# An option-based partial credit item response model

**Yuanchao (Emily) Bo*, Charles Lewis and David V. Budescu**

**Dept. of Psychology, Fordham University, Bronx, NY**

***Correspondence should be addressed to Yuanchao (Emily) Bo, Department of Psychology, Fordham University, Rose Hill Campus, 441 East Fordham Road, Bronx, NY, USA. (e-mail: ybo@fordham.edu)**

## Abstract

We introduce a new, option-based, partial credit IRT model for multiple-choice items with, possibly, more than one correct response, and describe a new scoring procedure based on a weighted Hamming distance between the vector of option responses and the correct option response vector. We report results of simulations and real data analysis that support the feasibility of the approach and the superiority of the new scoring rule, relative to several alternatives.

Key words: Item response theory, partial credit, partial knowledge, Hamming distance, multiple choice, scoring rule.

Multiple-choice (MC) tests have been consistently criticized for allowing guessing and the failure to credit partial knowledge (e.g., Budescu & Bar-Hillel, 1993). Several scoring methods have been proposed in the past decades including formula scoring (Thurstone, 1919) and elimination scoring (Coombs, Milholland & Womer, 1956). Ben-Simon, Budescu and Nevo (1997) have developed a classification system summarizing most of the alternative methods. Modern test theory also offers several alternatives to address the issue of guessing and partial knowledge in MC tests, such as the introduction of a guessing parameter in the 3-parameter IRT model (Birnbaum, 1968), and a variety of categorical response IRT models, including the partial credit model (Masters, 1982), graded response model (Samijima, 1969), and the nominal response model (Bock, 1972).

We propose a new MC partial credit IRT model (for items where the number of correct options can be 1 *or* 2) and a new scoring rule based on a weighted Hamming distance between the option key and the option response vectors.  These modifications reduce the test taker's (TT's) ability to guess and credit the TT's partial knowledge.  Unlike binary IRT models, the estimated TTs' ability is based on information from both correct options and distracters. The proposed new model can be tailored to different formats (one correct option per item, multiple correct options per item, or option elimination responses). The 2PL IRT model is a special case of the proposed model.

*Simulation studies*

Consider an item with 5 options (A, B, C, D, E) that may have 2 or 1 correct options. Suppose the response key for the item is (1, 1, 0, 0, 0). As shown in Table 1, there are 15 possible response patterns. Each of these patterns can be scored using a regular Hamming distance (simple count of correct classifications of options) or a weighted Hamming distance where option discrimination parameters ($a_k$) serve as weights.

Figure 1 shows the option response curves for the model. The curves are based on the sums of the probabilities for the appropriate response patterns. The marginal probabilities for the two correct options (option 1 and option 2) are monotonically increasing functions of the TT's ability.  Conversely, the marginal probabilities for the three incorrect options (options 3-5) are monotonically decreasing functions of the TT's ability.

We simulated responses for a test with 10 identical items. The option difficulty parameters used in the simulation are (1,1,0,0,0) and the option discrimination parameters  are (1.2,2,0.3,0.4,0.8).  TTs are instructed to choose two options per item. We compare the correlation between the actual ability and the scores from different scoring rules (weighted Hamming distance scores, Hamming distance scores, weighted number right scores and grouped number right scores). As shown in Figure 2, the weighted Hamming distance scores have higher correlation with TTs' true ability than the other three scores. The weighted Hamming distance scores have a distribution that is also less skewed than those of the other scores.

Table 1. Possible response patterns for items with 5 options and scoring rules.

| | | | | | | Hamming | Weighted Hamming |
|---|---|---|---|---|---|---|---|
| | A | B | C | D | E | distance | distance |
| case 1 | 1 | 1 | 0 | 0 | 0 | 5 | $a_1$ +$a_2$+$a_3$+$a_4$+$a_5$ |
| *case2* | 1 | 0 | 1 | 0 | 0 | 3 | $a_1$ + $a_4$+$a_5$ |
| *case3* | 1 | 0 | 0 | 1 | 0 | 3 | $a_1$ +$a_3$ +$a_5$ |
| *case4* | 1 | 0 | 0 | 0 | 1 | 3 | $a_1$ +$a_3$+$a_4$ |
| *case5* | 0 | 1 | 1 | 0 | 0 | 3 | $a_2$ +$a_4$+$a_5$ |
| *case6* | 0 | 1 | 0 | 1 | 0 | 3 | $a_2$+$a_3$ +$a_5$ |
| *case7* | 0 | 1 | 0 | 0 | 1 | 3 | $a_2$+$a_3$+$a_4$ |
| *case8* | 0 | 0 | 1 | 1 | 0 | 1 | $a_5$ |
| *case9* | 0 | 0 | 1 | 0 | 1 | 1 | $a_4$ |
| *case10* | 0 | 0 | 0 | 1 | 1 | 1 | $a_3$ |
| *case11* | 1 | 0 | 0 | 0 | 0 | 4 | $a_1$ +$a_3$+$a_4$+$a_5$ |
| *case12* | 0 | 1 | 0 | 0 | 0 | 4 | $a_2$+$a_3$+$a_4$+$a_5$ |
| *case13* | 0 | 0 | 1 | 0 | 0 | 2 | $a_4$+$a_5$ |
| *case14* | 0 | 0 | 0 | 1 | 0 | 2 | $a_3$ +$a_5$ |
| *case15* | 0 | 0 | 0 | 0 | 1 | 2 | $a_3$+$a_4$ |

Note: The correct response pattern is (1,1,0,0,0)

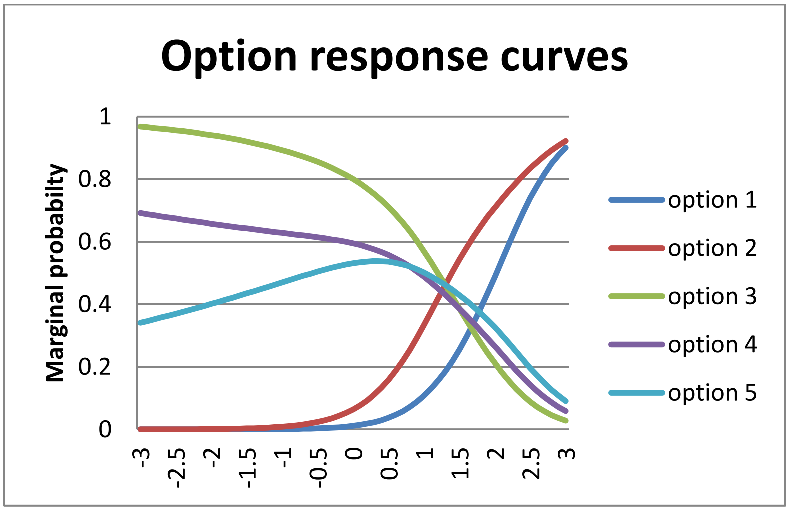

Figure 1. Option response curves for one item.

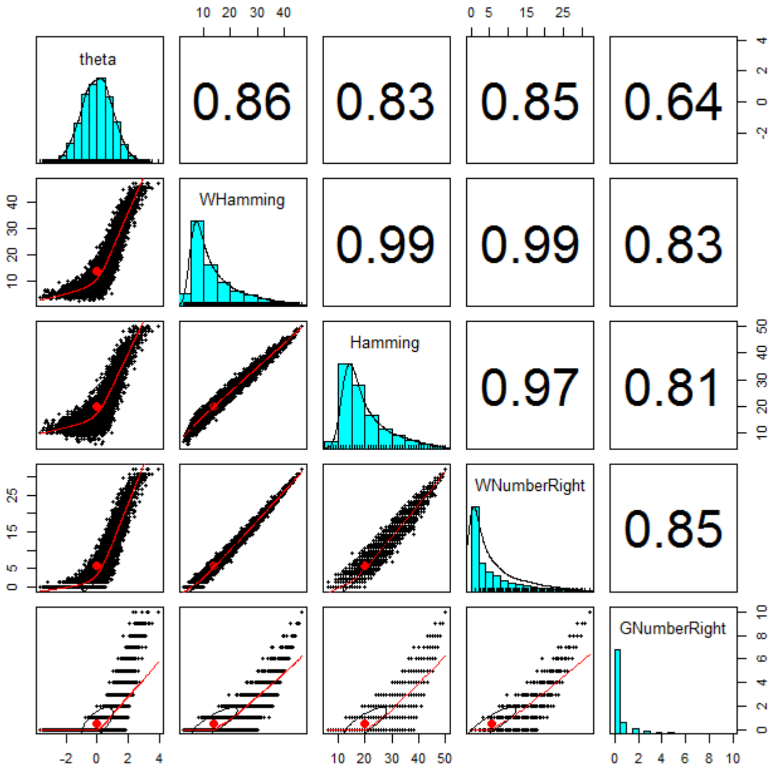

Figure 2. SPLOM of estimates for a test with 10 identical items (Option difficulty parameters (1,1,0,0,0), and option discrimination parameters (1.2,2,0.3,0.4,0.8))

*Real data analysis*

We also applied the model to a real data set which contains 4,393 TTs' responses on a GRE data set with 15 items. A marginal maximum likelihood approach using Haberman's stabilized Newton Raphson algorithm (Haberman, 1988) was used to estimate the model parameters. The algorithm was implemented in the MIRT software (Haberman, 2013). Eleven of the 15 items are the Single Selection MC (SSMC) items and the other 4 items call for choosing 2 out of 6 Multiple Selection MC (MSMC) items. Omissions and the responses-578 cases-which don't follow the specific item instructions, are treated as missing data.

We compared (1) the estimated and empirical response pattern curves, (2) the estimated and empirical option response curves and (3) the various score distributions and the estimated ability distribution. The estimated curves are close to the empirical curves in all items. Figure 3 shows the estimated and the empirical response pattern curves of the 15 response patterns for items 8 (one of the 4 MSMC items). Each panel represents one response pattern and the last panel represents the correct response pattern. The panels are ordered in ascending order of their (observed) response frequencies. The empirical response pattern curves are shown in red dots. The black dots represent the response pattern curves of the model that assumes option local independence. Among the wrong, or partially correct, response patterns, the model replicates well the observed response pattern curves at the higher end of the ability levels. The misfit of the estimated response pattern curves at the lower ability range could be due to asymmetrical ability distribution of the test-takers' in this data set as shown in the first cell of Figure 5. Most of the estimated ability parameters are between -1 to 2. Thus there may not be sufficient information to reproduce the observed response pattern curves at the lower end (theta < -1) of the ability range.

Figure 4 presents the option response plots of item 8. The option response plots are created by aggregating the corresponding response pattern plots. Each panel represents one option of the item. The two colors - black and red - the option local independence model and the empirical data, respectively. The two curves overlap at the higher end (above Theta = -1 ) of the ability in all the panels of the figure. In most panels the two curves diverge from each other at the lower end (below Theta = -1) of the ability. As discussed above, most of the test-takers' ability estimates are above -1, thus there is not sufficient data for the model to reproduce the observed scores.

The scatter plot matrix of the four types of test scores of the 15 items and the estimated ability parameter based on the option-based partial credit model is shown in Figure 5. We (list-wise) deleted omissions and the responses that don't follow the item instructions. The data set analyzed consists of 3,815 test takers' responses. The highest correlation, 0.95, is between the weighted Hamming distance scores and the estimated ability distribution . The correlation between the grouped number right scores and the estimated ability distribution is 0.85. The difference between the two correlations are not as big as the one in the simulation study presented above. This could due to the fact that 11 of the 15 items are the single selection MC items. The weighted Hamming distance scoring rule provides more benefits in the MSMC items than the grouped number right scores because it addresses the issue of guessing and partial credit. In the current real data set in which most items are SSMC items and only a subset are MSMC and, we believe, that the full benefits of the model cannot be detected. .

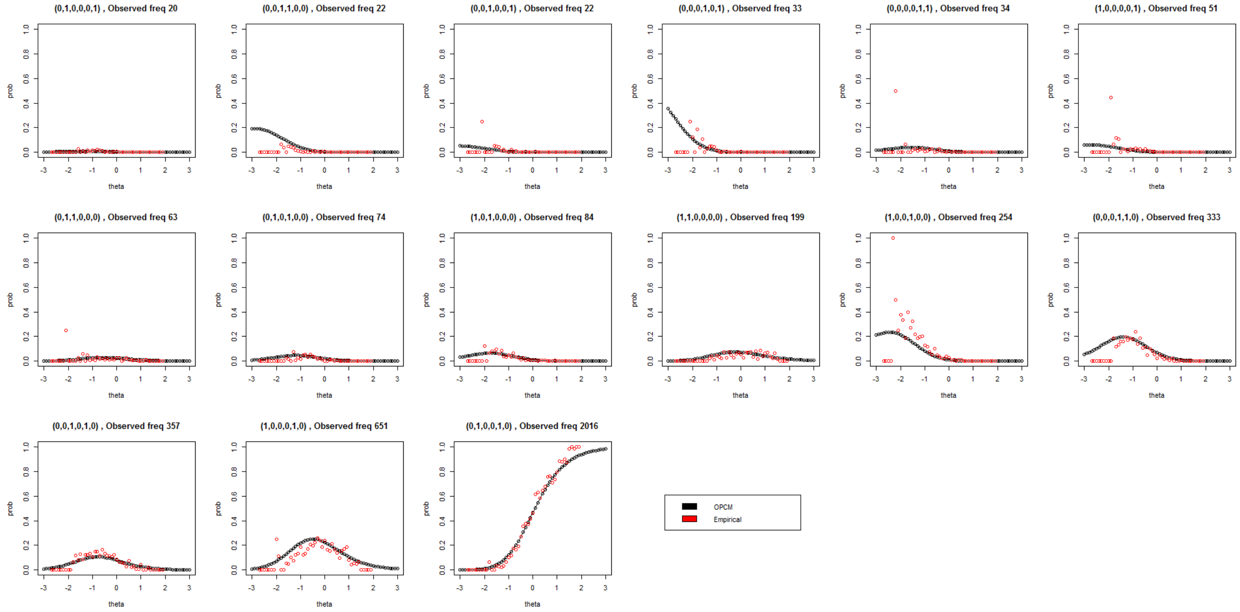

Figure 3. The estimated and the empirical response pattern curves of Item 8.

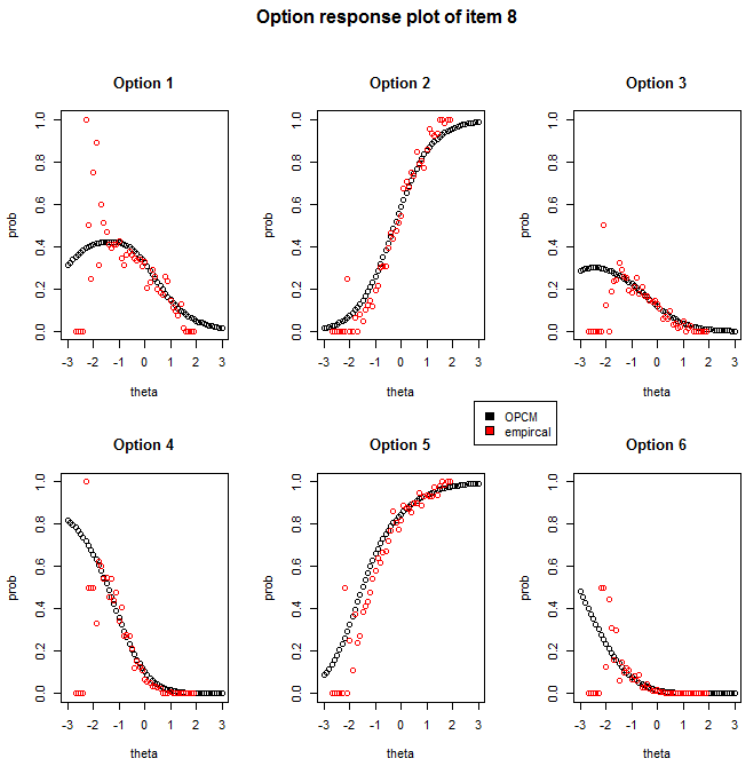

Figure 4. The estimated and the empirical option response curves of Item 8.

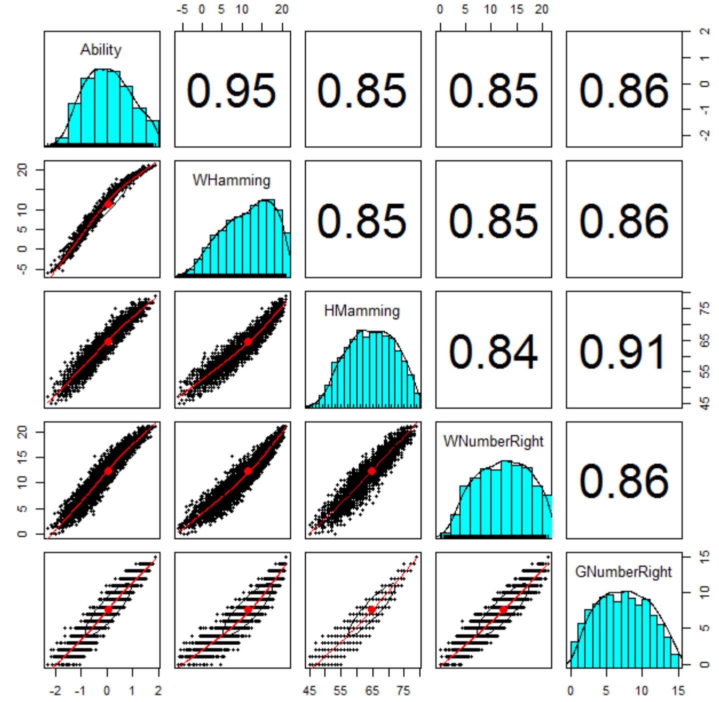

Figure 5. The scatter plot matrix of the score distributions and the estimated ability distribution.

## *Conclusions*

The stimulation study confirms the superiority of the weighted Hamming distance over grouped number correct scores in estimating TTs' abilities when the proposed model is used to generate the item responses. The weighted Hamming distance scoring improves estimation of TTs' abilities by assigning partial credit and extracting information from distracters.

The analysis of the GRE data also shows that the weighted Hamming distance scoring rule has a great potential to improve the precision of the estimation of the TTs' latent traits, compared to the grouped number right scores. The estimated curves based on the option-based partial credit model are very close to the corresponding empirical curves confirming the practicality of the model.

The option-based partial credit model and its underlying scoring mechanism-weighted Hamming distance scores are the very first attempt in the psychometric literature to provide both (1) a feasible and simple scoring procedure for multiple selection MC items and (2) partial credit to TTs by using the information in the distracters. The results from both the simulation studies and the analysis of the GRE data confirm the unique contribution of the weighted Hamming distance scores in manifesting TTs' latent traits.

## References

Ben-Simon, A., Budescu, D. V., & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*, 65-88.

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

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37, 29-51.

Budescu, D. V., & Bar-Hillel, M. (1993). To guess or not to guess: a decision theoretic view of formula scoring. *Journal of Educational Measurement, 30*, 227-291.

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

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.

Haberman, S. J. (1988). A stabilized Newton-Raphson algorithm for log-linear models for frequency tables derived by indirect observation. *Sociological Methodology, 18*, 193-211.

Haberman, S. J. (2013). A General Program for Item-response Analysis that Employs the Stabilized Newton-Raphson Algorithm. *MIRT software Menu.* Princeton, NJ: ETS.

Samejima, F. (1969). Estimation of ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 18.

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