Psicológica (2017), 38, 93-109.

Does the position of response options in multiple-choice tests matter?

Christine Hohensinn[1] and Purya Baghaei[2]

[1]*University of Vienna, Austria*

[2]*Islamic Azad University, Mashhad Branch, Mashhad, Iran*

In large scale multiple-choice (MC) tests alternate forms of a test may be developed to prevent cheating by changing the order of items or by changing the position of the response options. The assumption is that since the content of the test forms are the same the order of items or the positions of the response options do not have any effect on item difficulty and other psychometric characteristics of the test. The purpose of the present investigation is to model the difficulty of the options' positions (*a, b, c*, and *d*) in a high-stakes MC test using the linear logistic test model (Fischer, 1973). Findings revealed that options' positions have very slight differences in difficulty and as the position of the correct option moves toward the end of the set of response options it becomes slightly more difficult.

Multiple-choice (MC) test format is the most commonly used test format in large scale educational testing. The ease of administration and scoring and high reliability of MC items are amongst the reasons why MC items are so popular (Baghaei & Amrahi, 2011). However, MC format has also been criticized for being sensitive to several construct-irrelevant factors including testwiseness, number of response options, susceptibility to cheating and guessing, and pattern guessing (Baghaei & Amrahi, 2011).

Researchers have previously addressed some of these issues extensively but some other issues are under-researched. For example the issue of the optimal number of response option in MC items has been extensively addressed (see Haladyna, 2004 for a complete review). The final conclusion of most of these studies is that the number of options (3, 4,

[2] Corresponding author: Purya Baghaei, English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran. Email: pbaghaei@mshdiau.ac.ir

or 5) has little impact on item difficulty and other characteristics such as reliability and item discrimination with the 3-option items being slightly superior in terms of item discrimination.

The other commonly addressed issue, which is not necessarily specific to MC items but most commonly arise in large scale testing, is the issue of item order. Research in educational and psychological testing shows that items may be influenced by their location in the test. Wainer and Kiely (1987) refer to this as ‘context effect’ and define it as “any influence or interpretation that an item may acquire purely as a result of its relationship to the other items making up a specific test” (p.187). Previous research on the impact of item order, although not being in complete agreement, indicates none or only small effects of item position on the psychometric properties of test items and examinees’ performance (e.g. Hohensinn, Kubinger, Reif, Schleicher, & Khorramdel, 2011). Nevertheless, some researchers have also argued for the development of psychometric models which incorporate context effect in order to parameterize it (Brennan, 1992; Davey & Lee, 2011).

Likewise, in large scale assessments where MC items are employed alternate forms of a test may be prepared by changing the position of response options to prevent answer copying. The common assumption is that when preparing MC items the position of response options seems to be completely irrelevant to test design as long as answers are randomly assigned to positions or are equally distributed among them (Attali & Bar-Hillel, 2003). In their revised taxonomy of MC item writing guidelines Haladyna, Downing, and Rodriguez (2002) list five major categories of concern in MC item writing: content concerns, formatting concerns, style concerns, writing the stem, and writing the choices. Each category has a number of subcategories in the form of guidelines or tips where ‘writing the choices’ category has the highest number of guidelines (14 tips). Two of these guidelines refer to the option positions: *vary the location of the right answer according to the number of choices* and *place choices in logical or numerical order* (p.312). However, there is no specific guideline on the exact location of the correct response.

When alternate test forms, which are only different in the order of items or the location of the response options, are used the assumption is that a common metric is maintained across forms and there is no need for equating; equating is only needed when partially or entirely different sets of items are presented in different test forms. The problem, however, is that even when identical items are presented in all forms it is difficult to maintain a common metric to compare examinees who take different forms of a test since the difficulty parameters of the items may change across forms due to context effect.

The issue of item position and option order has been broached from another perspective by other researchers. Bachman (1990) considers item sequence an aspect of test method which could contaminate the measurement of the construct. Addressing test method is essential for investigating validity. To ensure validity the test method should not affect examinee performance or interfere with the measurement of the construct. By implication, if response position alters item difficulty it can be a component of test method that affects the measurement of the construct and should be taken into consideration.

Although the effect of item order or item position effect has been investigated to a relatively large extent (Hahne, 2008; Hohensinn, et al., 2011), research on option position effects in MC tests is scarce. Cizek (1994) investigated such effects in a 20-item test from a certification examination for medical students. The structure of the test in Cizek's study was not canonical MC. In his test, examinees were required to find answers to 20 questions from a list of 30 options. Examinees (*n*=759) were randomly assigned to two response booklets which differed only in the order of the 30 options. Results showed that classical item difficulty and discrimination values change across the two forms for some items, although they remain highly correlated (*r*=.99). Furthermore, no relationship was found between the position of the correct response and item difficulty.

Attali and Bar-Hillel (2003) introduce the concept of 'edge aversion' and 'middle bias' and state that test constructors tend to 'hide' the correct response in the middle position and examinees tend to seek it in exactly the same position. They argue that guessing examinees are more likely to select options in the middle positions and, therefore, items with middle correct answers are easier and less discriminating compared to items whose answers are placed in the edges. With data from real large scale exams they show that middle position items are significantly easier than edge position items. Since guessing is more prevalent in harder items they demonstrate that percentage correct due to the middle bias increases as difficulty increases. They also demonstrate that classical item discrimination (point-biserial correlations) indices decrease for items when correct options are in the middle position. The effect is more pronounced for harder items. They also study the effect of option position on IRT item parameters. They demonstrate that middle-keyed items are significantly easier, less discriminating, and more susceptible to guessing compared to extreme-keyed items.

Since the effect of response option positions in MC items has not been investigated thoroughly the purpose of the preset investigation is to model such effects with the linear logistic test model (Fischer, 1973). More specifically, the aim of the present study was to examine whether the position of the solutions in MC items contribute to the difficulty of the items. In other words, we studied whether the difficulty of items is only due to the content of the task – that is the item stem – or rather the position of the correct answer has also an effect on the difficulty parameter of the item.

METHOD

Data source and material

The test analysed in this study is the Iranian National University Entrance Test, a four-option multiple-choice high-stakes test held annually to admit candidates to master's programmes in English studies. The test is composed of four sections of grammar (10 items), vocabulary (20 items), cloze (10 items), and reading comprehension (20 items). The candidates are supposed to answer the test in 60 minutes. Responses of 21642 candidates (73% male) who took the test in 2013 were analysed. The test is prepared in four booklets which are only different in the position of the response options and are randomly assigned to test takers.

Linear logistic test model

Linear logistic test model (LLTM; Fischer, 1973) was employed to analyse the data. LLTM is an extension of the Rasch model (Rasch, 1960/1980) which imposes some linear constraints on item parameters. That is, item difficulty parameter is hypothesized to be a linear combination of the difficulty of several basic parameters. LLTM assumes that item difficulty parameterβ_i is a weighted sum of the basic parameters η_j. The item response function for the standard dichotomous Rasch model (Rasch, 1960/1980) is expressed as follows:

$$P\ (X_{vi}=1|\ \theta_v\ ,\beta_i)=\frac{exp\ (\theta_v-\beta_i)}{1+exp\ (\theta_v-\beta_i)} \qquad (1)$$

with $\theta_v$being the the person parameter of person v.

LLTM imposes the following linear constraint on the difficulty parameterβ_i:

$$\beta_i=\sum_j^p q_{ij}\,\eta_j+c \qquad (2)$$

where q_{ij} is the given weight of the basic parameter j on item i, η_j is the estimated basic parameter j reflecting the difficulty of the basic parameter, and c is a normalization constant.

The basic parameters are the cognitive operations or steps involved in solving the items which contribute to item difficulty. The model allows parameterizing these steps. By adding up the difficulty of the steps which are involved in solving an item the overall difficulty of the item, estimated by standard Rasch model, should be approximated. The closer the Rasch model based item parameters and the LLTM reconstructed item parameters (cf. Equation 2) are, the better the fit of the LLTM and hence stronger support for the substantive theory which guides item decomposition (Baghaei & Ravand, 2015; Baghaei, & Kubinger, 2015).

For instance, for correctly answering an elementary math item like$\frac{(4+5)*3}{3}$, we hypothesize that, examinees have to master three basic operations, namely, addition, multiplication, and division. LLTM assumes that the difficulty of this item is the sum of the difficulty of these three basic operations. LLTM estimates the difficulty of these three operations, referred to as basic parameters η, and then computes the difficulty of the item by adding up the difficulty of these three operations. The difficulty of this item is also independently estimated with the standard Rasch model. If our theory, which postulates that the difficulty of this item is the sum of the difficulty of the three basic operations of addition, multiplication, and division, is correct then the LLTM reconstructed item difficulty should be close to the difficulty estimated by the standard Rasch model. The difficulty of the basic operations show to what extent each operation contributes to the overall item difficulty.

Study design

LLTM can also be used to model construct irrelevant factors such as item format effects and item position effects (Kubinger, 2009). By forming virtual items and entering positions as basic operations which contribute to item difficulty they can be parameterized and their impact on difficulty be estimated. In this study, items with the same stem but different correct response positions or item locations were considered different items, i.e., “virtual items”. The following example illustrates how LLTM can be used to model response position effects.

Table 1 represents a hypothetical data matrix containing 12 students’ responses to a test with two items in four different booklets. Each booklet is taken by three students. The booklets contain the same items but only differ in the position of the correct option. Here 1s and 0s indicate whether the examinee has got the item right or wrong. The first three examinees have taken Booklet 1; the second three have taken Booklet 2, and so on. The shaded areas are missing responses. Note that this is just a hypothetical example and in a real study some items must have the same correct response positions in different test forms so that the data matrix is linked to avoid estimation problems.

Table 1. Hypothetical data matrix with two items and four booklets. The items in different booklets differ only in the position of the correct answer.

	1.1	2.1	1.2	2.2	1.3	2.3	1.4	2.4
1	1	1						
2	0	1						
3	0	0						
4			1	0				
5			0	1				
6			1	0				
7					1	0		
8					1	0		
9					0	0		
10							1	0
11							0	0
12							0	1

To develop a design matrix **Q**, each response position is considered a basic operation. Therefore, when there are four positions for the correct option in an MC item there are four basic parameters plus the number of items. Items 1 and 2 measure real substantive operations 1 and 2 (CO1 and CO2 in the Table 2) with their corresponding item stems. These two items 1 and 2 are in four different booklets and in each booklet the position of the correct response changes. In Table 2, 1.1 indicates item1 in Booklet 1, and 1.2 indicates item 1 in Booklet 2. That is, for the LLTM each item with each response position is considered a new item ("virtual item"). P1 to P4 refer to the correct options positions *a, b, c*, and *d* in a typical four-option MC item. The first row shows that item 1 in Booklet 1 measures cognitive operation 1 and the correct response is in position 1. The same item in Booklet 2 measures the same operation but the correct response has moved to position 2, and so on. With this scheme the difficulty parameters of the response positions can be estimated and their contribution to item difficulty be assessed.

Table 2. Example of a design matrix Q for analyzing response position effects. The rows display the virtual items and the columns show the basic parameters (there are two item stems and four response positions).

Virtual item	CO1	CO2	P1	P2	P3	P4
1.1	1		1			
1.2	1			1		
1.3	1				1	
1.4	1					1
2.1		1	1			
2.2		1		1		
2.2		1			1	
2.2		1				1

ANALYSES AND RESULTS

The main part of the sample, i.e., 97 % of the candidates responded to each of the 60 items. Two percent of the students had one missing response and the maximum number of missing responses was 19 missing items in one observation.

In the first step of data exploration, one observation had to be excluded because no demographic variables were available for this person. Therefore the sample size of the LLTM reduced to 21641. In one of the booklets, for two items there were only missing responses. This was strange as none of the 5400 respondents who had taken this booklet had tried this item. This was considered a problem in the data input. To be on the safe side, these two items were excluded and the remaining 58 items were analyzed. The missing responses were left as missing data because the extremely low percentage of missings should not have a noticeable effect on the results of the data analysis.

For analyzing the data with LLTM and the Rasch model eRm package (Mair, Hatzinger, & Maier, 2014) in R software package, version 3.11 (R Core Development Team, 2015) was used. As described in the materials part, the test consists of four different subsections. To ensure the unidimensionality, the Rasch model was estimated for the 60 items as a first step. Item fit measures – the mean-squared infit and mean-squared outfit (Bond & Fox, 2007) were calculated to assess the model fit of the items. Taking the often mentioned values 0.7 and 1.3 as cut-off values, only three of 60 items exceeded the cut-off 1.3 for the outfit MSQ only slightly (with values: 1.46, 1.51, 1.38). The Rasch model fit was assessed furthermore: Next, the items with correct answers in different positions in different booklets were expanded in the data set by creating 'virtual items'. That is, for example, if for item 1 the correct answer was in the first position in booklet 1 and in the third position in the other booklets, this item was treated as if it was two separate items – item 1.1 and item 1.3 (virtual item for item 1). The whole data set was expanded in this way. The 58 items were split into a total of 116 items. Because six items had the solution on the same position in each of the four booklets, these items served as linking items making it possible to estimate a Rasch model for the whole expanded data set.

As a first step, the Rasch model for the expanded data set including the virtual items was estimated. The fit of the Rasch model was assessed using item fit statistics (infit MSQ and outfit MSQ). For the Infit mean-square and the Outfit mean-square, the often mentioned cut-off values 0.7 and 1.3 were chosen. Additionally, with a graphical model check the fit of individual items were inspected visually. For the graphical model check (Kubinger, 2005), the sample was split according to the mean of the total score (i.e. into high and low achievers). For each subsample, item parameters were estimated and cross plotted graphically. Due to the feature of sample independence of the Rasch model item parameter estimates must be the same (except for the error).

Because of the very large sample size we decided not to apply inference statistics such as the Andersen Likelihood Ratio test (Anderson, 1973) for testing the Rasch model and a Likelihood Ratio test for comparing Rasch model and LLTM. Obviously, the tests would have been significant only because of the large sample size. Instead, infit and outfit indices, graphical model checks, comparing the reconstruction of the Rasch model based item parameters by the LLTM, and information criteria AIC and BIC were used to assess model fit.

The infit MSQ values showed no aberrant response behavior for the items (see Table 3); the outfit MSQ values indicated seven of the 116 expanded items had aberrant responses. These seven items were also misfitting in the graphical model check, too (see Figure 1). For example, items 51 (with correct response in position 4) and item 22 were among them.

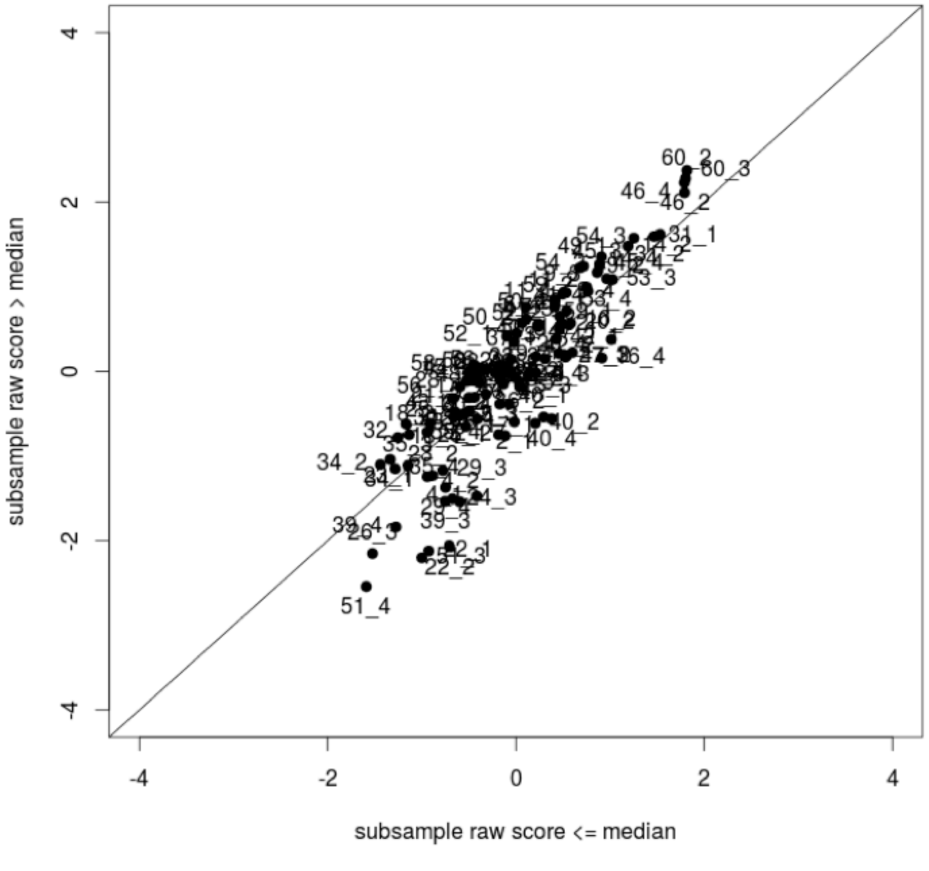
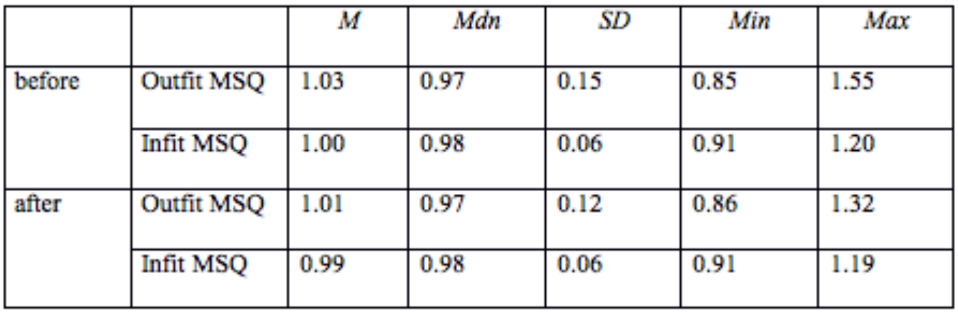


Figure 1. Graphical model check of the expanded data set with 116 items. The item parameter estimates of the two subsamples (divided by the median of the raw score) are cross plotted here. The label indicates the item number and the position of the solution (e.g. 51_4 means item number 51 with the correct answer in position 4).

These seven items consisted of only four different items stems, three of them were the same items that were misfitting in the first Rasch model analyses of the 60 items. As a consequence, for the LLTM analysis these seven virtual items were excluded and the Rasch model was reestimated without these items. In this reduced item set, the outfit MSQ for one item exceeded the cut-off value, but because of the very small deviation (the outfit was 1.32), this item was left in the item pool. The reduced item pool with 109 virtual items was used for further analyses.

Table 3. Summary statistics for Outfit and Infit MSQ. The two upper rows contain the statistics for the whole item set. The two lower rows display the results after eliminating seven non-fitting items.



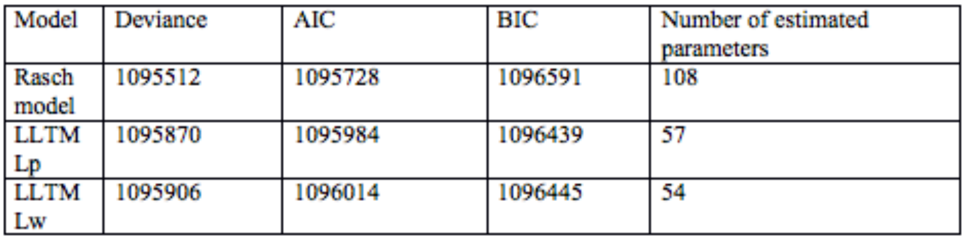
		M	*Mdn*	*SD*	*Min*	*Max*
before	Outfit MSQ	1.03	0.97	0.15	0.85	1.55
	Infit MSQ	1.00	0.98	0.06	0.91	1.20
after	Outfit MSQ	1.01	0.97	0.12	0.86	1.32
	Infit MSQ	0.99	0.98	0.06	0.91	1.19

Next, two LLTMs were estimated to see if a position effect occurs in this data set. For the first LLTM (Lp) the design matrix was established in such a way that the content of the items – the item stems – were modeled as well as each answer position. For each position one basic parameter was assumed – consequently, this model formulation includes the assumption that the effect of a position is homogeneous for each item. Finally, the formulation of Lp resulted in a total of 55 plus 4 basic parameters to estimate. Thus in this model the item difficulty is assumed to result from the difficulty of the item stem and the difficulty of the position. Avoiding overparameterization of the model, the design matrix had to be reduced to a 109 by 57 matrix. One item stem basic parameter and one position basic parameter had to be set to zero. A second LLTM (Lw) was estimated that included only the 55 item stem basic parameters without modeling an effect for the positions.

That is, the design matrix of Lw only differed from the design matrix of Lp by eliminating the position basic parameters. With Lw, it is assumed that the difficulty of the item depends only on the difficulty of the item stem. Again, one of the 55 basic parameters was set to zero resulting in 54 estimated parameters.

To summarize, three hierarchically nested models were estimated: the most general of the three is the Rasch model for the expanded data set – that is, each virtual item was modeled as a separate item. The LLTM Lp modeled each item stem and the position of the correct answer and the LLTM Lw modeled only the item stems.

Table 4. Number of estimated parameters, deviance, and information criteria AIC and BIC for the estimated models.



Model	Deviance	AIC	BIC	Number of estimated parameters
Rasch model	1095512	1095728	1096591	108
LLTM Lp	1095870	1095984	1096439	57
LLTM Lw	1095906	1096014	1096445	54

Table 4 presents the deviances, information criteria, AIC and BIC and the number of estimated parameters for the three estimated models. From the number of estimated parameters it can be seen that the two LLTM models are considerably more parsimonious than the Rasch model; the differences in the number of estimated parameters between the Rasch model and these two models are 54 and 51, respectively.

The results of the AIC and BIC are in agreement: they reveal that the LLTM that includes position effects has the best fit. But the difference to the LLTM without the position effects is low (only 6 points in the BIC, 30 points in the AIC).

The item difficulty parameters estimated by the Rasch model were reconstructed by the LLTM basic parameters according to Equation 2. These reconstructed item parameters are graphically plotted against the Rasch model based item parameters in Figure 2. The plot contains the item parameters reconstructed by the model Lp as well as Lw. It can be seen that Lp offers only slightly better item parameter recovery. This confirms the small differences in the information criteria, AIC and BIC reported earlier.

Table 5. Estimates of the position basic parameters from model Lp: the estimated parameter value, the SE of parameter estimation and the 95% confidence interval are presented.

Position	Estimate	SE	95 % CI	
			Lower	Upper
2	.011	.008	-.005	.027
3	.057	.013	.033	.082
4	.067	.012	.043	.091

Note that the first position basic parameter had to be set to zero due to over-parameterization. That is, the value of the other basic parameters modeling answer position can be interpreted in comparison to the fixed parameter, i.e., the first position parameter can be seen as "reference". The estimated position basic parameters are very small (see Table 5), with only the parameters for position 3 and 4 showing an estimate distinct from 0. Of course, as pointed out earlier, because of the very large sample, any inferential statistical interpretation of the confidence interval must be done with caution. However, in summary, the influences of the positions seem very small with a tendency that the item gets more difficult if the correct answer is located toward the end of the set of options.

For each item that was presented with at least two different correct response positions, the difference of the parameter estimates were calculated. For example item 2 was administrated in different booklets with correct answer position in positions 1, 2 and 3 (thus item 2 was split into three virtual items 2.1, 2.2. and 2.3 in the expanded data matrix). For each of these virtual items, the item difficulty parameter was estimated with the Rasch model. Then, the differences between these parameter estimates were calculated. For example, for item 2 the difference of $(\beta_{2.1})$-$(\beta_{2.2})$, $(\beta_{2.2})$-$(\beta_{2.3})$ and $(\beta_{2.1})$-$(\beta_{2.3})$ were calculated. All the differences calculated in this way are displayed in Figure 3.

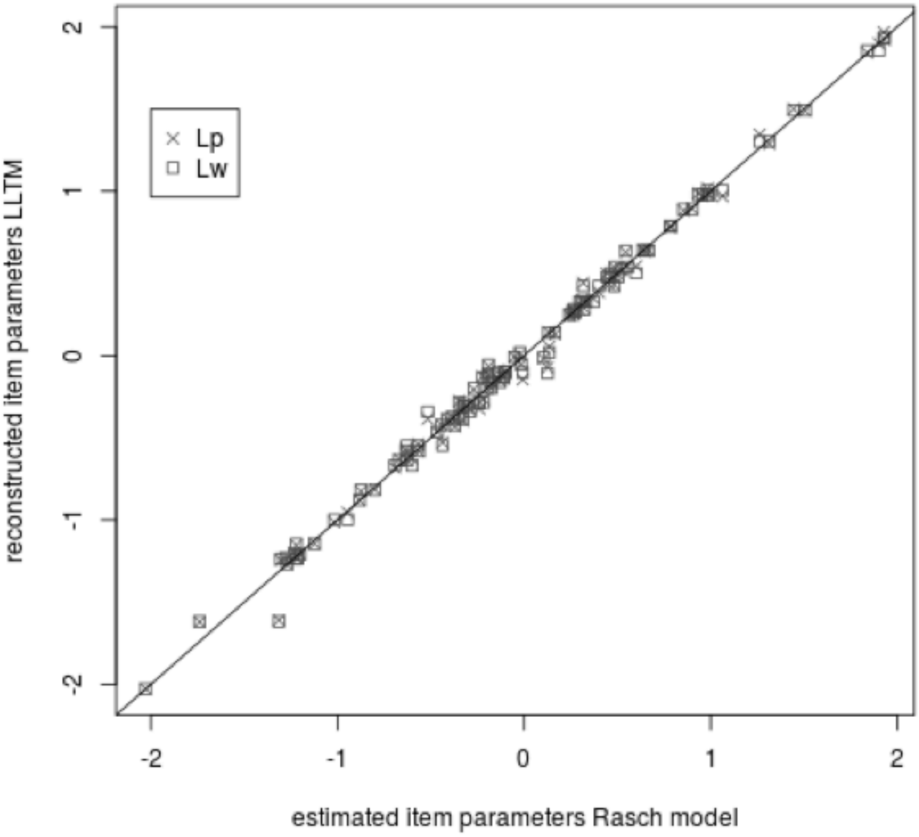


Figure 2. Reconstruction of item parameters estimated by the Rasch model by means of the basic parameters of the two LLTM models Lp and Lw.

The abscissa displays between which positions the differences were calculated. Each point in the graphic displays one difference. The boxplots show that all medians of differences lie a little bit under the point 0. This means that there is a slight tendency towards higher difficulty for the item as the solution position moves toward the end of the set of response options.

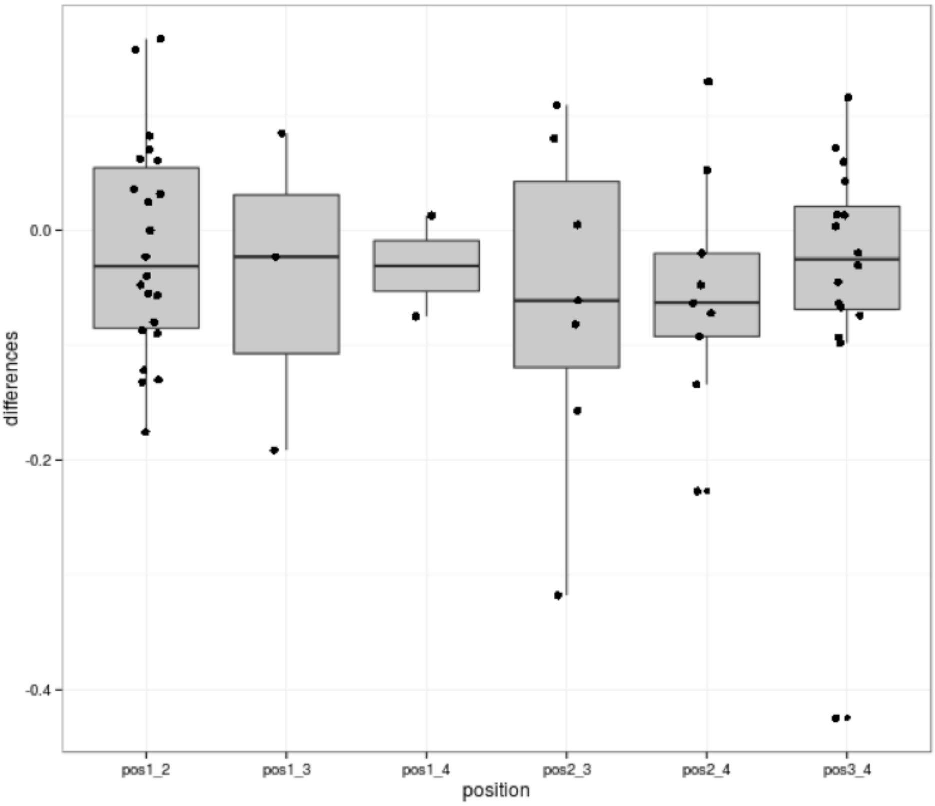


Figure 3. Graphical display of differences between positions' estimates.

DISCUSSION AND CONCLUSION

It is commonly believed that the position of the correct response in multiple choice items has no effect on item difficulty as long as the correct options are randomly or evenly assigned to different positions. However, empirical research on the effect of the response positions in MC items is scarce. If the correct response position affects item difficulty a common metric cannot be maintained across alternate forms of the test which are only different in the location of the response options.

In this study an attempt was made to estimate the contribution of the position of the response options to item difficulty in MC items. LLTM (Fischer, 1973) was employed to model and parameterize response positions in a four-option high-stakes MC test of English as a foreign language.

Three models were estimated: a standard Rasch model for the expanded data set with 109 virtual items, an LLTM with 55 item stems plus four option positions, and an LLTM with the 55 item stems without the option positions. According to the BIC the Rasch model with 109 virtual items had the worst fit. The model which included the option positions had a slightly better fit than the model without modeling the option positions. However, the very small parameter estimates for different response positions suggested that the location of the correct response contributes little to the overall item difficulty.

A close examination of the changes in the difficulty estimates of the items which are presented in different booklets with correct replies in different positions (Figure 3) showed that as the correct option moves toward the end of the set of options the item gets slightly more difficult. This effect is not very noticeable, though.

The findings of this study suggest that the position of the correct option has very little effect on MC item difficulty and the common practice of distributing correct options randomly is valid. Nevertheless, in this study, the four-option MC format was examined. The effect is small, but it seems that as the correct option moves to the end of the set of options the greater the effect on the item difficulty. Thus, further research on MC format with more response options seems important to find out whether this trend significantly affects response options located further toward the edge. Using an MC format with only up to four response options, item developers need not be very much concerned about the position of the correct options as long as pattern guessing is prevented by randomizing the answer key. However, note that Attala and Bar-Hillel (2002) state that knowledge of the fact that the answer key is balanced, i.e., correct responses are equally assigned to the positions, can be exploited by testwise candidates. They argue that when examinees are aware that the answer key is balanced they can add between 10 and 16 points to their final SAT score, on average, depending on their knowledge level and, therefore, recommend randomized answer keys instead of balanced.

REFERENCES

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38*, 123-140.

Attali, Y. & Bar-Hillel, M. (2002). Seek whence: Answer sequences and their consequences in key-balanced multiple-choice tests. *American Statistician, 56*, 299-303.

Attali, Y. & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*, 109-128.

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

Baghaei, P., & Kubinger, K. D. (2015). Linear logistic test modeling with R. *Practical Assessment, Research & Evaluation*, *20*, 1-11. Available online: http://pareonline.net/getvn.asp?v=20&n=1

Baghaei, P., & Ravand, H. (2015). A cognitive processing model of reading comprehension in English as a foreign language using the linear logistic test model. *Learning and Individual Differences, 43,* 100-105.

Baghaei, P., & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53*, 192-211.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model. Fundamental Measurement in the Human Sciences. Mahwah, New Jersey:* Lawrence Erlbaum Associates.

Brennan, R. L. (1992). The context of context effects. Applied Measurement in Education, 5 , 225-264.

Cizek, G.J. (1994). The effects of altering the position of options in a multiple-choice examination. *Educational and Psychological Measurement, 54*, 8-20.

Davey, T., & Lee, Y. H. (2011). *Potential impact of context effects on the scoring and equating of the multistage GRE revised general test*. ETS Research Rep. No. RR-11-26. Princeton, NJ: ETS.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Hahne, J. (2008). Analyzing position effects within reasoning items using the LLTM for structurally incomplete data. *Psychology Science Quarterly, 50*, 379-390.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15,* 309-334.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E, & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation, 17*, 497-509.

Kubinger, K. D. (2009). Applications of the linear logistic test model in psychometric research. *Educational and Psychological Measurement, 69*, 232-244.

Kubinger, K. D. (2005). Psychological test calibration using the Rasch model: some critical suggestions on traditional approaches. *International Journal of Testing, 5*, 377-394.

Mair, P., Hatzinger, R., & Mair, M. J. (2014). *eRm: extended Rasch modeling* [Computer software]. R package version 0.15-4. http://CRAN.R-project.org/package=eRm.

R Core Development Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. Expanded Ed. University of Chicago Press, Chicago, IL. (Originally published 1960, Pædagogiske Institut, Copenhagen.)

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

(Manuscript received: 22 February 2016; accepted: 13 June 2016)