit criteria for the classification of examinees in proficiency levels are provided in the official documentation for the test, in the annual INVALSI report students' ability distributions are compared across regional areas by taking four specific percentile cut-offs as a reference — that is, the 5th, 25th, 75th, and 95th percentiles. As a means to further investigate across-models agreement in the classification of students, two 5-level classifications were obtained and compared by using these cut-off points on the EAP ability measures for the unidimensional model and the main dimension of the bidimensional model.

RESULTS

Dimensionality Analyses

The model fit statistics for the unidimensional model and bidimensional model are presented in Table 1. For these two models, the difference in the deviances resulted in a χ^2 of 273.95 with three degrees of freedom; this is statistically significant at the .01 level indicating that by implementing a bidimensional compensatory model on the data a significant but small increase in

fit over the unidimensional model is obtained. Inspection of the AIC and BIC confirmed this interpretation. Table 2 reports the descriptive statistics for item difficulty parameters and item fit statistics as estimated implementing both models. As can be seen, fit statistics are in the range of 0.7-1.3 (Wright et al., 1994), indicating a good fit of both models. No relevant differences in estimated difficulty and fit to the model were observed across the models. Moreover, under both models no significant differences in mean difficulty were observed when comparing the MC and CR sections of the test — unidimensional model: $F(1, 35) = 0.16$, $p = .69$; bidimensional model: $F(1, 35) = 0.33$, $p = .57$. Correlation between item difficulty parameters as estimated in the two models was + .98 ($p < .01$), indicating a strong consistency across the models.

TABLE 1
Model fit statistics for the unidimensional and bidimensional model

Model	Unidimensional model	Bidimensional model
–2*log likelihood	132265.53	131991.58
Parameters	38	41
G^2 LR Test	$\chi^2(3) = 273.95$, $p < .01$	
AIC	132341.53	132073.58
BIC	132569.77	132319.84

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion.

TABLE 2
Descriptives: Item difficulty estimates and fit statistics under the unidimensional and the bidimensional model

	Unidimensional model			Bidimensional model		
	Difficulty	Outfit	Infit	Difficulty	Outfit	Infit
M	0.00	0.99	1.00	0.00	1.03	1.02
Min	–1.74	0.87	0.90	–1.66	0.91	0.92
Max	1.49	1.17	1.10	1.62	1.12	1.09
SD	0.86	0.07	0.05	0.87	0.04	0.03

Table 3 shows the descriptive and reliability statistics for students' EAP ability estimates as computed by ConQuest. The average ability as estimated using the unidimensional model was +.08 logit ($SD = 0.73$); under the bidimensional model, the average ability on the main dimension was –.05 logit ($SD = 0.66$), while for the secondary CR dimension was +.29 logit ($SD = 0.28$). Compared to the variability in the dimension for the unidimensional model ($\sigma^2 = .63$), a minor change in variability is observed on the primary math dimension after the inclusion in the model of the additional dimension for the CR items ($\sigma^2 = .55$). However, compared to the latent dimension as modeled in the unidimensional model and with primary dimension of the bidimensional model, the variability observed in the auxiliary CR dimension was very low ($\sigma^2 = .22$), indicating

the additional local dependency across CR items to be limited. Examination of the relationship between the latent ability traits modeled with the bidimensional model revealed the CR auxiliary dimension to be positively yet weakly correlated with the primary math ability dimension (+.22, at $p < .01$). The correlation between the EAP ability estimates for the two dimensions was significantly larger ($\rho = +.60$, at $p < .01$). Finally, the correlation between the EAP ability measures for the unidimensional model and the primary dimension of the bidimensional model indicated almost complete linearity between the measures ($\rho = +.99$, $p < .01$).

TABLE 3
Descriptives: EAP ability estimates under the unidimensional and the bidimensional model

	Min	Max	*M*	*SD*	Reliability
Unidimensional model	–1.94	2.19	.08	0.73	.83
Bidimensional model: Main dimension (MC + CR)	–1.92	1.84	–.05	0.66	.78
Bidimensional model: Secondary dimension (CR)	–0.58	1.22	.29	0.28	.37

Group Ability Differences

Table 4 reports the result of the two regression models implemented to test the relationship between students' gender, citizenship status, and regularity in studies and their ability as estimated on the primary math dimension and the secondary CR dimension of the bidimensional model. For both the models, the observed explanatory power was very limited (Adjusted $R^2 \leq .04$). Still, significant but small effects were found for all the predictors for both the models. More in detail, when controlling for gender, being female was found to predict lower ability estimates on both the dimensions, while being Italian and having a regular study track were found to predict higher ability estimates on both the dimensions. Examination of the direction and size of such effects revealed a similar pattern across the dimensions.

TABLE 4
Regression models: EAP ability estimates (bidimensional model) on gender, regularity in study, and citizenship status (dummy coding)

	Main (MC + CR)			Secondary (CR)		
	B	*SE*	β	B	*SE*	β
Intercept	–0.37*	0.04	–	+0.15*	0.02	–
Gender (1 = Female, 0 = Male)	–0.13*	0.02	–0.10	–0.04*	0.01	–0.08
Regularity in studies (1 = Regular, 0 = Repeating)	+0.20*	0.04	+0.10	+0.10*	0.02	+0.08
Citizenship status (1 = Italian, 0 = Non-Italian)	+0.23*	0.04	+0.10	+0.08*	0.02	+0.11
Adjusted R^2 (Cohen's f^2)		.04 (.04)			.03 (.04)	

*$p < .01$.

Classification Analyses

The proficiency classifications obtained on the standardized scores as computed on the ability measures for the unidimensional model and the primary dimension of the bidimensional model are compared in Table 5. The kappa statistic for the two classifications was .91 ($p < .01$), revealing a strong degree of agreement between the classifications. Consistently, the number of examinees for which we found a discrepancy in classification across the models accounted for only 6% of the sample: specifically, 4.3% of the examinees were categorized in a lower proficiency level under the unidimensional model, while for 1.7% of them the opposite pattern was observed. Comparable results (not reported in tables) emerged when comparing the 5-level classifications obtained using the 5th, 25th, 75th, and 95th percentiles as cut-off points on the ability measures. Both the kappa statistic ($k = .91$, $p < .01$) and the existence of a very limited number of misclassifications (4.37%) further indicated a strong agreement across the compared models in the classification of students in different proficiency levels.

TABLE 5
Proficiency level classifications under the unidimensional model versus the first dimension of the bidimensional model based on the standardized distributions of the ability estimates ($N = 3000$)

		Classification based on bidimensional model			
		Level 1	Level 2	Level 3	Level 4
Classification based on unidimensional model	Level 1	457 (15.2%)	37 (1.2%)	0 (0.0%)	0 (0.0%)
	Level 2	18 (0.6%)	1068 (35.6%)	63 (2.1%)	0 (0.0%)
	Level 3	0 (0.0%)	18 (0.6%)	813 (27.1%)	31 (1.0%)
	Level 4	0 (0.0%)	0 (0.0%)	15 (0.5%)	480 (16.0%)

DISCUSSION

The main aim of the present study was to investigate the existence of format-related multidimensionality in students' responses to a mixed-format standardized math test administered in Italy as part of the statewide INVALSI assessment program. Specifically, we hypothesized that modeling students' responses to the test by distinguishing between a primary math ability dimension related to both the MC and CR sections of the test and a secondary compensatory dimension accounting for the presence of additional local dependencies across CR items would be more appropriate for the examined test than a unidimensional modeling approach. In this study, however, the existence of a secondary CR ability dimension was not fully supported by the results. An increase in fit over the unidimensional model was observed when implementing the bidimensional model on the data; under this model, a secondary CR dimension was found to be positively, yet

weakly correlated with the primary math ability dimension. However, when compared to other studies using the same modeling approach for format-related multidimensionality (Rauch & Hartig, 2010; Walker & Beretvas, 2003), the relatively low variability observed in the secondary CR dimension suggests the violation of unidimensionality due to the inclusion of CR items to be negligible and ultimately does not justify the use of a multidimensional model for the current test. The results of the classification analyses also support this interpretation. Specifically, we found a high degree of concordance when comparing the proficiency classifications of students as obtained on the ability estimates for the unidimensional model and the primary dimension of the bidimensional model, resulting in only a very limited number of inconsistencies across the compared classifications. This finding indicates that the distortion introduced in the unidimensional classification by not considering the residual local dependency existent among CR items is very limited; as a result, a revision of the scoring procedure to account for format differences is not required for the examined test. Furthermore, the presence of a similar pattern of ability differences when comparing students by gender, citizenship status, and regularity of studies on the primary math dimension and the secondary CR dimension further suggested a general lack of distinctiveness in the performance required by the CR items when compared to MC section of the examined test. Given the specific characteristics of the CR items included in the examined test, our results appear to be in line with findings reporting overall construct equivalence between the abilities assessed by CR and MC items when CR items are mainly designed as short-answer items, as opposed to open-ended questions (Bacon, 2003; Bible, Simkin & Kuechler, 2008; Rodriguez, 2003).

Still, this study has several limitations. Due to the specific characteristics related to the data collection design implemented by INVALSI for the 8th grade education level — that is, the lack of a student questionnaire like those administered in the context of the PISA and TIMSS assessment programs — relevant individual characteristics (e.g., students' motivation, previous academic performances in mathematics) were not considered in the present study. For this reason, we could not examine further hypotheses concerning the nature of the construct underlying the secondary CR dimension. Moreover, even though the MC and CR sections of the examined test were comparable in terms of difficulty and investigated math content proportion, the existence of across-format differences concerning the specific cognitive skills elicited by the items was not controlled. Further research is needed to study (e.g., by implementing experimental designs) how the interaction between different cognitive requirements and item formats affects test dimensionality.

NOTE

1. Preliminary analyses revealed the scores on one item of the MC subset of the test (i.e., item E1) to be characterized by an extreme score in the extracted sample, that is, all examinees answered the item correctly. This is consistent with the documentation for the test, which indicates the item as extremely easy compared to the rest of the test (INVALSI, 2012). As a result, the item was not included in subsequent analyses, reducing the examined item-pool to 37 items.

ACKNOWLEDGEMENTS

The authors would like to thank the Italian National Institute for the Evaluation of the Education System (INVALSI), and in particular Dr. Roberto Ricci and Dr. Patrizia Falzetti, for their effective support.

REFERENCES

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23. doi:10.1177/0146621697211001

Adams, R. J., Wu, M. L., Haldane, S., & Sun, X. X. (2012). *ConQuest 3.0: Generalised item response modelling software* [Computer Software]. Camberwell, AU: Australian Council for Educational Research.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716-723. doi:10.1109/TAC.1974.1100705

Arthur, N., & Everaert, P. (2012). Gender and performance in accounting examinations: Exploring the impact of examination format. *Accounting Education*, *21*, 471-487. doi:10.1080/09639284.2011.650447

Bacon, D. R. (2003). Assessing learning outcomes: A comparison of multiple-choice and short-answer questions in a marketing context. *Journal of Marketing Education*, *25*, 31-36. doi:10.1177/0273475302250570

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement*, *28*, 23-35. doi:10.1111/j.1745-3984.1991.tb00341.x

Bennett, R. E., Rock, D. A., & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement*, *28*, 77-92. doi:10.1111/j.1745-3984.1991.tb00345.x

Bible, L., Simkin, M. G., & Kuechler, W. L. (2008). Using multiple-choice tests to evaluate students' understanding of accounting. *Accounting Education: An International Journal*, *17*, S55-S68.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York, NY: David McKay.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444. doi:10.1177/014662168200600405

Boodoo, G. M. (1993). Performance assessments or multiple choice? *Educational Horizons*, *72*, 50-56.

Bridgeman, B. (1991). Essays and multiple-choice tests as predictors of college freshman GPA. *Research in Higher Education*, *32*, 319-332. doi:10.1007/BF00992895

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement*, *4*, 87-100.

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education*, *11*, 279-299. doi:10.1207/s15324818ame1103_4

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*, 55-77. doi:10.1207/s15324818ame1301_3

Ercikan, K., Sehwarz, R. D., Julian, M. W., Burket, G. R., Weber, M. M., & Link, V. (1998). Calibration and scoring of tests with multiple-choice and constructed-response item types. *Journal of Educational Measurement*, *35*, 137-154. doi:10.1111/j.1745-3984.1998.tb00531.x

Ghorpade, J., & Lackritz, J. R. (1998). Equal opportunity in the classroom: Test construction in a diversity-sensitive environment. *Journal of Management Education*, *22*, 452-471. doi:10.1177/105256299802200402

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education*, *62*, 143-157. doi:10.1080/00220973.1994.9943836

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*, 57-63. doi:10.1016/j.stueduc.2009.10.002

Ilich, M. O. (2013). Differential Item Functioning (DIF) among Spanish-Speaking English Language Learners (ELLs) in State Science Tests (Doctoral dissertation). Retrieved from https://www.digital.lib.washington.edu/dspace/bitstream/handle/1773/23633/Ilich_washington_0250E_12054.pdf?sequence=1

INVALSI (National Institute for the Evaluation of the Education System). (2012). *Rilevazioni nazionali sugli apprendimenti 2011-2012. Rapporto tecnico*. [National learning assessments 2011-2012. Technical Report]. Retrieved from http://www.invalsi.it/snv2012/documenti/Rapporti/Rapporto_rilevazione_apprendimenti_2012.pdf

Kuechler, W. L., & Simkin, M. G. (2010). Why is performance on multiple-choice tests and constructed-response tests not more closely related? Theory and an empirical test. *Decision Sciences Journal of Innovative Education*, *8*, 55-73. doi:10.1111/j.1540-4609.2009.00243.x

Lissitz, R. W., Hou, X., & Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact. *Journal of Applied Testing Technology*, *13*, 1-52.

Manhart, J. J. (1996, April). *Factor analytic methods for determining whether multiple-choice and constructed-response tests measure the same construct*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*, 207-218. doi:10.1207/s15326985ep3404_2

Miceli, R., Marengo, D., Molinengo, G., & Settanni, M. (2015). Emerging problems and IRT-based operational solutions in large-scale programs of student assessment: The Italian case. *TPM – Testing, Psychometrics, Methodology in Applied Psychology*, *22*, 53-70. doi:10.4473/TPM22.1.5

Perkhounkova, Y., & Dunbar, S. B. (1999, April). *Influences of item content and format on the dimensionality of tests combining multiple-choice and open-response items: An application of the Poly-DIMTEST procedure*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Rauch, D., & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*, 354-379.

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, *40*, 163-184. doi:10.1111/j.1745-3984.2003.tb01102.x

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461-464. doi:10.1214/aos/1176344136

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education*, *3*, 73-98. doi:10.1111/j.1540-4609.2005.00053.x

Taylor, C. S., & Lee, Y. (2012). Gender DIF in reading and mathematics tests with mixed item formats. *Applied Measurement in Education*, *25*, 246-280. doi:10.1080/08957347.2012.687650

Thissen, D., Wainer, H., & Wang, X. B. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement*, *31*, 113-123. doi:10.1111/j.1745-3984.1994.tb00437.x

Wainer, H., & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a Marxist theory of test construction. *Applied Measurement in Education*, *6*, 103-118. doi:10.1207/s15324818ame0602_1

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of Educational Measurement*, *38*, 147-163. doi:10.1111/j.1745-3984.2001.tb01120.x

Walker, C. M., & Beretvas, S. N. (2003). Comparing multidimensional and unidimensional proficiency classifications: Multidimensional IRT as a diagnostic aid. *Journal of Educational Measurement*, *40*, 255-275. doi:10.1111/j.1745-3984.2003.tb01107.x

Walstad, W., & Robinson, D. (1997). Differential item functioning and male-female differences on multiple choice tests in economics. *Journal of Economic Education*, *28*, 155-171. doi:10.1080/00220489709595917

Wang, Z. (2002). *Comparison of different item types in terms of latent trait in mathematics assessment* (Doctoral dissertation). Retrieved from https://circle.ubc.ca/handle/2429/12796

Wilson, L. D., & Zhang, L. (1998, April). *A cognitive analysis of gender differences on constructed-response and multiple-choice assessments in mathematics*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA, USA.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago, IL: MESA Press.

Zimmerman, D. W., & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement*, *27*, 357-371. doi:10.1177/0146621603254799