Information value of Multiple Response questions

Srđan Verbić

Institute for Education Quality and Evaluation, Belgrade, Serbia

Answers to Multiple Response (MR) questions carry more information than we usually utilize. Simple idea that all options of MR questions should be scored as independent test items has two major difficulties: 1) false options have item-response characteristics that are hard to model and use with other items; and 2) responses to individual options within the same MR question could be too dependent on each other. These difficulties lead to an overestimation of item discrimination and test information function. A few scoring methods that could increase information value obtained from MR questions are proposed and evaluated in this paper.

Keywords: multiple response questions, item information, test scoring, item response theory

In formative tests, like tests for standards-setting or diagnostic tests, it is more important to find out what exactly examinees do or do not know than how much they know. For this purpose, because of interpretation simplicity, test makers usually utilize only two item types: Multiple Choice (MC) and Constructed Response (CR) questions. However, there is a need for other question types, especially in computer-based testing, that would bridge the gap between completely closed and completely open questions. Multiple Response (MR) questions have been examined as a promising candidate in a few recent studies (Eggen & Lampe, 2011; Hohensinn & Kubinger; Jiao, Liu, Haynie, Woo, & Gorham, 2012).

MULTIPLE RESPONSE QUESTIONS

Multiple Response (MR) questions are a type of selected response questions. They are similar to Multiple Choice (MC) questions with the exception that more than one of the provided options are (or may be) true (Ebel, 1951; Kurz, 1999; Parshall, Davey, & Pashley, 2000; Wesman, 1971). MR questions are also known as selected response items (Sireci & Zenisky, 2006), multiple-mark items (Pomplun & Omar, 1997; Scalise & Gifford, 2006), or multiple multiple-choice (Cronbach, 1941) and multiple answer (Dressel & Schmid,

Corresponding author: sverbic@ceo.edu.rs
1953) in older literature. Today, MR questions represent a common question type in all computer-based testing software. MR questions sometimes request exact number of selections (Bauer, Holzer, Kopp, & Fischer, 2011; Eggen & Lampe, 2011), but more common is unconstrained variant where examinees should select all true options (Parshall, Stewart, & Ritter, 1996).

MR question consists of a stem, an instruction, and a few true or false options. Examinees respond correctly to MR question if they select true options and leave false options unmarked. MR question may be treated as a set of true-false options where examinee is supposed to select only true options while all unselected options are considered to be marked as false. This item type is very similar to Multiple True-False (MTF) questions. In MTF questions, examinees respond to each of separate true-false statements by selecting either true or false option. MTF questions, just like MR questions, have common stem and instructions for all statements. Sometimes test-makers identify MR as MTF questions and treat them in the same manner (Hsu, Moss, & Khampalikit, 1984; Tsai & Suen, 1993). Main difference between these two item types is that omitting to mark false option does not imply that examinee considers the option as false. We present these two item types in Table 1.

Table 1. Examples of a MR and the corresponding MTF question

<table>
<thead>
<tr>
<th>Question type</th>
<th>Multiple Response question</th>
<th>Multiple True-False question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stem</td>
<td>Which elements make water?</td>
<td>Which elements make water?</td>
</tr>
<tr>
<td>Instruction</td>
<td>Select all true options.</td>
<td>Select either true or false for each option.</td>
</tr>
<tr>
<td>Options</td>
<td></td>
<td>True</td>
</tr>
<tr>
<td>Hydrogen</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Helium</td>
<td>□</td>
<td></td>
</tr>
<tr>
<td>Oxygen</td>
<td>✗</td>
<td></td>
</tr>
<tr>
<td>Nitrogen</td>
<td>□</td>
<td></td>
</tr>
</tbody>
</table>

In spite of availability and notable educational measurement potentials of MR questions, psychometric characteristics of MR questions have not been systematically investigated yet (Eggen & Lampe, 2011; Kastner & Stangla, 2011; Parshall et al., 2000; Scalise & Gifford, 2006; Tsai & Suen, 1993). Reliability and efficiency of MTF questions were reported in a few studies (Albanese & Sabers, 1988; Dudley, 2006; Emmerich, 1991; Frisbie & Sweeney, 1982), but adequate studies of MR questions are still missing.

SCORING METHODS

A general problem with Multiple Response, as well as with Multiple True-False questions, is how to score them. Reader should note that these two types of questions have the same response format. Therefore, scoring methods applied for one of these types can be automatically applied for the other type as well.
The most common scoring method is “all or nothing”. This method awards full credit (one point) to an examinee only if all options of that question are responded correctly. Otherwise, the examinee does not receive any credit (zero points). This method of scoring is labeled as the cluster scoring by Frisbie & Druva (1986). It is also known as multiple response scoring (Albanese, Kent, & Whitney, 1979) or rigid scoring (McCabe & Barrett, 2003). The cluster scoring method is based on the assumption that reliability of a test is increased if the probability of chance success is reduced (Frary & Zimmerman, 1970). A weakness of the cluster scoring is that useful information about responses to any particular option is ignored. That way the whole set of responses is treated binary.

Cluster scoring method does not have to be applied only to all options, but also to various combinations of options. Full credit, for instance, can be given to examinees that had four or more correctly responded options, to those who responded correctly to all true options, etc. That way we can get a set of dichotomous scoring methods, which can be more reliable than “all or nothing”. Studies concerning such scoring methods are rare, but they clearly demonstrate that the most rigid “all or nothing” variant is the least reliable one (Tsai & Suen, 1993).

Simple alternative to the cluster scoring is the item scoring method where an examinee is awarded with one point for each correctly responded option. If we score response data this way, it is as if we increased number of items in the same test, which then increases reliability of the results (Dudley, 2006; Frisbie & Sweeney, 1982; Kreiter & Frisbie, 1989).

Previous studies on item scoring demonstrated that two main disadvantages of this method are local dependency of the options and high level of guessing in responding to individual options. The lack of local independency required by both classical and item response testing theories, can be a major source of bias in determination of the item parameters. Positive local item dependence (LID) increases the strength of the relationship between some items. Therefore, positive LID increases the correlation between any of the items and the total test score. This occurrence, to some extent, violates the assumption of local independency and consequently influences item parameter estimation (Hambleton & Swaminathan, 1985). This effect generates higher estimates of item discrimination for LID items (Masters, 1988) and higher reliability estimates.

The second disadvantage of MR or MTF questions is high level of guessing to their binary options. For MTF options, Emmerich (1991) had shown that examinees more frequently answer “true” than “false”. In the case of MR questions, options can also be either true or false. Correct response would be to mark an option if it is true or to leave the option unmarked if it is false. The probabilities of these two ways of responding correctly are not the same. It was reported that low-achieving student often leave both, true and false, options unmarked (Pomplun & Omar, 1997). If this happens, we can get false impression that student responded correctly to all false but not to the true options. This occurrence should warn us that false options might not be as useful and informative as true options in MR questions are.
Comparison of reliability or test information functions for different scoring methods should reveal conditions for optimal scoring of MR questions.

INFORMATION VALUE

The statistical meaning of information is credited to Ronald A. Fisher, who defined information as the reciprocal value of the precision with which a parameter could be estimated (Baker, 1985). If we take estimate’s variance as the measure of precision, information can be easily defined for both Classical Test Theory (CTT) and Item Response Theory (IRT). The main difference between the information that we obtain using IRT and CTT is that estimate’s variance can be associated with every item and all ability levels within IRT while the estimate’s variance is a global characteristic of the entire test within CTT.

In practice, we usually compare reliability of tests or scoring methods in CTT or their test information functions in IRT analysis. Tests that are more informative have greater reliability as well as greater values of test information function. Both, reliability and information function, are simply related to estimate’s standard error (SE).

In CTT, we estimate examinee’s true score using different scoring methods. Standard error of that estimate can be calculated from reliability obtained for that scoring method, represented by Cronbach’s alpha (α), and the standard deviation (σ) of examinees’ scores (Embretson & Reise, 2000):

\[ SE = \sigma \sqrt{1 - \alpha}, \]  

(1)

In CTT, obviously, estimate’s standard error does not depend on the examinee’s score. Nevertheless, in IRT, standard error is not uniform across the entire range of ability parameter (θ). Instead of a single number in CTT, information in IRT becomes a function of ability parameter. That way, IRT advances the concept of item and test information and upgrades classical concept of reliability. Using the appropriate formula I(θ), the information can be calculated for each ability level on an ability scale. Corresponding standard error of the ability estimate is given by:

\[ SE(\theta) = \frac{1}{\sqrt{I(\theta)}}, \]  

(2)

where I(θ) represents Fisher’s test information function at the ability level θ. Since a test is a set of items, the test information value at a given ability level is simply the sum of the item information values at that level (Baker, 1985). Hence, the test information function can be calculated as:

\[ I(\theta) = \sum_{i=1}^{n} I_i(\theta), \]  

(3)

where \( I_i(\theta) \) is the amount of information for item i at the ability level θ, n is the number of items in the test. The test information function is a particularly
useful feature of Item Response Theory. It tells us how well a test measures examinees’ abilities over the whole range of ability levels.

Calculation of Fisher information function depends on the chosen IRT model. The most general IRT model has three parameters (3PL model): discrimination $a$, difficulty $b$, and pseudo-guessing $c$. Probability that an examinee with the ability $\theta$ would give a correct answer to item $i$ with parameters $(a, b, c)$ is given by formula:

$$P_i(\theta) = c_i + (1 - c_i) \frac{1}{1 + e^{-a_i(\theta-b)}},$$  \hspace{1cm} (4)

Finally, the item information function for the item $i$ is given by:

$$I_i(\theta) = \frac{P_i'(\theta)^2}{P_i(\theta)Q_i(\theta)},$$  \hspace{1cm} (5)

where $P_i'(\theta)$ is the first derivative of $P_i(\theta)$ with respect to $\theta$ and

$$Q_i(\theta) = 1 - P_i(\theta).$$  \hspace{1cm} (6)

RESEARCH QUESTIONS

Current national assessment studies rely on large-scale knowledge testing. That practice – started in Serbia with international studies PISA and TIMSS – imposes new challenges concerning the selection of item types and scoring methods, as well as their information properties. Good estimation of items’ psychometric properties early in test production process is essential for test writers and test constructors. This analysis attempts to evaluate characteristics of various scoring methods for Multiple Response questions. The research questions we are trying to evaluate are:

1) Is there a significant difference in MR questions discrimination, calculated as item-total correlations, if we apply different scoring methods?

2) Is there a significant difference in test reliability for different scoring methods?

3) Is there a difference in item characteristics of true and false options in MR questions?

4) Do we increase the value of test information function across the ability scale by using item instead of cluster scoring?

5) Can we estimate effects of inter-item dependency on test reliability or test information function?

Methods

Results shown in this study were obtained as a secondary analysis of an on-line pilot-test for 4th-grade pupils within Nature & Society course (test PD09). The study was conducted at the Institute for Education Quality and Evaluation in 2009. Main results of this on-line test
were reported in (Verbić, Tomić, & Kartal, 2010). The purpose of this study was to explore possibilities of on-line pretesting pupils’ knowledge in the subject and to help establish a framework for an annual national-level formative test. Pupils were asked to answer everything they know in order to assess their own knowledge and help develop new, better tests. They were informed that no penalty will be applied for wrong answers and that they will receive no grades for their test results.

Sample. The sample was stratified – primary schools were the first stratum and students were the second. The school sample was designed to be convenient since not all schools had facilities to participate in on-line testing. At the students’ level, the sample was created according to willingness of students to participate and number of available computers in the school. In total, 926 students from 50 schools participated in the testing.

Instruments. The test consisted of 32 questions: 29 Multiple Choice (MC) and 3 Multiple Response (MR) questions (labeled with #4, #16, and #22) with 5 options each (labeled as items #4.1, #4.2, ..., #22.5). Among 15 MR options in the test, 8 options were true and 7 false. The questions assessed school knowledge about nature and society acquired during the first four years of primary school in Serbia. Difficulty of the questions varied from very easy to moderately difficult.

The software environment for the on-line test delivery was Moodle, open-source software for producing internet-based courses (Dougiamas, 2001) with a testing module adjusted for a large-scale assessment (Verbić & Tomić, 2009). MC and MR questions were presented visually in such a way that examinees could easily tell apart questions with one and more than one correct options.

Scoring methods. Test and questions’ properties were analyzed using several scoring methods divided in two categories: cluster and item scoring. In the first category, there are four dichotomous and two polytomous scoring methods where responses to all options of a MR question are represented by a single number:

<table>
<thead>
<tr>
<th>Method</th>
<th>Scoring key</th>
</tr>
</thead>
<tbody>
<tr>
<td>all 5</td>
<td>Score is 1 only if responses to all 5 options are correct. Otherwise, 0 points.</td>
</tr>
<tr>
<td>4+</td>
<td>Score is 1 if examinee correctly answered at least 4 options. Otherwise, 0 points.</td>
</tr>
<tr>
<td>3+</td>
<td>Score is 1 if examinee correctly answered at least 3 options. Otherwise, 0 points.</td>
</tr>
<tr>
<td>all T</td>
<td>Score is 1 only if responses to all true options are correct. Otherwise, 0 points.</td>
</tr>
<tr>
<td>mean</td>
<td>Score is number of correctly responded options divided by the number of options.</td>
</tr>
<tr>
<td>mean T</td>
<td>Score is number of correctly responded true options divided by the number of true options.</td>
</tr>
</tbody>
</table>

Since MC questions were always scored in the same way, dichotomously like independent items, within the cluster scoring, we always had 29 MC and 3 MR, which is 32 scoring items.

For the item scoring, all options were treated as independent dichotomous items, where the following two methods were applied:

<table>
<thead>
<tr>
<th>Method</th>
<th>Scoring key</th>
</tr>
</thead>
<tbody>
<tr>
<td>item</td>
<td>All options are items scored 1 for correct and for 0 incorrect responses.</td>
</tr>
<tr>
<td>item T</td>
<td>All true options are items scored 1 or 0. False options are omitted.</td>
</tr>
</tbody>
</table>

Here MR questions were treated as sets of items. Dataset for “item” scoring had 44 scoring items, while for “item T” scoring dataset had 37 aggregated items.
Properties of all scoring methods can be analyzed using Classical Test Theory (CTT). Dichotomous scores can also be analyzed using Item Response Theory (IRT). In order to see the difference between guessing characteristics of MR options, we have chosen three-parameter variant of IRT model (3PL IRT) without prior distribution for $c$-parameter (Partchev, 2008; Zimowski, Muraki, Mislevy, & Bock, 1998) and applied it to all scoring methods.

Data Analysis. Effects of different scoring methods on discrimination and reliability were analyzed using bootstrapping method (Efron & Tibshirani, 1986). List of examinees were resampled 100 times in order to estimate the standard error for all parameters for the given scoring method. Since parameters’ estimates are based on the same set of bootstrap samples for all scoring methods, estimations can be treated as paired. Therefore, differences between discrimination indices for all enlisted scoring methods and the default “all or nothing” scoring method were calculated for all questions. Their statistical significance is tested using series of paired $t$-tests.

The response data were analyzed using R: a language for statistical computing (R Development Core Team, 2007) with ltm package for Latent Variable Modelling and Item Response Theory Analyses (Rizopoulos, 2006), irtoys interface package (Partchev, 2008), bootstrap package (Canty & Ripley, 2009), and BILOG routines (Zimowski et al., 1998).

Results

Multiple Response (MR) and Multiple Choice (MC) questions in the test PD09 have been examined concurrently using several basic scoring methods in Classical Test Theory (CTT) and Item Response Theory (IRT).

Difference in question discrimination for different scoring methods

Responses to individual options of MR questions were aggregated in different ways in order to obtain discrimination indices of these questions for each scoring method. Discrimination index of a MR question is calculated as an item-total correlation, where a total score excludes that MR question. In Table 2, we can see item-total correlations for all six cluster-scoring methods and all three MR questions. In order to demonstrate whether the difference in discrimination index between any of these scoring methods and the reference “all 5” method is significant, we have generated 100 bootstrapped response datasets using random samples of examinees. Hence, we obtained simulated distributions for all discrimination indices and compared their mean values using paired $t$-test. The item-total correlation values that are significantly greater than corresponding values for “all 5” cluster-scoring (at level $p<0.01$) are displayed in boldface in Table 2. Alternatively, we have tested effect size of different scoring methods on bootstrapped response data using Friedman test, a non-parametric version of repeated-measure ANOVA, which does not assume normality of distributions. Friedman test revealed that the effect of scoring method on item-total correlation for all three MR questions is significant ($\chi^2(5)>350, p<10^{-15}$). In order to increase rigor of testing because of simultaneous comparisons, we have introduced Wilcoxon signed-rank test, a post-hoc test with Bonferroni correction. The post-hoc test confirmed significance of difference ($p<0.01$) for all pairs whose difference was already labeled as significant according to paired $t$-test.
We observe that both variants of polytomous scoring have significantly better discrimination than “all 5” cluster scoring for all three MR questions. Among dichotomous methods, “mean T” scoring method, which disregards responses to false options, and “4+” appear to be more discriminative than “all 5”. The worst solution for a scoring method in this case would be “3+” method where a guessing probability becomes critical factor. Note that examinee, which does not mark any MR option at all, already has 2, 3, and 2 “correctly responded” options for questions #4, #16, and #22, respectively.

**Difference in test reliability for different scoring methods**

The increasing of question discrimination with different scoring methods should be noticeable throughout the entire test. Greater discrimination of a few questions should result in slightly greater Cronbach’s alpha as a measure of test reliability. The values of Cronbach’s alpha for different scoring methods are shown in Table 3. Friedman test revealed that the effect of various cluster scoring methods on Cronbach’s alpha of PD09 test is significant ($\chi^2(5)=393, p<10^{-15}$). The values that are significantly greater than a corresponding value for “all 5” cluster scoring, according to paired $t$-test, are displayed in boldface. Estimated standard error for all values of item-total correlation is approximately 0.09.

<table>
<thead>
<tr>
<th>Cluster scoring</th>
<th>Item scoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dichotomous</td>
<td>Polytomous</td>
</tr>
<tr>
<td>all 5</td>
<td>0.785</td>
</tr>
<tr>
<td>4+</td>
<td>0.780</td>
</tr>
<tr>
<td>3+</td>
<td>0.790</td>
</tr>
<tr>
<td>all T</td>
<td>0.792</td>
</tr>
<tr>
<td>mean T</td>
<td>0.805</td>
</tr>
</tbody>
</table>

Beside the cluster scoring methods, we have also shown examples of item scoring. If we consider MR options as individual items, Cronbach’s alpha of the test significantly increases, which decreases measurement error for examinees’ total scores. We can see that “mean” variant of cluster scoring method does not produce significantly greater reliability although questions scored in this way have greater discrimination value. Increasing discrimination for three (MR) questions seems to be insufficient to significantly increase the reliability of the entire 32-question test.

**Difference in item characteristics of true and false options**

Scoring methods that disregard false options in MR questions give better question discrimination than methods that equally involve both true and false options. Therefore, we can expect that false options have characteristics that
obscure information contained in MR questions. The main reason for this is the occurrence that low-achieving students often leave options blank rather than to mark them (Pomplun & Omar, 1997). Such a behavior in test PD09 can be observed on Figure 1, where proportion of marked options for all students that participated in the test is given against the total score, calculated using item scoring method. Since the unmarked false options are considered as correctly responded ones, this occurrence destroys our ability to estimate proportion of examinees that deliberately left false options unmarked. Difference in item characteristics between true and false options could be an important issue in IRT analysis where prior estimation of guessing probability plays an important role.

Figure 1: Proportion of marked MR options against the total score for the test PD09

**IRT analysis**

In order to compare information functions for different scoring functions, we have determined IRT parameters as if both IRT assumptions, unidimensionality of the construct and local independence of the responses, are fulfilled. The results of Principal Component Analysis show that the first eigenvalue was five or more times as big as the second eigenvalue, while the second eigenvalue was not distinguishable in size from the other eigenvalues for all cluster-scoring methods for 1PL and 2PL IRT models. Similar analyses for item-scoring methods are not purposeful since the further study demonstrates that we still do not have adequate IRT model for both true and false options of MR questions. Therefore, we paid more attention to the analysis of the second, presumably more critical, IRT assumption – local independence of items.

Estimated values of 3PL IRT model parameters (discrimination $a$, difficulty $b$, and pseudo-guessing $c$) are given in Table 4 for two cluster scoring methods (“all 5” and “all T”), and the item scoring method. The third parameter
in the model (c) indicates how likely it is that examinee will respond correctly by the chance. In the case of cluster-scoring, we can see that pseudo-guessing parameter has values close to zero for all three MR questions and both cluster scoring methods. Estimated difficulty and discrimination parameters for “all T” scoring method that are significantly greater than corresponding parameters for “all 5” scoring method, are printed in boldface.

IRT parameters are estimated for all options of MR questions for item scoring. Estimated pseudo-guessing values for MC questions in PD09 are predominantly between 0.1 and 0.3, while pseudo-guessing parameter for MR options depends heavily on the truth-value of an option. As we can anticipate from results given on Figure 1, guessing factor is much greater for the false than for the true options. BILOG estimates of pseudo-guessing for the true MR options are close to 0 for all options but two, while this parameter is always 0.5 for the false options, except for one (a false option of the question #16, where the BILOG algorithm does not converge at all). Convergence to 0.5 is rather a consequence of the algorithm’s internal constraints (disregards possibility that guessing can be greater than 0.5), than a precise estimation. More detailed analysis of item response characteristics of false options would require dataset with greater number of MR questions.

The presence of one inadequate option (#16.4) causes very different estimates of difficulty parameters (b) for two cluster scoring methods (1.78 for “all 5” and 0.20 for “all T”). Since all other options of question #16 have negative difficulty, obtained difficulty for the whole question, in the case of “all 5” scoring, can be explained only as a consequence of inadequate option presence. On the other hand, “all T” method cannot be affected by the presence of that item since the dubious item is a false option.

Table 4: Test PD09, parameters for 3PL IRT model without prior for pseudo-guessing

<table>
<thead>
<tr>
<th>question</th>
<th>cluster scoring</th>
<th>item scoring</th>
<th>item truth value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all 5</td>
<td>all T</td>
<td>a</td>
</tr>
<tr>
<td>#4</td>
<td>0.75</td>
<td>1.96</td>
<td>0.00</td>
</tr>
<tr>
<td>#16</td>
<td>0.65</td>
<td>1.78</td>
<td>0.00</td>
</tr>
<tr>
<td>#22</td>
<td>0.77</td>
<td>-1.73</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Determining parameters of all 15 MR options, along with all 29 MC questions, enables us to treat them as 44 individual items and use all of them to estimate examinee’s ability. Using item parameters given in Table 4, we can estimate the test information function value across the ability scale (Figure 2). It appears that MR questions (gray areas on the figure) for cluster scoring do not contribute to the test information function more than other (MC) questions. In the case of item scoring, MR options (gray areas divided into sections) appear to contribute much more. We can also see that true options (light gray areas) have greater contribution than false options (dark gray).

Figure 2: Test information function for three scoring methods and 3PL IRT model. Gray areas represent MR questions. Light gray area indicates true options of MR questions, while dark gray represents false options.
Looking at Fisher’s test information function, item scoring of MR questions seems to be superior in comparison to the cluster scoring methods. This impression could be a consequence of inter-item dependency of responses to the options in the same question that cannot be neglected.

Inter-item correlations given on Figure 3 show how often examinees who answer correctly to one option, answer correctly to the other options of the same question. Generally, items in the same test correlate positively since they measure the same construct. The options’ correlations of MR questions are expected to have greater values than correlations between different MC questions (Albanese & Sabers, 1988). For the sake of comparison, distribution of inter-item correlations for MC questions in PD09 test is displayed as a set of lines in the background of Figure 3. While the inter-item correlations for questions #4 and #16, in average, have similar values to those for MC questions, the question #22 has inter-item correlations that go above the values we consider to be normal for this test. In the similar study for MTF questions, Frisbie and Druva (1986) reported inter-item correlations within cluster close to 0.009.

![Figure 3: Inter-item correlations of MR options. Long horizontal lines (in the background) display distribution of correlations between response-vectors to different MC questions.](image-url)

Previous research has demonstrated that local dependency causes an overestimation of discrimination parameter (Tuerlinckx & De Boeck, 2001; Yen, 1993), but we cannot measure precisely the magnitude of this effect. In order to avoid artificial boost of discrimination parameter estimate, we have calculated IRT parameters for each MR option as if it was the only option in that MR question. In other words, during the estimation procedure, we have excluded...
all the options but one from MR question. For the reference, we will call this method the **single-option scoring approximation**. Results of estimation \((a', b',\text{ and } c')\) obtained through the single-option scoring approximation are given in Table 5 and contrasted to results obtained through the item scoring \((a, b,\text{ and } c)\). We have estimated both sets of parameters on the same bootstrap sample and compared those results. Discrimination parameters obtained for the item scoring are always greater than those for the single-option approximation. This difference is not stochastic, but systematic artifact. Parameters obtained through the single-option scoring approximation that differ significantly from the parameters obtained through the item scoring are displayed in boldface.

### Table 5: Estimations of IRT parameters (3PL IRT model) for the item scoring \((a, b, c)\) and the single-option scoring approximation \((a', b', c')\)

<table>
<thead>
<tr>
<th>item</th>
<th>(a)</th>
<th>(a')</th>
<th>(b)</th>
<th>(b')</th>
<th>(c)</th>
<th>(c')</th>
</tr>
</thead>
<tbody>
<tr>
<td>#4.1</td>
<td>0.62</td>
<td>0.53</td>
<td>-0.42</td>
<td>-0.47</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>#4.2</td>
<td>0.61</td>
<td>0.59</td>
<td>1.20</td>
<td>1.11</td>
<td>0.43</td>
<td>0.42</td>
</tr>
<tr>
<td>#4.3</td>
<td>0.78</td>
<td>0.69</td>
<td>0.48</td>
<td>0.33</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>#4.4</td>
<td>0.95</td>
<td>0.84</td>
<td>-2.93</td>
<td>-3.23</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>#4.5</td>
<td>0.96</td>
<td>0.96</td>
<td>-1.23</td>
<td>-1.23</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>#16.1</td>
<td>0.88</td>
<td>0.84</td>
<td>-0.71</td>
<td>-0.68</td>
<td>0.12</td>
<td>0.14</td>
</tr>
<tr>
<td>#16.2</td>
<td>0.82</td>
<td>0.73</td>
<td>-1.10</td>
<td>-1.20</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>#16.3</td>
<td>0.98</td>
<td>0.79</td>
<td>-1.01</td>
<td>-1.18</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>#16.4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>#16.5</td>
<td>0.85</td>
<td>0.78</td>
<td>-0.79</td>
<td>-0.85</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>#22.1</td>
<td>0.57</td>
<td>0.46</td>
<td>-3.70</td>
<td>-4.52</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>#22.2</td>
<td>1.33</td>
<td>1.00</td>
<td>-2.03</td>
<td>-2.66</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>#22.3</td>
<td>0.70</td>
<td>0.56</td>
<td>-2.51</td>
<td>-3.06</td>
<td>0.50</td>
<td>0.50</td>
</tr>
<tr>
<td>#22.4</td>
<td>1.42</td>
<td>1.10</td>
<td>-2.11</td>
<td>-2.50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>#22.5</td>
<td>1.38</td>
<td>1.03</td>
<td>-2.18</td>
<td>-2.65</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

As it was expected, discrimination parameter of a MR option has lower value in the single-option scoring approximation than when all options are present. This stands for all MR options in PD09 test. Presence of other MR options increases estimates of discrimination parameter for each particular MR option. This occurrence is particularly important for question #22 where inter-item correlations are greatest. Fisher’s test information function depends on the square of discrimination and hence test information is much affected by difference in discrimination parameter estimation.

Fisher’s test information function has lower values for the single-option scoring approximation than for item-scoring, but still higher than for both cluster-scoring methods (Figure 4). It should be noted that the difference between test information functions for different scoring methods is less visible for high-ability examinees.
DISCUSSION

Answers to Multiple Response (MR) questions carry more information than we usually utilize. Similarly to the scoring of Multiple True-False (MTF) questions, we can score MR questions in two ways: to assign scores to MR questions, or to assign scores to all their options. This way, we virtually increase the number of scoring items and consequently increase reliability of the test.

If we analyze test responses using Classical Test Theory, we can see there are a few simple scoring methods that could outperform coarse-grained binary outcome of “all or nothing” cluster-scoring method. Common practice to score MR question with one point only if all options are correct is one of the least informative choices as a scoring method, since a lot of information acquired through responses to the separate options is lost. A major weakness of “all or nothing” scoring method is that the score within an MR question heavily depends on the question’s least discriminative options. Failing to respond correctly to one dubious, and possibly invalid, option collapses examinee’s score on that question to zero even if more demanding and more discriminative options are correctly responded. That way, one bad option can ruin measuring potential of the whole question.

Two important assumptions of Item Response Theory (IRT) are unidimensionality of the construct and local independence. None of these
assumptions has been investigated thoroughly for MR questions thus far. A recent study by Hohensinn and Kubinger (2011) demonstrated that MR questions (with specific instruction to select 2 out of 6 options) measure the same latent trait as the same questions given in MC questions format. Similar analysis in this work would be quite difficult since we compare several scoring methods in combination with different IRT models, which would require a series of analyses and that way would overcome the scope of this paper. Keeping in mind that “it is hard to provide an explicit quantitative criterion for deciding whether an item pool is sufficiently unidimensional to allow application of unidimensional IRT model” (Drasgow & Lissak, 1983), we have explored other quantitative test characteristics that could endanger applicability of IRT to a greater extent.

MR questions scored as clusters of options make guessing less probable than in the case of MC questions. In three-parameter IRT analysis of cluster-scored responses, pseudo-guessing parameter was always close to zero. Hence, we can suppose that two-parameter IRT model with pseudo-guessing parameter preset to zero (2PL IRT) can be very useful for analysis of scores obtained using cluster-scoring methods.

Format of MR questions enables examinees to guess the correct answer for false options easily – if they do not mark those options, their responses will be scored as correct. For true options, unmarked options are always scored as incorrect. This technical difference causes different examinees’ behavior in interaction with true and false options. Tendency that examinees leave too many options unmarked makes correct responses to false options more probable and successful guessing more likely. IRT analysis easily detects difference between item-response characteristics of true and false options: discrimination of false options is lower than of true options, while pseudo-guessing parameters become polarized and, in most cases, converge either to 0, for true, or to 0.5 for false options. Pseudo-guessing parameter for false options might have even greater values, but BILOG imposes 0.5 as maximal plausible value for this parameter. Although interpretations of this occurrence are limited by small number of MR questions in PD09 test, it seems that item-response characteristics of false options cannot be adequately described by ordinary three-parameter IRT models.

High level of inter-item dependency decreases our ability to accurately estimate item parameters. Local dependency of options in the same MR question is a source of bias in estimation of item parameters: discrimination, as well as item difficulty, seems to be higher than it is. Imminent consequences of this bias are apparently greater values of item and test information functions. Simple adjustment of item parameters estimation procedure could diminish these effects of inter-item dependency. Therefore, we propose approximation method where parameters of individual MR option are estimated as if it was the only option in that MR question. Cluster-scoring methods are not affected by inter-item dependence of MR questions since the aggregation of responses to individual options masks internal structure of the question.
CONCLUSIONS

Using MR questions in educational testing makes scoring procedure more complicated than for MC questions. Generally, there is a variety of scoring methods for MR questions and we are supposed to choose the optimal solution for the purpose of a test. Although MR questions have been neglected for decades, computer-based testing has restored interest in scoring of this type of test questions. Recently, a few studies have explored benefits and disadvantages of different cluster-scoring methods for MR questions with fixed number of true options (Bauer et al., 2011; Eggen & Lampe, 2011; Jiao et al., 2012; Kastner & Stangla, 2011). To our knowledge, more general evaluation studies of scoring methods for MR questions are not published yet.

Cluster-scoring methods are more robust than methods for item-scoring. Therefore they are a natural choice for high-stakes exams and operational tests, which we use to measure how much students know. However, if we want to pull out information what they know, like in diagnostic or trial tests, cluster-scores would be difficult to interpret since we do not know which of options they responded correctly. Instead, we should record all responses to individual options, which is a trivial task within computer-based tests, and try different scoring methods in order to choose the most appropriate one. The chosen scoring method has to be a compromise between loosing information for responses to individual options and dealing with nontrivial internal structure of response patterns.

Various cluster-scoring methods applied on MR question produce scoring-items with various difficulty and discrimination. The most common, “all or nothing” cluster-scoring method has a serious disadvantage that one indiscriminative option can destroy measuring potential of the whole question. Less rigid scoring methods, like “4+” or “all T” analyzed in this paper, might be more useful solutions.

If the purpose of an operational test is to rank examinees according to knowledge or ability, responses can be also scored polytomously. Although it is expected that polytomous scoring increases reliability and information function, at least for MR questions with fixed number of true options, difference between these measures for polytomous and dichotomous scoring appears to be small (Jiao et al., 2012) or even negative (Eggen & Lampe, 2011). For the test analyzed in this paper polytomous cluster-scoring methods show greater item discrimination and test reliability than corresponding dichotomous methods. This issue deserves additional research and deeper analysis, especially in the context of different instructions given to examinees concerning number of true options.

For explorative or pilot tests, it is important to determine properties of all options of MR questions in order to describe what students really know, how do they interact with particular options, and finally to collect information for questions revision. That is why we need a scoring method that would preserve information about individual responses instead of aggregated scores. Usage of item-scoring instead of cluster-scoring methods certainly brings more information
about student-item interactions. We can use test information function to estimate how big this gain is across the ability scale.

The simple idea that all options of MR questions should be scored as individual test items has two major difficulties: 1) false options have item-response characteristics that are hard to model and use along with the other items; and 2) responses to individual options within the same MR question could be too dependent on each other. These difficulties lead to an overestimation of item discrimination and test information function.

We cannot distinguish responses where examinees left false options unmarked because they thought that statements are false, and where they failed to mark them for some other reasons. Consequently, responses to false options have much greater noise than true options within MR questions, which represents a challenge for data analysis. This problem might be solved with simple heuristics, i.e. to assign lower weight to noisy false options. However, item response curve of false options is still not described well, which prevents us from calculating weights properly. An idea pursued in this paper – to neglect false items during the scoring procedure – represents a special case of weighted options model, which appears to slightly improve metric characteristics of MR questions. Further research should find adequate item-response model for false options in various contexts and various kinds of tests.

Responses to MR questions' options depend on each other more than responses to different MC questions in the test. This occurrence causes bias in the estimation of MR options discrimination. Extreme cure for this bias is to detect such a question and abandon it from the test. If we have to keep them, we should decrease the number of such options in the question (Yen, 1993). Alternatively, we can use single-option approximation, which seems to be less biased way to estimate item parameters.

It is important to emphasize that conclusions given in this paper are results of secondary analysis of a test with small number of MR questions. Although most of findings have statistical significance, it will be of practical significance to explore characteristics of various scoring methods on a larger sample and much bigger number of MR questions in a test in future studies.

Finally, this kind of secondary analysis can be useful for other types of test questions as well. Matching questions, for instance, which are common in many paper & pencil and computer-based tests, can be taken as more complex variant of MR questions, which have similar problems with inter-item dependency.

REFERENCES


