Application of the Hierarchical Item Response Model to Computer Adaptive Test of Graded Response Data

Purpose

The hierarchical item response model (Rijmen, 2011; Yung, Thissen, & Mcleod, 1999) is one of the confirmatory multidimensional item response models and the items have zero loading on some dimensions. The best-know hierarchical model is the bifactor model (Gibbons & Hedeker, 1992). However, sometimes bi-factor (Gibbons & Hedeker, 1992) model cannot fully model the latent construct structure because constructs of more than two levels need to be modeled. For instance, in a reading motivation assessment, items all measure the reading motivation construct: some items measure intrinsic motivation, other items measure extrinsic motivation, and every item measures one submotivation within intrinsic and extrinsic motivation. In this situation, the trilevel model can model this three-level latent construct structure, which is another example of the hierarchical item response model. The hierarchical item response model is more flexible compared to bifactor or trilevel models, which can allow models with a mix of different levels and models with structures including more than three levels. In this study, the model we studied measures 15 constructs and has one general-level factor, four second-level factors, and 14 third-level factors. The hierarchical model applied to the computer adaptive test, which we use in this study, is a mixed model of both bifactor and trilevel models. This model constrains each item to have a nonzero loading on the general-level factor, one second-level factor, and one third-level factor (except the item measuring construct 1, which has only a nonzero loading on the general-level factor and one second-level factor). In this hierarchical item response model, every item has two or three nonzero loading, and theta estimates of 19 dimensions can be reported for each student. When this model is applied to the computer

adaptive test (CAT), different item selection methods can be used. In this study, we compare three different item selection methods.

Theoretical Framework

Item selection method. Previous studies applied only the bifactor model to CAT. Weiss and Gibbons (2007) use a unidimensional item response model, item selection method to select items for the bifactor model CAT. The results of the simulation study show that the correlation between the true general-factor theta and the estimated general-factor theta is about 0.92 and the correlation between the true second-level theta and the estimated second-level theta is about 0.95 after 615 items were administered. The results indicate that this item-selection method can provide accurate estimates for both general-level and second-level thetas; however, students could not take a test with 615 items. This method needs to be modified before being applied in real tests. Seo (2011) in his dissertation used the multidimensional item selection method (Segall, 1996) to select items during the application of bifactor model to CAT. This method selects the item that can provide the best estimate on all thetas. Seo (2011) also compared the Weiss and Gibbons (2007) item selection method with the multidimensional item selection using the fixed test length as the termination rule. The comparison result indicates that the accuracy and efficiency of the general factor theta estimation is similar of these two methods but the multidimensional item selection method provides more accurate and efficient estimation of the second-level theta estimate. When 20 items were administered for each second-level factor, the correlation between the estimated and true general-factor theta is about 0.94 for both methods; the correlation between the estimated and true second-level theta is about 0.5 for the Weiss and Gibbons (2007) item selection method and 0.73 for the multidimensional item selection method. However, these two methods applied only the bifactor model to CAT and used binary data.

In this study, the hierarchical item response model will be applied to CAT, using the graded response data. The Bayesian estimation method (Segall, 1996) will be used to estimate theta and calculate the information matrix. The termination rule sets the length of the test. The **multidimensional item selection method** is the first method compared in this study, which considers and estimates 19 dimensions together during the item selection. The **local multidimensional item-selection method**, the second method compared, considers and estimates only the dimensions on which the current item has nonzero loadings during the item selection. The same multidimensional item selection method will be used, but only two or three dimensions will be considered during item selection. The **unidimensional item selection method**, the third item selection method, considers and estimates only the most specific dimension (lowest-level dimension) on which the current item loads. This study aims to compare these three methods, using simulation data, to find an efficient and accurate method on 19 dimensions of theta estimation.

Method

Hierarchical graded response model. The hierarchical graded response model extends the graded response model (Samejima, 1969) and the item discrimination parameters, and the theta becomes the vector. Thus, the probability of response to item *j* in or above category *t* of the hierarchical graded response model is:

$$P_{jt}(\theta) = \frac{1}{1 + exp(-\boldsymbol{a}'_j\boldsymbol{\theta} + b_{jt})} \quad (1)$$

In equation 1, a_j is the discrimination vector with the nonzero a_{jg} and a_{jks} and other 0 elements, and θ is the latent proficiency level vector with all nonzero elements. The intercept

parameter, b_{jt}, is for score category *t*. The probability of response to item *j* in category *t* follows equation 2:

$$p_{jt} = \begin{cases} 1 - P_{jt+1} & \text{if } t = 1 \\ P_{jt} - P_{jt+1} & \text{if } t \text{ in } (2, \ldots, m-1) \\ p_t = P_t & \text{if } t = m. \end{cases} \quad (2)$$

Conditions. Although it is true that the more items administered, the more accurate the estimation, a long item test is not reasonable for real testing situations. In order to select a reasonable test length that also provides enough accurate theta estimation, five conditions demonstrating difference in test lengths are studied. Since the hieratical graded response model in this study has 19 dimensions, and this model has 15 specific dimensions, at least 15 items need to be administered in order to get an estimate on all 19 dimensions. The administration of 15 items is called a stage. Condition 1 administered one stage (15 items). The number of stages increase according to the order of conditions. Table 1 presents the number of stages and number of items administered for each condition. For each item selection method and each condition, 200 replication studies were conducted. During item selection, item exposure rate was not controlled.

Evaluation criteria. Three criteria were used to compare three item selection methods and find a reasonable test length that provides sufficient accurate theta estimation. These criteria are the Pearson correlation between the true and estimated theta of each dimension for each method and each condition; the root mean square error (RMSE) of the estimation; and the determinant of the covariance matrix of theta. The equation for calculating the RMSE of one dimension is:

$$RMSE(\theta_{kd}) = \sqrt{\frac{\sum_{k=1}^{K}\sum_{r=1}^{R}(\theta_{kd} - \hat{\theta}_{kdr})^2}{K * R}}$$

where *d* is one of the dimensions, *K* is the number of true theta vectors (50 in this study), and *R* is the number of replications (200 in this study). The covariance matrix of the theta decreases if the estimated theta is closer to the true theta. Since the determinant of a matrix provides very important information about the matrix and can be reported easily, the determinant of the covariance matrix is a criterion on the accuracy of estimation. The covariance matrix is the inverse of the information matrix. The covariance matrix was first averaged across 200 replications for each true theta vector, and then the determinant of the average covariance matrix was calculated. The determinants of the covariance matrix for 50 true theta vectors were averaged for every method and every condition.

Data

Item pool. The item pool used for CAT is simulated including 300 items measuring 15 constructs (20 items per construct). There are two or three nonzero item loadings, one for general factors and the others for factors of other levels, for each item. Every nonzero loading is drawn randomly and independently from the log-normal distribution, with a mean of 0 and a standard deviation of 0.4. There are five score category difficulty parameters for each item. These parameters are drawn randomly and independently from a standard normal distribution, with a mean of 0 and a standard deviation of 1.0. These five score category difficulty parameters are then sorted in *ascending order* to be the difficulty parameters for scores two to six.

Generating data. Once an item is selected, the response to this item is generated, using a model from equation 1 by the Monte Carlo simulation method. The 50 theta vectors, which are randomly generated from multivariate normal distribution, $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean vector elements equaling zero and diagonal covariance matrix with one diagonal element, are the true theta vectors used for simulation.

Results

Table 2 shows the correlations between the true theta and estimated theta for every dimension, every method, and every condition. For every item selection method, the correlation between the true and the estimate theta is increasing as the test length increases. However, the rate of increase is slowing down: the difference between condition 1 and condition 3 is greater than the difference between condition 3 and condition 5. When comparing three item selection methods under each condition, the multidimensional item selection method has high correlation on the general and second level factors and the unidimensional item selection method has high correlation on the third level factors. The differences among three methods are very large under condition 1, but the differences decrease as the test length increases. Under condition 5, three item selection methods have similarly high correlations on factors. Generally, factors of more general level (like the general factor and the second-level factors with third-level factors) have higher correlations.

Table 3 includes the RMSEs for every dimension, every method, and every condition. For every item selection method, the RMSE decreases as the test length increases. However, the rate of decrease is slowing down: the difference between condition 1 and condition 3 is greater than the difference between condition 3 and condition 5. When comparing the three item selection methods under each condition, the multidimensional item selection method has low RMSE on the general and second level factors and the unidimensional item-selection method has low RMSE on the third-level factors. There are differences among the three methods under condition 1, but the differences decrease as the test length increases. Under condition 5, the multidimensional item selection method and the unidimensional item selection method have the

low and similar RMSEs on factors. Generally, factors of more general level (like the general factor and the second-level factors with third-level factors) have smaller RMSEs.

Table 4 shows the determinant of the covariance matrix for every method and every condition. For every item selection method, the determinant of the covariance matrix decreases as the test length increases. However, the rate of decrease is slowing down: the difference between condition 1 and condition 3 is greater than the difference between condition 3 and condition 5. When comparing the three item selection methods under each condition, the multidimensional item selection method has the lowest determinant under conditions 3 to 5, and the unidimensional item selection method has the lowest determinant under conditions 1 and 2.

Significance

In this study, we compared three item selection methods. Results indicated that the multidimensional item selection method provides more accurate estimation on factors of more general, and the unidimensional item selection method provides more accurate estimation on specific factors; however, the method differences are very small under long test length conditions. If a long test is preferred, there is no difference on these three methods. The unidimensional item selection method should be used for administration of a short test where accurate specific factor estimation is desired. On the other hand, if accurate general estimation is preferred, the multidimensional item selection method should be used. Another finding is that the medium-length 45-item test, condition 3 in the simulation study, can provide sufficient accuracy of theta estimation. In the real testing program, the algorithm is usually implemented on a local server, which might not be able to conduct complicated calculation. Compared with the other two methods, the unidimensional item selection method requiring the easiest calculation is much easier to implement and saves time in a real test environment.

References

Gibbons R. D., Bock, R. D., Hedeker D, Weiss, D. J, Segawa, E., Bhaumik, D. K., ... Ellen Frank, V. J. (2007). Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19.

Gibbons, R. D., & Hedeker, D. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423-436.

Yung, Y.-F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika*, *64*, 113–128.

Samejima, F. (1969). Estimation of a latent ability using a response pattern of graded scores. *Psychometrika Monographs*, 34 (Suppl. 4).

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331-354.

Seo, D. G. (2011). *Application of the bifactor model to computerized adaptive testing* (Unpublished doctoral dissertation). University of Minnesota, Minneapolis. Available from http://conservancy.umn.edu/bitstream/101923/1/Seo_umn_0130E_11740.pdf.

Table 1

Number of Stages and Items for Each Condition

Condition	Number of Stages	Number of Items
1	1	15
2	2	30
3	3	45
4	4	60
5	5	75

Table 2

The Correlation Between the True and Estimated Theta of Each Dimension for Each Method and Each Condition

	Condition 1			Condition 2			Condition 3			Condition 4			Condition5		
Factor	M[a]	LM[b]	U[c]	M[a]	LM[b]	U[c]	M[a]	LM[b]	U[c]	M[a]	LM[b]	U[c]	M[a]	LM[b]	U[c]
G[d]	0.87	0.87	0.85	0.91	0.94	0.87	0.94	0.96	0.90	0.95	0.96	0.92	0.95	0.97	0.93
S[e]1	0.66	0.65	0.65	0.76	0.67	0.75	0.79	0.70	0.79	0.81	0.74	0.82	0.83	0.80	0.83
S[e]2	0.89	0.89	0.57	0.91	0.92	0.69	0.92	0.94	0.77	0.93	0.94	0.85	0.94	0.95	0.87
S[e]3	0.43	0.43	0.45	0.82	0.81	0.59	0.84	0.83	0.70	0.85	0.84	0.74	0.86	0.85	0.81
S[e]4	0.75	0.75	0.58	0.81	0.83	0.75	0.82	0.86	0.81	0.85	0.87	0.84	0.87	0.88	0.85
T[f]1	0.28	0.28	0.59	0.65	0.63	0.74	0.75	0.75	0.78	0.79	0.77	0.79	0.81	0.80	0.80
T[f]2	0.74	0.74	0.71	0.83	0.75	0.80	0.84	0.77	0.83	0.85	0.80	0.85	0.86	0.84	0.85
T[f]3	0.36	0.35	0.67	0.69	0.63	0.69	0.75	0.68	0.75	0.78	0.73	0.78	0.79	0.75	0.80
T[f]4	0.69	0.69	0.70	0.74	0.73	0.77	0.78	0.74	0.79	0.81	0.76	0.83	0.83	0.77	0.84
T[f]5	0.29	0.26	0.68	0.69	0.69	0.78	0.78	0.73	0.82	0.82	0.79	0.84	0.84	0.82	0.85
T[f]6	0.16	0.17	0.56	0.60	0.59	0.67	0.70	0.69	0.71	0.74	0.72	0.75	0.77	0.74	0.75
T[f]7	0.34	0.35	0.57	0.48	0.48	0.67	0.65	0.66	0.70	0.72	0.72	0.73	0.76	0.76	0.75
T[f]8	0.37	0.36	0.57	0.49	0.47	0.64	0.65	0.51	0.66	0.69	0.64	0.68	0.72	0.70	0.72
T[f]9	0.55	0.55	0.54	0.62	0.61	0.69	0.72	0.68	0.74	0.76	0.74	0.77	0.78	0.78	0.80
T[f]10	0.49	0.50	0.54	0.56	0.57	0.66	0.66	0.58	0.70	0.72	0.68	0.73	0.75	0.72	0.75
T[f]11	0.59	0.61	0.58	0.73	0.64	0.73	0.78	0.72	0.78	0.82	0.76	0.82	0.84	0.81	0.84
T[f]12	0.43	0.43	0.72	0.72	0.58	0.75	0.78	0.75	0.79	0.83	0.81	0.84	0.86	0.85	0.85
T[f]13	0.22	0.21	0.69	0.70	0.68	0.81	0.80	0.77	0.83	0.83	0.83	0.84	0.85	0.85	0.85
T[f]14	0.61	0.62	0.61	0.72	0.62	0.73	0.78	0.71	0.78	0.79	0.74	0.81	0.81	0.78	0.82

[a]Multidimensional item-selection method; [b]Local multidimensional item-selection method; [c]Unidimensional item-selection method; [d]General factor; [e]Second-level factor; [f]Third-level factor.

Table 3

The RMSE of Each Dimension for Each Method and Each Condition

	Condition 1			Condition 2			Condition 3			Condition 4			Condition5		
Factor	M[a]	LM[b]	U[c]	M[a]	LM[b]	U[c]	M[a]	LM[b]	U[c]	M[a]	LM[b]	U[c]	M[a]	LM[b]	U[c]
G[d]	0.61	0.62	0.67	0.52	0.44	0.61	0.45	0.37	0.55	0.40	0.34	0.48	0.37	0.31	0.44
S[e]1	0.73	0.74	0.75	0.64	0.73	0.65	0.60	0.70	0.60	0.57	0.65	0.57	0.55	0.59	0.54
S[e]2	0.61	0.61	1.03	0.54	0.51	0.90	0.49	0.45	0.78	0.45	0.42	0.66	0.41	0.39	0.61
S[e]3	0.96	0.96	0.95	0.62	0.64	0.87	0.59	0.59	0.77	0.57	0.58	0.72	0.55	0.57	0.62
S[e]4	0.64	0.64	0.80	0.57	0.53	0.64	0.55	0.49	0.57	0.50	0.47	0.53	0.48	0.46	0.51
T[f]1	1.01	1.02	0.86	0.81	0.83	0.72	0.70	0.71	0.67	0.65	0.69	0.66	0.62	0.65	0.64
T[f]2	0.72	0.72	0.75	0.60	0.70	0.63	0.57	0.67	0.59	0.55	0.64	0.56	0.53	0.57	0.55
T[f]3	0.88	0.88	0.69	0.68	0.73	0.68	0.62	0.69	0.63	0.59	0.64	0.58	0.57	0.61	0.56
T[f]4	0.80	0.80	0.80	0.75	0.76	0.72	0.69	0.75	0.69	0.65	0.72	0.63	0.62	0.70	0.61
T[f]5	0.85	0.85	0.64	0.64	0.63	0.55	0.55	0.60	0.50	0.50	0.54	0.47	0.48	0.50	0.46
T[f]6	1.00	1.00	0.84	0.81	0.82	0.75	0.73	0.73	0.71	0.68	0.70	0.68	0.64	0.68	0.67
T[f]7	0.91	0.91	0.80	0.85	0.85	0.72	0.74	0.72	0.69	0.67	0.67	0.66	0.63	0.63	0.64
T[f]8	0.84	0.85	0.74	0.78	0.80	0.69	0.68	0.78	0.68	0.66	0.69	0.66	0.63	0.65	0.62
T[f]9	0.71	0.71	0.72	0.67	0.68	0.62	0.59	0.63	0.57	0.55	0.57	0.55	0.53	0.54	0.52
T[f]10	0.85	0.85	0.81	0.80	0.79	0.73	0.72	0.78	0.69	0.67	0.70	0.66	0.63	0.66	0.64
T[f]11	0.91	0.90	0.91	0.76	0.86	0.76	0.70	0.78	0.70	0.64	0.72	0.65	0.61	0.66	0.61
T[f]12	1.07	1.07	0.85	0.84	0.97	0.81	0.76	0.80	0.75	0.68	0.71	0.67	0.63	0.65	0.63
T[f]13	1.04	1.04	0.77	0.76	0.79	0.64	0.64	0.68	0.59	0.59	0.60	0.58	0.57	0.57	0.56
T[f]14	0.72	0.71	0.72	0.62	0.71	0.62	0.56	0.63	0.57	0.55	0.61	0.53	0.53	0.56	0.51

[a]Multidimensional item-selection method; [b]Local multidimensional item-selection method; [c]Unidimensional item-selection method; [d]General factor; [e]Second-level factor; [f]Third-level factor.

Table 4

The Determinant of the Covariance Matrix for Every Method and Every Condition

Condition	Multidimensional	Local Multidimensional	Unidimensional
1	4.86E-05	4.60E-05	3.95E-06
2	3.65E-08	3.91E-07	2.26E-08
3	7.26E-10	1.17E-08	1.06E-09
4	5.62E-11	8.46E-10	1.08E-10
5	8.88E-12	8.85E-11	2.08E-11