Fall 10-21-2011

Item Response Theory (IRT) Analysis of Item Sets

Liwen Liu
University of Illinois at Urbana-Champaign, liwenliu36@gmail.com

Fritz Drasgow
University of Illinois at Urbana-Champaign, fdrasgow@cyrus.psych.illinois.edu

Rosemary Reshetar
The College Board, rreshetar@collegeboard.org

YoungKoung Rachel Kim
The College Board, rkim@collegeboard.org

Running head: IRT ANALYSIS OF ITEM SETS

Item Response Theory (IRT) Analysis of Item Sets

Liwen Liu

Fritz Drasgow

University of Illinois at Urbana-Champaign

Rosemary Reshetar

YoungKoung Rachel Kim

The College Board

Abstract

We examined whether set-based items affected IRT model-data fit. We also evaluated fit after combining dependent items into composites and treating them as polytomous items. Analysis of the 2009 AP English Literature and Composition Exam showed that some of the item pairs had major violations of local independence. Model fit improved when we analyzed the data using composites. Our findings suggest that conducting IRT analyses on composites provides a viable approach to circumventing problems of local dependence for set-based items.

More than three million Advanced Placement Program® (AP®) Exams are taken annually by high school students. These exams include constructed response (CR) items as well as multiple-choice (MC) items. Often the MC items use an "item set" format where several questions refer to the same stimulus material. For example, an assessment of listening comprehension of a Spanish language exam may utilize the same listening material followed by three or more questions. However, item sets may violate the fundamental local independence assumption of unidimensional item response theory (IRT) due to their shared content (Wainer & Kiely, 1987).

Item response theory (IRT) provides elegant solutions to many measurement problems (e.g., invariant item parameters, ability scores that can be computed from different sets of items), but it is not clear that an IRT model such as the three-parameter logistic is appropriate for tests with item sets. One of the fundamental assumptions of this model is local independence: after controlling for an examinee's ability, item responses should be statistically independent. However, it is possible that a correct response to one item on a set implies a higher probability of a correct answer to another item from that set. Therefore, in the current study we examined the extent to which set-based items led to violations of IRT assumptions. We also evaluated a method for combining items that violate local independence.

IRT and Model Fit

For MC items, the three-parameter logistic model (3PLM; Birnbaum, 1968) was used:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}},$$

where $P_i(\theta)$ is the probability that an examinee with ability θ answers item *i* correctly, a_i is the discrimination parameter for item *i*, b_i is the difficulty parameter, c_i is the "pseudo-guessing" parameter, and *D* is a constant set equal to 1.7.

For items with multiple ordered categories, we used Samejima's Graded Response Model (SGRM; Samejima, 1969). It uses two-parameter logistic response functions to model the probability that an examinee obtained a score of *k* or higher versus *k-1* or lower,

$$\tilde{P}_{ik}(\theta) = \frac{1}{1+e^{-Da_i(\theta - b_{ik})}},$$

where *k* is the *k*th response option of item *i* and b_{ik} is threshold parameter for option *k*. Then the probability of responding in category *k* is

$$P_{ik}(\theta) = \tilde{P}_{ik}(\theta) - \tilde{P}_{i(k+1)}(\theta).$$

We used a χ^2 statistic to assess goodness of fit. It summarizes the differences between the model-expected and observed frequencies of right and wrong responses. The expected frequency of a correct response to an individual MC item can be written as

$$E(u_i = 1) = N \int P_i(t) f(t) dt,$$

where $P_i(t)$ is the 3PLM probability of a correct response and $f(t)$ is the probability density function of the latent trait. Unfortunately, the χ^2 statistic for individual items allows compensation between local misfits, so χ^2 statistics were also computed for item pairs and triples. Finally, all χ^2 statistics were adjusted to what would be expected in a sample size of 3000 and

then divided by their degrees of freedom (*df*). Drasgow, Levine, Tsien, Williams and Mead (1995) suggested that values of adjusted χ^2/df smaller than 3 indicate good model-data fit.

Method

We analyzed data from the 2009 AP English Literature and Composition Exam with a random sample of 20,000 examinees. There were 55 passage-based set items with 8 to 15 MC items in each of the 5 item sets. In the original data set, there were 10 categories (i.e., 0-9 points) for the 3 CR items. To obtain more accurate estimation, we collapsed the 10 categories into 7 to ensure enough responses in each category.

We first ran MULTILOG 7.03 (Thissen, Chen, & Bock, 2003) to estimate the 3PLM and SGRM parameters for 55 MC and 3 CR items simultaneously. Then for the MC items, we used the MODFIT program (Stark, 2007) to examine the model fit of individual items, item pairs and item triples. We combined the MC item pairs whose adjusted χ^2/df values were larger than 3, which indicates a violation of local independence. These composites were modeled by the SGRM and the MC items not included in any item set, referred to as "discrete MC items," were modeled by the 3 PLM. After re-estimating item parameters, we again examined model fit. Any improvement of adjusted χ^2/df statistics suggests that item composites reduce problems of local dependence.

Results

Initial model fit analyses showed that some MC item pairs had major violations of local independence (see Table 1 for the adjusted χ^2/df values for a sampling of all the possible combinations of item pairs). For example, items 36 and 37 (i.e., pair 39) had an adjusted χ^2/df value of 43.43, items 43 and 44 (i.e., pair 46) had an adjusted χ^2/df value of 9.94, and items 11

and 12 (i.e., pair 12) had an adjusted χ^2/df value of 7.16. These values are much larger than the cutoff value of 3 and thus indicate violation of local independence. Table 2 shows a summary of the adjusted χ^2/df values for item singlets, item doublets, and item triplets. For the item doublets, there were three pairs with adjusted χ^2/df values over 7. Although the mean of the adjusted χ^2/df values is acceptable, the large SD confirms the existence of extreme adjusted χ^2/df values.

Also, note that all these item pairs with large adjusted χ^2/df values are composed of adjacent items and belong to the same item set (the items refer to a common stimulus). Instead of forming item sets based on the content of the items, we used an empirical approach. Based on the χ^2/df statistics for item pairs, we formed 9 composites, each containing 3 to 6 MC items. For example, both the pair of Item 46 and 49 and Item 49 and 52 had adjusted χ^2/df values larger than 3. Therefore we combined these three items (i.e., Item 46, 49, and 52) into an item set, which can be considered as a polytomous item with four categories (i.e., 0, 1, 2, 3). A value of zero indicates that none of the items was answered correctly while a value of 3 indicates that all the three items responses were correct. If any of the items in the item set had a missing value, we coded the response to the whole item set as missing.

Then we re-estimated parameters for the 9 composites and the remaining 26 discrete MC items simultaneously, and re-evaluated model fit. Table 3 shows the χ^2/df statistics for all 36 pairs of the 9 item composites. Most pairs had an adjusted χ^2/df value smaller than 3, only one item set pair had a value of over 6, and no extremely large values were found. Although the mean of the adjusted χ^2/df values only decreased slightly from 2.431 to 2.284, the SD decreased from 5.748 to 0.972 (see Table 4). This again indicates that there were no extremely large adjusted χ^2/df values for the model fit indices of the item set. Therefore, improved model fit was obtained by analyzing item composites.

Finally, we examined the correspondence between ability scores estimated in the original format (i.e., individual MC items) and ability scores estimated using the item composites and discrete MC items. The correlation was .990 for IRT ability score estimates. We also correlated the estimated standard errors for these two ability estimates; this correlation was .815. As expected, the average estimated standard error was slightly higher for ability estimates obtained in the analysis of item sets ($M = 0.378$) than for the original analysis ($M = 0.352$). This finding is logical because violations of local independence of items in the original format would result in an artificially inflated test information function and artificially reduced standard errors.

Conclusions

We found that some of the set-based items exhibited major violations of local independence. By combining dependent items, we obtained a better model fit. The correspondence of examinees' ability score estimates using the two item formats was very high.

Although we found some pairs of items from a given item set to substantially violate local independence, many other pairs of items did not. This is important because the pairs violating local independence are, at least statistically speaking, overly redundant and the second item provides little incremental information about the latent trait. Items satisfying local independence are not overly redundant and thus provide a more efficient use of testing time. It would be very helpful to have subject matter experts identify the differences between the problematic pairs of items and the remaining items that satisfy local independence. The next step in this line of research is for us to conduct this review and incorporate the findings into our test development work. If item writers can be trained to create set-based items with no local dependence, testing time could be utilized more effectively.

References

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

Drasgow, F., Levine, M. V., Tsien, S., Williams, B. A., & Mead, A. D. (1995). Fitting polytomous item response models to multiple-choice tests. *Applied Psychological Measurement, 19*, 145-165.

Samejima, F. (1969). Estimation of latent ability using response pattern of graded scores. *Psychometrika Monograph, 17*, 1-100.

Stark, S. (2007). MODFIT: Plot theoretical item response functions and examine the fit of dichotomous or polytomous IRT models to response data [computer program]. Champaign, IL: University of Illinois at Urbana-Champaign.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). MULTILOG 7 for Windows: Multiple-category item analysis and test scoring using item response theory [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

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

Table 1. Model fit statistics for individual MC item pairs of the AP English Literature and Composition Exam 2009.

Pairs	Item1	Item2	N	*df*	χ^2	χ^2/df	Adjusted χ^2	Adjusted χ^2/df
1	1	2	19358	3	12.625	4.208	4.492	1.497
2	1	5	19195	3	4.533	1.511	3.240	1.080
3	1	55	17559	3	8.116	2.705	3.874	1.291
4	2	5	18944	3	4.826	1.609	3.289	1.096
5	2	55	17334	3	5.990	1.997	3.517	1.172
6	5	55	17192	3	17.746	5.915	5.573	1.858
7	3	4	19643	3	12.472	4.157	4.447	1.482
8	3	7	19705	3	31.000	10.333	7.263	2.421
9	4	7	19505	3	0.426	0.142	2.604	0.868
10	6	11	19015	3	4.585	1.528	3.250	1.083
11	6	12	18091	3	5.686	1.895	3.445	1.148
12	11	12	17910	3	113.278	37.759	21.472	7.157
13	8	15	18749	3	1.342	0.447	2.735	0.912
14	8	16	19047	3	4.407	1.469	3.222	1.074
15	15	16	18429	3	12.494	4.165	4.545	1.515
16	9	17	19444	3	3.674	1.225	3.104	1.035
17	9	26	19454	3	1.893	0.631	2.829	0.943
18	17	26	19105	3	6.957	2.319	3.621	1.207
19	10	20	17856	3	0.829	0.276	2.635	0.878
20	10	27	17821	3	2.165	0.722	2.859	0.953
21	20	27	17469	3	0.452	0.151	2.562	0.854
22	13	21	18449	3	3.676	1.225	3.110	1.037
23	13	29	19200	3	5.491	1.830	3.389	1.130
24	21	29	18275	3	2.034	0.678	2.841	0.947
25	14	23	17620	3	1.707	0.569	2.780	0.927
26	14	32	17609	3	8.559	2.853	3.947	1.316
27	23	32	16910	3	2.631	0.877	2.935	0.978
28	18	28	18139	3	2.790	0.930	2.965	0.988
29	18	33	18803	3	7.686	2.562	3.748	1.249
30	28	33	18298	3	6.470	2.157	3.569	1.190
31	19	30	19192	3	3.527	1.176	3.082	1.027
32	19	34	19385	3	9.588	3.196	4.020	1.340
33	30	34	18811	3	0.558	0.186	2.611	0.870
34	22	31	17780	3	5.906	1.969	3.490	1.163
35	22	35	19032	3	4.892	1.631	3.298	1.099
36	31	35	17459	3	4.733	1.578	3.298	1.099
37	24	36	18860	3	5.184	1.728	3.347	1.116
38	24	37	18677	3	2.169	0.723	2.866	0.955
39	36	37	18696	3	796.243	265.414	130.285	43.428
40	25	38	19091	3	1.465	0.488	2.759	0.920
41	25	40	18349	3	3.511	1.170	3.084	1.028

42	38	40	18423	3	7.617	2.539	3.752	1.251
43	39	41	18879	3	46.629	15.543	9.933	3.311
44	39	42	18295	3	9.909	3.303	4.133	1.378
45	41	42	18529	3	20.785	6.928	5.880	1.960
46	43	44	18731	3	170.471	56.824	29.823	9.941
47	43	45	18754	3	39.353	13.118	8.815	2.938
48	44	45	18525	3	36.368	12.123	8.404	2.801
49	46	47	18435	3	4.792	1.597	3.292	1.097
50	46	48	18512	3	31.894	10.631	7.683	2.561
51	47	48	19021	3	4.460	1.487	3.230	1.077
52	49	50	17802	3	4.115	1.372	3.188	1.063
53	49	51	17613	3	6.371	2.124	3.574	1.191
54	50	51	18745	3	26.487	8.829	6.759	2.253
55	52	53	16598	3	31.108	10.369	8.080	2.693
56	52	54	16514	3	17.278	5.759	5.594	1.865
57	53	54	17908	3	89.945	29.982	17.565	5.855

Note. N = 20,000. These results were obtained from MODFIT (Stark, 2007) output.

Table 2. Summary of model fit statistics for individual MC items of the AP English Literature and Composition Exam 2009.

FREQUENCY TABLE OF ADJUSTED (N=3000) CHISQUARE/DF RATIOS									
	<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	SD
Singlets	35	17	3	0	0	0	0	1.097	0.356
Doublets	13	33	6	1	0	1	3	2.431	5.748
Triplets	1	15	1	2	0	1	1	2.707	4.037

Note. N = 20,000. These results were obtained from MODFIT (Stark, 2007) output.

Table 3. Model fit statistics for item composite pairs of the AP English Literature and Composition Exam 2009.

Pairs	Item1	Item2	N	*df*	χ^2	χ^2/df	Adjusted χ^2	Adjusted χ^2/df
1	1	2	17028	14	57.470	4.105	21.659	1.547
2	1	3	18089	14	154.269	11.019	37.263	2.662
3	1	4	15389	29	76.013	2.621	38.165	1.316
4	1	5	17103	19	103.618	5.454	33.843	1.781
5	1	6	17364	19	54.683	2.878	25.165	1.324
6	1	7	16846	24	265.443	11.060	66.997	2.792
7	1	8	14818	19	83.748	4.408	32.109	1.690
8	1	9	16183	19	70.727	3.722	28.589	1.505
9	2	3	17323	8	76.139	9.517	19.800	2.475
10	2	4	14843	17	51.592	3.035	23.992	1.411
11	2	5	16416	11	57.958	5.269	19.582	1.780
12	2	6	16633	11	31.792	2.890	14.750	1.341
13	2	7	16144	14	150.958	10.783	39.451	2.818
14	2	8	14279	11	101.750	9.250	30.066	2.733
15	2	9	15493	11	43.880	3.989	17.367	1.579
16	3	4	15590	17	195.875	11.522	51.421	3.025
17	3	5	17433	11	197.513	17.956	43.096	3.918
18	3	6	17718	11	104.013	9.456	26.749	2.432
19	3	7	17187	14	415.231	29.659	84.035	6.003
20	3	8	15031	11	98.397	8.945	28.443	2.586
21	3	9	16505	11	93.532	8.503	26.001	2.364
22	4	5	14968	23	241.176	10.486	66.728	2.901
23	4	6	15103	23	81.619	3.549	34.644	1.506
24	4	7	14735	29	322.094	11.107	88.673	3.058
25	4	8	13226	23	98.388	4.278	40.100	1.743
26	4	9	14029	23	71.060	3.090	33.277	1.447
27	5	6	16756	15	59.837	3.989	23.028	1.535
28	5	7	16246	19	285.609	15.032	68.232	3.591
29	5	8	14420	15	83.526	5.568	29.256	1.950
30	5	9	15588	15	72.632	4.842	26.092	1.739
31	6	7	16670	19	182.304	9.595	48.389	2.547
32	6	8	14589	15	59.757	3.984	24.204	1.614
33	6	9	15849	15	43.049	2.870	20.309	1.354
34	7	8	14258	19	255.410	13.443	68.743	3.618
35	7	9	15430	19	195.449	10.287	53.306	2.806
36	8	9	14451	15	67.794	4.520	25.960	1.731

Note. N = 20,000. These results were obtained from MODFIT (Stark, 2007) output.

Table 4. Summary of model fit statistics for item composites of the English Literature and Composition Exam 2009.

FREQUENCY TABLE OF ADJUSTED (N=3000) CHISQUARE/DF RATIOS									
	<1	1<2	2<3	3<4	4<5	5<7	>7	Mean	SD
Singlets	1	7	1	0	0	0	0	1.315	0.519
Doublets	0	19	11	5	0	1	0	2.284	0.972
Triplets	0	65	16	3	0	0	0	1.834	0.463

Note. N = 20,000. These results were obtained from MODFIT (Stark, 2007) output.