Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 56

Illustration of Factors Affecting Performance of the $S - X^2$ Item-Fit Index

Hyung Jin Kim†
Won-Chan Lee

June 2023

† Hyung Jin Kim is Associate Research Scientist, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: hyungjin-kim@uiowa.edu). Won-Chan Lee is Director, Center for Advanced Studies in Measurement and Assessment (CASMA), College of Education, University of Iowa (email: won-chan-lee@uiowa.edu).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5954
Web: https://education.uiowa.edu/casma

Contents

Abstract

Orlando and Thissen (2000) introduced the $S - X^2$ item-fit index by utilizing the likelihoods of number-correct (NC) scores derived from the dichotomous item response theory (IRT) model. For each possible NC score, the index provides a direct comparison between observed and model-based expected proportions for correct and incorrect responses. According to Kim and Lee (2022), there exists an alternative approach for computing $S - X^2$ values. Kim and Lee (2022) also identified that there are multiple procedures for collapsing a contingency table of observed and expected proportions of examinees (i.e., OE table) that is often necessary to remove sparseness in OE tables prior to computing $S - X^2$ values. This study compares various procedures for collapsing OE tables and handling score categories when collapsing. Additionally, step-by-step instructions are provided for each procedures, offering clear guidance for researchers and practitioners. Study results showed that, for real data, conclusions based on $S - X^2$ could depend on computational approaches and procedures for collapsing OE tables. The study also showed that a choice of a minimum cell value for collapsing OE tables was an important factor that affected the results of $S - X^2$.

1 Computation of $S - X^2$ Values

According to Kim and Lee (2022), there are at least two different approaches for computing the $S - X^2$ index.

1.1 Orlando & Thissen (2000)

Orlando and Thissen (2000) introduced two goodness-of-fit indices for dichotomous IRT models, $S - X^2$ and $S - G^2$. These indices utilize the likelihoods of number-correct (NC) scores derived from the IRT model and provide a direct comparison between observed and model-based expected proportions for correct and incorrect responses for each possible NC score.

According to Orlando and Thissen (2000), an expected proportion of examinees with NC score k who respond correctly to item i is defined as follow:

$$E_{ik1}^{P} = \frac{\sum_{q=1}^{Q} T_i(\theta_q) S_{k-1}^{*i}(\theta_q) w(\theta_q)}{\sum_{q=1}^{Q} S_k(\theta_q) w(\theta_q)}, \tag{1}$$

where the subscript 1 refers to score category 1 (i.e., answering item i correctly); θ_q refers to the q^{th} quadrature point where $q = 1, ..., Q$; $w(\theta_q)$ is the quadrature weight associated with θ_q; $T_i(\theta)$ is the probability that examinees with ability θ_q respond correctly to item i; $S_{k-1}^{*i}(\theta_q)$ is the probability that examinees with ability θ_q obtain NC score $k-1$ without item i; and, $S_k(\theta_q)$ is the probability that examinees with ability θ_q obtain NC score k including item i[1].

The item-fit index $S - X^2$ for item i has the form

$$S - X_i^2 = \sum_{k=1}^{n-1} \sum_{z=0}^{1} N_k \frac{(O_{ikz}^{P} - E_{ikz}^{P})^2}{E_{ikz}^{P}} \tag{2}$$

where z is the item score (0 or 1); N_k is the number of examinees with NC score k; and O_{ikz}^{P} and E_{ikz}^{P} are observed and expected proportions of examinees with NC score k who obtain score z on item i, respectively.

As depicted in Kim and Lee (2022), Figure 1 displays a table of NC scores by observed/expected proportions for responding correctly and incorrectly to item i. The table of NC scores by observed and expected values responding correctly/incorrectly to item i is referred to as the OE table, hereafter. Rows for the OE table represent NC scores ($k = 1, ..., n-1$) including item i. Note that, in both Equation (2) and Figure 1, NC scores range from 1 to $n - 1$.

[1] See Appendix A for detailed procedures for computing S_k and S_{k-1}^{*i}.

Figure 1: OE Table based on Orlando and Thissen (2000)

		Correct		Incorrect		
	NC Score k	Observed O^P_{ik1}	Expected E^P_{ik1}	Observed 1 - O^P_{ik1}	Expected 1 - E^P_{ik1}	N_k
Possible NC Scores with Item i	1					
	2					
	:					
	$n-1$					

Note: In this OE table, E_{ikz} and O_{ikz} are in proportion-metric.

1.2 Alternative Approach

According to Kim and Lee (2022), the IRTFIT manual (Bjorner, Smith, Stone, & Sun, 2007) suggests an alternative approach for computing expected *numbers* of examinees[2]. It is argued that OE tables constructed according to Orlando and Thissen (2000) have cells with zero probabilities of observing examinees. For NC score of 0, it is impossible to observe examinees responding correctly to item i. Similarly, for a perfect NC score, there should be no examinee who responded incorrectly to item i. In order to avoid having such cells, Bjorner et al. (2007) suggested computing $S - X^2$ values based on OE tables where rows represent NC scores without item i (i.e., $k^{'} = 0, 1, ..., n-1$). More details can be found in Kim and Lee (2022).

It appears that the flexMIRT program and the R `mirt` package use different approaches for computing $S - X^2$ values (Kim & Lee, 2022). The flexMIRT program uses the alternative approach suggested by Bjorner et al. (2007), whereas the R `mirt` package (Chalmers, 2019) uses the original approach introduced by Orlando and Thissen (2000).

2 Other Factors Affecting OE Tables

Besides the two aforementioned approaches for computing $S - X^2$ values, Kim and Lee (2022) considered other factors that can possibly affect results of the $S - X^2$ item-fit index. Those other factors included the procedures for collapsing OE tables and approaches to dealing with score categories when collapsing. Note that all procedures are labeled with the names used in Kim and Lee (2022).

[2]In Orlando and Thissen (2000), observed and expected values are in *proportions*. See Appendix 1 in Kim and Lee (2022) for the modified $S - X^2$ where proportions are replaced by total *numbers*.

2.1 Procedures for Collapsing OE Tables

Expected numbers for correct and incorrect responses are often small for some scores, which could decrease the accuracy of the χ^2 approximation for their distribution. In order to avoid this problem, it is often suggested that cells with small frequencies are collapsed with other cells.

This section describes two different collapsing procedures that are currently available to the public and presents step-by-step procedures with examples. For the examples, the study selected 45 items at random from a large-scale assessment, and calibrated them using a 3 parameter-logistic (3PL) IRT model. In order to avoid any confounded effects arising from different settings, the examples in this study used the original Orlando and Thissen (2000)'s approach for computing $S - X^2$ values and conducted collapsing if *any* of score categories did not satisfy the minimum cell (MIN) value of one.

2.1.1 Mid Procedure

For the mid procedure, collapsing starts from the first and last rows of an OE table and the collapsing process progresses towards the middle of the score list. Suppose that there are n items such that NC scores range from 0 to n (i.e., $k = 0, \ldots\ n$). For item i, the collapsing procedure towards the middle of scores can be described as follows:

1. Define the middle of scores, M. Note that the middle of a score distribution can be defined using the mean, median, or mode.

2. Remove scores without examinees (i.e., $N_k = 0$) from the initial uncollapsed OE table. Let R and r denote the number of rows remaining in the updated OE table and the row index, respectively.

3. For $r = 1$, consider the first and last rows with the row numbers of r and $R-r+1$, respectively, in the updated OE table. Let k_r represent the score associated with the r^{th} row.

   (a) For the r^{th} row, if any of expected numbers of examinees responding correctly/incorrectly to item i (i.e., E_{ir1} and $N_{k_r} - E_{ir1}$) is less than a user-specified MIN value (m), collapse the row with the $(r+1)^{th}$ row .

   (b) At the same time, for the $(R-r+1)^{th}$ row, if any of the expected numbers (i.e., $E_{i(R-r+1)1}$ and $N_{k_{R-r+1}} - E_{i(R-r+1)1}$) is less than m, collapse the row with the $(R-(r+1)+1)^{th}$ row. The OE table should be updated if collapsing occurs in Steps 3(a) and/or 3(b).

4. Increase r by 1 and repeat Step 3 using the updated OE table from the previous step until $k_r > M$ and $k_{R-r} < M$.

Example

Table 1 presents an OE table for Item A after scores with $N_k = 0$ were removed. Note that there were no examinees with NC scores of 0, 1, 2, and 45. Suppose that the mid procedure is used to collapse the OE table for Item A with the MIN value of one. An NC score whose cumulative frequency started exceeding half of the total number of examinees (i.e., median; NC score 25) was defined as the middle of the score list and cells were collapsed towards the score 25.

Based on Table 1, starting from the two end-rows of the OE table, both scores 3 and 44 have at least one expected frequency less than one, suggesting that these rows should be collapsed with the rows for NC scores 4 and 43, respectively. Figure 2(a) presents the OE table after collapsing occurred for those two end-rows. Based on Figure 2(a), scores 4 and 43 still have at least one expected frequency less than one, indicating that those rows should also be collapsed with the rows for scores 5 and 42. Figures 2(b) and 2(c) present the OE tables after the second and third iterations, respectively. The iteration should continue until all cells satisfy the MIN value.

2.1.2 Even Procedure

In the R `mirt` package, the `collapseCells` function collapses OE tables following three main stages. During the first stage, rows for scores with $N_k = 0$ are removed from the OE table. During the second stage, rows for scores with $N_k = 1$ are collapsed with adjacent rows. Note that, for $N_k = 1$, the number of examinees responding incorrectly to item i is zero if the observed number of examinee responding correctly to item i is one, and vice versa. During the third stage, it examines expected numbers and collapses corresponding rows if the expected numbers do not satisfy the MIN value.

The collapsing procedure defined in the `collapseCells` function can be summarized as follows.

1. Remove rows of NC scores with $N_k = 0$ from the initial uncollapsed OE table.

2. Collapse rows of NC scores with $N_k = 1$ with adjacent rows of higher scores. For example, if the row for NC score 1 has $N_1 = 1$, the row is collapsed with a row for a higher score, say NC score 2. However, if the last row for NC score n has $N_n = 1$, the row is collapsed with an adjacent row for a smaller score (e.g., $n - 1$). Steps 1 and 2 result in an updated OE table.

3. For the updated OE table from Step 2, examine expected numbers of examinees responding correctly or incorrectly to item i. Again, let R and r denote the number of rows remaining in the updated OE table and the row index, respectively. And, let k_r represent the score associated with the r^{th} row.

(a) For the first row (i.e., $r = 1$), if any of expected numbers of examinees responding correctly/incorrectly to item i (i.e., E_{ir1} and $N_{k_r} - E_{ir1}$) is less than the MIN value (m), collapse the row with the second row. Collapsing may update the OE table. Using the updated OE table, repeat Step 3(a) until the first row satisfies the MIN value requirement.

(b) For $r = 2$, if any of E_{ir1} and $N_{k_r} - E_{ir1}$ is less than m, collapse the row with one of two adjacent rows that has a smaller number of examinees. Collapsing may update the previous OE table from Step 3(a).

(c) Repeat Step 3(b) until the second row (i.e., $r = 2$) satisfies the MIN value.

(d) Increase r by 1 and repeat Step 3(b)-(c) until all rows except the last row satisfy the MIN value.

(e) For the last row, if any of E_{ir1} and $N_{k_r} - E_{ir1}$ is less than m, the row is collapsed with an adjacent row that is associated with a smaller NC score.

Note that the collapsing procedure described above applies to dichotomous items only. For polytomously-scored items, the collapsing procedure is more complicated than the procedure described above.

For a simpler procedure to collapse OE tables, Step 2 can be omitted. For $N_k = 1$, it is highly likely that an expected number of examinees responding correctly or incorrectly is less than 1. Considering that cells with expected frequencies less than a MIN value are collapsed with one of two adjacent rows, the rows of $N_k = 1$ will be eventually collapsed with an adjacent row when a MIN value ≥ 1. Furthermore, if the MIN value is 0 (i.e., no collapsing), it does not make sense to collapse such scores with $N_k = 1$ with other scores.

Example

For the same example presented above, the simpler version of the even procedure was considered for collapsing OE tables with the MIN value of one. As already mentioned, the simpler procedure skips the second stage and moves on to the third stage directly. Based on Table 1, the expected number of examinees who obtain score 3 and respond correctly to Item A is 0.1275, which is smaller than one; thus, the row for score 3 should be collapsed with the row for score 4. Figure 3(a) presents an OE table after the first iteration and suggests that the row for score 4 should be collapsed with the row for score 5. Based on Figure 3(b) for an OE table after the second iteration, the expected number of examinees who obtain the NC score 5 and respond correctly to Item A is again smaller than 1, suggesting that the row for score 5 should be collapsed with the row for score 6. Figure 3(c) presents an updated OE tables after the third iterations. Based on Figure 3(c), the row for NC score 31 has the expected number smaller than one and should be collapsed with the row for NC score 32 which has a smaller expected number than the one for NC score 30. Figure 3(d) presents an updated OE tables after the fourth

iterations; and, the iteration continues until the MIN value requirement is satisfied for all remaining cells.

2.2 Approaches to Dealing with Score Categories when Collapsing

When collapsing occurs for OE tables, there are two different approaches to dealing with score categories: concurrent and separate (Kim & Lee, 2022). The following subsections describe procedures for the two approaches and present examples using Item B from the 45-item form. For the examples, $S - X^2$ values were computed based on the original computational approach, and cells were collapsed with one of adjacent rows with a smaller number of examinees (i.e., the even procedure).

2.2.1 Concurrent Collapsing

For concurrent collapsing, all score categories (e.g., 0 and 1 for dichotomous items) are collapsed simultaneously if there is at least one score category whose expected number does not satisfy the MIN value. Consequently, the completion of concurrent collapsing results in OE tables with the same number of remaining rows for all score categories. $S - X^2$ values are computed separately for each score category; and the overall $S - X^2$ is the sum of the $S - X^2$ values across all score categories. The *df* is the number of remaining rows in the OE table after collapsing minus the number of item parameters.

Example

As an example, concurrent collapsing is applied to the score categories of 1 (i.e., correct) and 0 (i.e., incorrect) for Item B. Table 2 presents a final OE table after concurrent collapsing was completed with the MIN value of 5. As noted earlier, the number of remaining rows is the same for both score categories. As a result, the $S - X^2$ values for the score categories of 1 and 0 were 17.294 and 24.642, respectively. The overall $S - X^2$ was 41.935 with 28 (= 31 - 3) for the *df*. This resulted in a *p*-value of 0.044

2.2.2 Separate Collapsing

Collapsing can also be conducted separately for each score category. The overall $S - X^2$ is still the sum of $S - X^2$ values across all score categories. After separate collapsing is completed for each score category, it is possible that the number of remaining rows in the OE table differs across different score categories. Thus, for separate collapsing, the *df* associated with the overall $S - X^2$ equals the number of remaining NC scores across *all* score categories minus the number of item parameters, which is equivalent to how the *df* for concurrent collapsing is obtained.

Example

For the same Item B, separate collapsing was applied with the MIN value of 5. Table 3 presents the final OE table for the score categories of 0 and 1. Since collapsing was

completed separately for each score category, the number of rows for the remaining NC scores could differ between the two score categories. In cases where cells were already collapsed with one of the adjacent cells, they were denoted as NA in Table 3.

As a result, the $S - X^2$ values for the score categories of 1 and 0 were 17.338 and 24.502, respectively; and the overall $S - X^2$ was 41.839. Since there were 32 NC scores remaining for both score categories after collapsing, the *df* became 29 (= 32 - 3). The final p-value of 0.058 indicates that the item should not be flagged for misfit at the 5% significance level. Note that the item was flagged for misfit when concurrent collapsing was applied with the same MIN value.

The R `mirt` package (Chalmers, 2019) uses concurrent collapsing and cells are collapsed with one of the two adjacent rows with a smaller number of examinees. However, the flexMIRT program collapses OE tables towards the middle of scores and collapsing is done separately for each score category. In order to investigate the potential impact of different choices in computational approach, collapsing procedure, and approach to dealing with score categories when collapsing on the conclusions based on $S - X^2$, a small study was conducted using real data.

3 Method

The study conducted a series of three small comparisons between (1) different approaches for computing $S - X^2$ values, (2) procedures for collapsing OE tables, and (3) approaches to dealing with score categories when collapsing. For constructing OE tables and obtaining observed and expected values, the study considered two procedures: the original approach (Orlando & Thissen, 2000) and the alternative approach (Bjorner et al., 2007). For collapsing cells in OE tables, the study considered two procedures: the mid and even procedures. And, for dealing with score categories when collapsing, the study considered the concurrent and separate approaches.

Other than the study factors of interest, conditions were set to default settings of the original computational approach, the even procedure for collapsing OE tables, and the concurrent collapsing procedure to deal with score categories. For example, when the original and alternative approaches were compared for computing $S - X^2$ values, the study used the even procedure for collapsing OE tables and concurrent collapsing to deal with score categories. Similarly, for comparing the two collapsing procedures (mid vs. even), the study used the original computational approach and collapsed OE tables concurrently for all score categories.

For comparison purposes, this study selected one operational form from a large-scale assessment. Item parameters were estimated for the 3PL IRT model using flexMIRT.

Note that separate R functions were developed to accommodate different procedures for computing $S - X^2$ values, collapsing OE tables, and dealing with score categories. The performance of these functions was verified by comparing their results with those obtained using the flexMIRT program and the `mirt` package.

In the three comparison studies, two conditions for the MIN value were considered for collapsing OE tables: 1 and 5. The MIN value of 1 was used as the smallest possible value, and the study also considered the MIN value of 5 as suggested by Cochran (1952). Results were compared in terms of $S - X^2$, df, and p-value.

4 Results

This section consists of three subsections. The first subsection compares results for different computational approaches, followed by a subsection comparing results for different approaches for collapsing OE tables. The last subsection presents results for different approaches to dealing with score categories when collapsing OE tables.

4.1 Comparison of Computational Approaches

Table 4 presents the $S - X^2$, df, and p-value for both the original and alternative computational approaches. Note that the results in Table 4 were obtained using the MIN value of 1. The first column represents item numbers; the second to fourth columns display the $S - X^2$, df, and p-value using the original approach; and the fifth to seventh columns display the $S - X^2$, df, and p-value using the alternative approach. The light grey color refers to items flagged for misfit at the 5% significance level only, while the dark grey color refers to items flagged for misfit at both 1% and 5% significance levels.

Based on Table 4, it can be observed that the alternative approach tended to yield smaller p-values compared to the original approach, indicating that the alternative approach flagged more items. Indeed, the flagged items for misfit were not the same when comparing the two computational approaches, especially at the 5% significance level. At the 5% significance level with the MIN value of 1, the original and alternative approaches flagged 10 and 18 items, respectively; and, among those flagged items, eight were flagged by both approaches.

Table 5 presents the $S - X^2$, df, and p-value for the two computational approaches when the MIN value of 5 was used for collapsing. Similar patterns were observed. The two approaches flagged different sets of items for misfit, and the alternative approach flagged eight more items for misfit at the 5% significance level. Note, however, that the sets of flagged items using the MIN value of 5 were somewhat different from those flagged using the MIN value of 1, suggesting that conclusions based on $S - X^2$ depend on the choice of MIN value for collapsing OE tables as well as the computational approaches.

4.2 Comparison of Procedures for Collapsing OE Tables

Table 6 presents the $S - X^2$, *df*, and *p*-value for the mid and even collapsing procedures with the MIN value of 1, and Table 7 presents results for the MIN value of 5. In both tables, the first column represents item numbers; the second to fourth columns present the $S - X^2$, *df*, and *p*-value results obtained using the mid procedure; and the fifth to seventh columns present results for the even procedure. Similar to previous tables, the light and dark grey colors refer to items flagged for misfit at the 5% significance level only and at both 1% and 5% significance levels, respectively.

Results in Tables 6 and 7 suggest that differences in $S - X^2$ and *df* values were smaller when comparing the two collapsing procedures than the differences observed between the two computational approaches. However, there were still differences in the flagged items for misfit between the two collapsing procedures. Using the MIN value of 1, the mid procedure flagged two more items for misfit than the even procedure did. However, with the MIN value of 5, the number of flagged items was larger by one for the even procedure compared to the mid procedure. Furthermore, the sets of flagged items using the MIN value of 5 were not the same as those flagged using the MIN value of 1.

4.3 Comparison of Approaches to Dealing with Score Categories when Collapsing

Results for the two approaches to dealing with score categories are summarized in Tables 8 and 9. Table 8 presents the $S - X^2$, *df*, and *p*-value when the MIN value of 1 was used for collapsing, while Table 9 shows results using the MIN value of 5. The columns and shaded cells in Tables 8 and 9 represent the same meaning as mentioned earlier.

Based on Table 8 for the MIN value of 1, the items flagged for misfit were not the same for the two approaches. The separate approach flagged one more item for misfit at the 5% significance level. When the MIN value of 5 was used for collapsing, the two approaches flagged the same items for misfit. However, the $S - X^2$, *df*, and *p*-value were not exactly the same for the two approaches. This suggests that the two approaches could still flag different items for misfit, particularly for items with *p*-values close to the threshold of being flagged (e.g., 0.01 or 0.05). Indeed, the simulation study conducted by Kim and Lee (2022) found notable differences in results between the concurrent and separate approaches. Similar to the findings presented in the previous sections, using different MIN values flagged different sets of items for misfit.

5 Summary and Conclusions

Kim and Lee (2022) noted that different software programs implement different approaches for computing $S - X^2$ values, collapsing OE tables, and dealing with score categories when collapsing. The R `mirt` package implements the Orlando and Thissen (2000)'s approach for computing $S - X^2$ values. Cells in OE tables are collapsed with one of the two adjacent cells that has a smaller number of examinees, and collapsing is done concurrently for all score categories. Whereas, for the flexMIRT program, $S - X^2$ values are obtained using the alternative computational approach, and OE tables are collapsed towards the middle of scores, separately for each score category.

However, there is currently limited information available regarding the specific procedural differences. Thus, this report aims to provide a detailed description of the step-by-step processes for the various procedures. Moreover, a real-data study was conducted to demonstrate the computational processes and to illustrate the potential differences in results based on choices for the computational approach (original and alternative), procedure for collapsing OE tables (mid and even), and procedure for dealing with score categories when collapsing (concurrent and separate). For the datasets considered in this study, the results showed that the conclusions based on $S-X^2$ indeed depended on how $S - X^2$ values were computed, how OE tables were collapsed, and how score categories were handled during collapsing OE tables. Furthermore, the results also depended on the choice of MIN value used for collapsing OE tables.

It is worth noting that the mid procedure, by its definition, tends to accumulate more frequencies towards the middle of a score list. On the contrary, the even procedure tends to spread frequencies evenly throughout the score scale by collapsing cells with one of two adjacent cells with a *smaller* number of examinees. Thus, frequency distributions in collapsed OE tables are expected to be different for different collapsing procedures, and, for each score, squared differences between observed and expected proportions relative to the expected proportions (i.e., the ratio in Equation (2)) could vary depending on the collapsing procedures.

In many testing context, it is common to observe more examinees towards the middle of a score range and less examinees towards the two ends of the score range. By transferring information from the ends towards the middle, the mid procedure has a tendency to overlook potential differences between observed and expected values at the score-ends. In other words, when cells with smaller numbers remain towards the two ends, the magnitude of the ratio (i.e., squared differences between observed and expected proportions relative to expected proportions) can be large and their contributions to $S-X^2$ can also be large. However, when those cells are collapsed towards the middle of scores, their contributions to the final $S - X^2$ could become smaller. Therefore, the extent to which the results for the mid procedure differ from those for the even procedure could depend

on the shape of a score distribution.

Table 1: *Example: OE Table for Item A*

Score k	Correct		Incorrect		N_k
	O_{Ak1}	E_{Ak1}	$N_k - O_{Ak1}$	$N_k - E_{Ak1}$	
3	1	0.1275	0	0.8725	1
4	0	0.3548	2	1.6452	2
5	0	0.4625	2	1.5375	2
6	3	3.4680	9	8.5320	12
7	5	4.9042	9	9.0958	14
8	7	8.7036	14	12.2964	21
9	19	18.2533	19	19.7467	38
10	29	29.5092	25	24.4908	54
11	47	46.4357	29	29.5643	76
12	57	63.1726	37	30.8274	94
13	55	60.4181	28	22.5819	83
14	65	68.4119	23	19.5881	88
15	112	104.9454	16	23.0546	128
16	101	100.9354	17	17.0646	118
17	135	126.4712	8	16.5288	143
18	124	125.2727	14	12.7273	138
19	149	145.4393	8	11.5607	157
20	132	134.5723	11	8.4277	143
21	134	133.3714	6	6.6286	140
22	166	165.4270	6	6.5730	172
23	149	147.2876	3	4.7124	152
24	186	189.0951	8	4.9049	194
25	149	152.7675	7	3.2325	156
26	180	180.8615	4	3.1385	184
27	165	162.6749	0	2.3251	165
28	173	171.9691	1	2.0309	174
29	149	148.5474	1	1.4526	150
30	179	179.5450	2	1.4550	181
31	140	139.0662	0	0.9338	140
32	147	148.1768	2	0.8232	149
33	159	159.2699	1	0.7301	160
34	146	146.4485	1	0.5515	147
35	121	121.6260	1	0.3740	122
36	125	124.6895	0	0.3105	125
37	101	100.7991	0	0.2009	101
38	110	109.8276	0	0.1724	110
39	104	103.8744	0	0.1256	104
40	93	92.9161	0	0.0839	93
41	66	65.9575	0	0.0425	66
42	77	76.9672	0	0.0328	77
43	49	48.9879	0	0.0121	49
44	54	53.9944	0	0.0056	54

Note. O_{Ak1} and E_{Ak1} refer to the observed and expected numbers of examinees with NC score k who obtain score 1 on item A, respectively.

Table 2: *Item B: OE Table after Concurrent Collapsing (MIN value = 5)*

	Correct		Incorrect	
Score	O_{Bk1}	E_{Bk1}	O_{Bk0}	E_{Bk0}
8	8	5.381	44	46.619
9	6	5.482	32	32.518
10	7	9.036	47	44.964
11	18	14.666	58	61.334
12	19	20.819	75	73.181
13	21	21.007	62	61.993
14	27	25.330	61	62.670
15	28	41.669	100	86.331
16	50	43.164	68	74.836
17	47	58.346	96	84.654
18	58	62.300	80	75.700
19	68	77.761	89	79.239
20	77	77.039	66	65.961
21	88	81.345	52	58.655
22	113	106.906	59	65.094
23	112	100.290	40	51.710
24	129	134.925	65	59.075
25	126	113.639	30	42.361
26	143	139.596	41	44.404
27	134	129.726	31	35.274
28	132	141.153	42	32.847
29	122	125.078	28	24.922
30	156	154.627	25	26.373
31	123	122.179	17	17.821
32	128	132.504	21	16.496
33	146	144.671	14	15.329
34	135	134.886	12	12.114
35	115	113.416	7	8.584
36	119	117.559	6	7.441
37	195	201.463	16	9.537
40	435	434.200	8	8.800

Note. O_{Bk1} and E_{Bk1} refer to the observed and expected numbers of examinees with NC score k who obtain score 1 on item B, respectively. Subscripts with 0 refer to the observed/expected numbers of examinees with NC score k who obtain score 0 on item B.

Table 3: *Item B: OE Table after Separate Collapsing (MIN value = 5)*

	Correct		Incorrect	
Score	O_{Bk1}	E_{Bk1}	O_{Bk0}	E_{Bk0}
6	NA	NA	17	15.670
7	NA	NA	12	12.540
8	8	5.381	15	18.409
9	6	5.482	32	32.518
10	7	9.036	47	44.964
11	18	14.666	58	61.334
12	19	20.819	75	73.181
13	21	21.007	62	61.993
14	27	25.330	61	62.670
15	28	41.669	100	86.331
16	50	43.164	68	74.836
17	47	58.346	96	84.654
18	58	62.300	80	75.700
19	68	77.761	89	79.239
20	77	77.039	66	65.961
21	88	81.345	52	58.655
22	113	106.906	59	65.094
23	112	100.290	40	51.710
24	129	134.925	65	59.075
25	126	113.639	30	42.361
26	143	139.596	41	44.404
27	134	129.726	31	35.274
28	132	141.153	42	32.847
29	122	125.078	28	24.922
30	156	154.627	25	26.373
31	123	122.179	17	17.821
32	128	132.504	21	16.496
33	146	144.671	14	15.329
34	135	134.886	12	12.114
35	115	113.416	7	8.584
36	119	117.559	6	7.441
37	92	95.972	9	5.028
38	103	105.491	NA	NA
39	101	100.561	10	7.948
40	90	90.590	NA	NA
41	65	64.715	NA	NA
42	77	75.950	NA	NA
43	48	48.589	5	5.361
44	54	53.796	NA	NA

Note. O_{Bk1} and E_{Bk1} refer to the observed and expected numbers of examinees with NC score k who obtain score 1 on item B, respectively. Subscripts with 0 refer to the observed/expected numbers of examinees with NC score k who obtain score 0 on item B. NA refers to cells that were collapsed with one of adjacent cells.

Table 4: *Comparison in $S-X^2$ Results between Two Computational Approaches (MIN Value = 1)*

Item	Original			Alternative		
	$S - X^2$	*df*	p-val	$S - X^2$	*df*	p-val
1	32.015	39	0.778	44.568	41	0.324
2	55.713	39	0.040	62.325	39	0.010
3	57.458	37	0.017	59.560	39	0.019
4	44.673	39	0.246	53.332	40	0.077
5	49.050	39	0.130	56.956	41	0.050
6	44.584	32	0.069	49.139	33	0.035
7	48.336	35	0.066	44.318	36	0.161
8	42.160	40	0.378	37.791	41	0.614
9	33.632	39	0.713	46.507	40	0.222
10	50.983	39	0.095	49.015	40	0.155
11	31.647	36	0.676	47.246	35	0.081
12	50.194	35	0.046	50.916	36	0.051
13	35.537	39	0.629	44.682	40	0.282
14	32.731	38	0.711	46.450	40	0.224
15	37.412	40	0.587	59.358	41	0.032
16	28.311	38	0.874	35.815	39	0.616
17	34.297	38	0.641	56.083	39	0.037
18	42.240	36	0.219	45.710	38	0.182
19	45.936	39	0.207	62.667	40	0.012
20	39.759	39	0.436	46.706	40	0.216
21	40.774	39	0.392	54.662	39	0.049
22	39.601	37	0.355	41.976	38	0.303
23	52.765	35	0.027	60.186	37	0.009
24	48.010	30	0.020	62.709	31	0.001
25	52.164	28	0.004	53.387	30	0.005
26	35.552	37	0.537	59.102	38	0.016
27	24.877	34	0.873	38.784	36	0.345
28	41.341	39	0.369	40.696	40	0.440
29	45.368	38	0.192	58.764	39	0.022
30	25.793	37	0.917	40.883	39	0.388
31	43.785	39	0.276	53.325	40	0.077
32	50.447	39	0.104	49.945	40	0.135
33	54.465	39	0.051	40.965	39	0.384
34	66.219	33	0.001	83.909	35	0.000
35	41.877	40	0.389	50.719	41	0.142
36	44.041	39	0.267	54.641	41	0.075
37	36.204	37	0.506	40.541	38	0.359
38	67.240	38	0.002	58.734	40	0.028
39	36.686	35	0.391	38.009	37	0.423
40	49.589	31	0.018	67.120	33	0.000
41	40.908	36	0.264	46.700	38	0.157
42	49.186	33	0.035	43.997	33	0.096
43	43.595	37	0.211	40.831	38	0.347
44	28.296	32	0.655	30.544	33	0.590
45	39.909	38	0.385	44.101	39	0.265
46	52.373	38	0.060	53.935	39	0.056
47	51.558	37	0.056	55.445	38	0.034
48	53.624	39	0.060	66.384	40	0.005
49	31.767	39	0.788	49.424	40	0.146
50	31.324	38	0.770	48.994	39	0.131
51	42.194	37	0.256	44.871	38	0.206
52	37.244	36	0.412	38.972	37	0.381

: Refers to items flagged at the 5% significance level only.

: Refers to items flagged at both 1% and 5% significance levels.

Table 5: *Comparison in $S-X^2$ Results between Two Computational Approaches (MIN Value = 5)*

Item	Original			Alternative		
	$S - X^2$	*df*	p-val	$S - X^2$	*df*	p-val
1	31.436	37	0.727	38.705	38	0.438
2	45.611	34	0.088	52.950	35	0.026
3	45.397	31	0.046	54.454	32	0.008
4	37.208	33	0.281	44.778	34	0.102
5	29.888	35	0.713	48.950	37	0.090
6	30.443	23	0.137	30.505	23	0.135
7	38.491	28	0.089	35.627	30	0.221
8	37.639	36	0.394	37.287	36	0.410
9	31.209	35	0.652	42.929	37	0.232
10	43.472	34	0.128	45.052	36	0.143
11	26.223	29	0.614	40.246	29	0.080
12	37.439	28	0.109	43.083	30	0.058
13	33.449	33	0.445	41.939	34	0.165
14	31.613	33	0.536	42.366	34	0.154
15	36.278	36	0.456	56.001	36	0.018
16	25.730	31	0.734	33.331	31	0.354
17	31.027	32	0.516	54.615	34	0.014
18	39.374	31	0.144	44.343	31	0.057
19	42.718	34	0.145	61.400	36	0.005
20	37.075	35	0.373	44.169	36	0.165
21	38.610	34	0.269	53.492	35	0.024
22	37.518	30	0.163	39.650	32	0.166
23	45.046	30	0.038	47.978	31	0.026
24	42.312	23	0.008	49.700	24	0.002
25	32.897	21	0.047	43.208	23	0.007
26	30.137	30	0.459	52.459	32	0.013
27	18.032	29	0.944	31.803	30	0.377
28	31.810	34	0.575	37.730	35	0.346
29	41.422	32	0.123	54.170	33	0.012
30	22.295	32	0.899	35.696	33	0.343
31	39.150	33	0.213	43.047	33	0.113
32	45.323	35	0.114	39.186	35	0.288
33	49.891	33	0.030	36.269	34	0.363
34	36.263	25	0.068	52.527	27	0.002
35	37.484	35	0.356	46.427	36	0.114
36	40.410	35	0.244	45.830	35	0.104
37	26.712	32	0.731	35.059	33	0.371
38	61.837	34	0.002	56.570	34	0.009
39	23.838	28	0.690	31.755	31	0.429
40	40.089	23	0.015	59.928	26	0.000
41	40.408	29	0.077	43.920	31	0.062
42	39.058	26	0.048	33.288	27	0.188
43	36.504	29	0.159	35.911	31	0.249
44	25.526	26	0.489	27.566	28	0.488
45	38.210	34	0.284	43.134	35	0.163
46	51.795	34	0.026	53.585	36	0.030
47	41.956	32	0.112	50.709	33	0.025
48	45.227	35	0.115	56.758	36	0.015
49	30.266	35	0.696	45.471	35	0.111
50	26.497	31	0.697	44.273	33	0.091
51	36.038	30	0.207	40.842	32	0.136
52	34.504	31	0.304	38.470	32	0.200

: Refers to items flagged at the 5% significance level only.

: Refers to items flagged at both 1% and 5% significance levels.

Table 6: *Comparison in $S - X^2$ Results between Two Collapsing Procedures (MIN Value = 1)*

Item	Mid			Even		
	$S - X^2$	*df*	p-val	$S - X^2$	*df*	p-val
1	32.015	39	0.778	32.015	39	0.778
2	55.713	39	0.040	55.713	39	0.040
3	57.458	37	0.017	57.458	37	0.017
4	44.673	39	0.246	44.673	39	0.246
5	49.050	39	0.130	49.050	39	0.130
6	46.967	32	0.043	44.584	32	0.069
7	50.484	35	0.044	48.336	35	0.066
8	42.160	40	0.378	42.160	40	0.378
9	33.632	39	0.713	33.632	39	0.713
10	50.983	39	0.095	50.983	39	0.095
11	31.647	36	0.676	31.647	36	0.676
12	50.194	35	0.046	50.194	35	0.046
13	35.537	39	0.629	35.537	39	0.629
14	32.731	38	0.711	32.731	38	0.711
15	37.412	40	0.587	37.412	40	0.587
16	28.311	38	0.874	28.311	38	0.874
17	34.297	38	0.641	34.297	38	0.641
18	42.190	36	0.221	42.240	36	0.219
19	45.936	39	0.207	45.936	39	0.207
20	39.759	39	0.436	39.759	39	0.436
21	40.774	39	0.392	40.774	39	0.392
22	39.601	37	0.355	39.601	37	0.355
23	53.097	35	0.026	52.765	35	0.027
24	48.010	30	0.020	48.010	30	0.020
25	46.330	28	0.016	52.164	28	0.004
26	35.552	37	0.537	35.552	37	0.537
27	24.877	34	0.873	24.877	34	0.873
28	41.341	39	0.369	41.341	39	0.369
29	45.368	38	0.192	45.368	38	0.192
30	25.793	37	0.917	25.793	37	0.917
31	43.785	39	0.276	43.785	39	0.276
32	50.447	39	0.104	50.447	39	0.104
33	54.465	39	0.051	54.465	39	0.051
34	65.081	33	0.001	66.219	33	0.001
35	41.877	40	0.389	41.877	40	0.389
36	44.041	39	0.267	44.041	39	0.267
37	36.204	37	0.506	36.204	37	0.506
38	67.240	38	0.002	67.240	38	0.002
39	37.075	35	0.373	36.686	35	0.391
40	50.656	31	0.014	49.589	31	0.018
41	40.908	36	0.264	40.908	36	0.264
42	48.192	33	0.043	49.186	33	0.035
43	43.595	37	0.211	43.595	37	0.211
44	28.337	32	0.653	28.296	32	0.655
45	39.909	38	0.385	39.909	38	0.385
46	52.373	38	0.060	52.373	38	0.060
47	51.558	37	0.056	51.558	37	0.056
48	53.624	39	0.060	53.624	39	0.060
49	31.767	39	0.788	31.767	39	0.788
50	31.324	38	0.770	31.324	38	0.770
51	42.194	37	0.256	42.194	37	0.256
52	37.244	36	0.412	37.244	36	0.412

: Refers to items flagged at the 5% significance level only.

: Refers to items flagged at both 1% and 5% significance levels.

Table 7: *Comparison in $S - X^2$ Results between Two Collapsing Procedures (MIN Value = 5)*

Item	Mid			Even		
	$S - X^2$	*df*	p-val	$S - X^2$	*df*	p-val
1	31.436	37	0.727	31.436	37	0.727
2	45.611	34	0.088	45.611	34	0.088
3	44.091	31	0.060	45.397	31	0.046
4	37.208	33	0.281	37.208	33	0.281
5	29.888	35	0.713	29.888	35	0.713
6	30.362	23	0.139	30.443	23	0.137
7	38.491	28	0.089	38.491	28	0.089
8	37.639	36	0.394	37.639	36	0.394
9	31.209	35	0.652	31.209	35	0.652
10	43.734	34	0.122	43.472	34	0.128
11	25.537	29	0.650	26.223	29	0.614
12	38.948	28	0.082	37.439	28	0.109
13	33.449	33	0.445	33.449	33	0.445
14	31.613	33	0.536	31.613	33	0.536
15	36.278	36	0.456	36.278	36	0.456
16	24.333	31	0.797	25.730	31	0.734
17	31.226	32	0.506	31.027	32	0.516
18	37.258	31	0.203	39.374	31	0.144
19	41.362	34	0.180	42.718	34	0.145
20	37.075	35	0.373	37.075	35	0.373
21	38.610	34	0.269	38.610	34	0.269
22	37.518	30	0.163	37.518	30	0.163
23	45.046	30	0.038	45.046	30	0.038
24	42.312	23	0.008	42.312	23	0.008
25	36.329	21	0.020	32.897	21	0.047
26	30.137	30	0.459	30.137	30	0.459
27	18.303	29	0.938	18.032	29	0.944
28	32.275	34	0.552	31.810	34	0.575
29	39.977	32	0.157	41.422	32	0.123
30	21.199	32	0.927	22.295	32	0.899
31	39.150	33	0.213	39.150	33	0.213
32	45.323	35	0.114	45.323	35	0.114
33	50.120	33	0.028	49.891	33	0.030
34	33.982	25	0.108	36.263	25	0.068
35	37.484	35	0.356	37.484	35	0.356
36	41.077	35	0.222	40.410	35	0.244
37	26.712	32	0.731	26.712	32	0.731
38	62.477	34	0.002	61.837	34	0.002
39	23.847	28	0.690	23.838	28	0.690
40	39.915	23	0.016	40.089	23	0.015
41	40.038	29	0.083	40.408	29	0.077
42	39.058	26	0.048	39.058	26	0.048
43	37.462	29	0.135	36.504	29	0.159
44	25.526	26	0.489	25.526	26	0.489
45	38.210	34	0.284	38.210	34	0.284
46	51.795	34	0.026	51.795	34	0.026
47	42.552	32	0.101	41.956	32	0.112
48	45.227	35	0.115	45.227	35	0.115
49	30.266	35	0.696	30.266	35	0.696
50	27.211	31	0.662	26.497	31	0.697
51	33.559	30	0.299	36.038	30	0.207
52	34.504	31	0.304	34.504	31	0.304

: Refers to items flagged at the 5% significance level only.

: Refers to items flagged at both 1% and 5% significance levels.

Table 8: *Comparison in $S - X^2$ Results between Two Approaches to Dealing with Score Categories (MIN Value = 1)*

Item	Concurrent			Separate		
	$S - X^2$	*df*	p-val	$S - X^2$	*df*	p-val
1	32.015	39	0.778	33.360	39	0.724
2	55.713	39	0.040	55.923	39	0.039
3	57.458	37	0.017	57.614	37	0.017
4	44.673	39	0.246	45.411	39	0.222
5	49.050	39	0.130	49.597	39	0.119
6	44.584	32	0.069	44.806	32	0.066
7	48.336	35	0.066	45.785	34	0.085
8	42.160	40	0.378	42.723	40	0.355
9	33.632	39	0.713	34.750	39	0.664
10	50.983	39	0.095	51.568	39	0.086
11	31.647	36	0.676	31.803	35	0.623
12	50.194	35	0.046	51.032	35	0.039
13	35.537	39	0.629	35.770	39	0.618
14	32.731	38	0.711	32.999	38	0.700
15	37.412	40	0.587	37.477	40	0.584
16	28.311	38	0.874	28.527	37	0.840
17	34.297	38	0.641	34.345	38	0.639
18	42.240	36	0.219	44.091	36	0.167
19	45.936	39	0.207	46.064	39	0.203
20	39.759	39	0.436	40.393	39	0.409
21	40.774	39	0.392	40.950	39	0.385
22	39.601	37	0.355	39.642	36	0.311
23	52.765	35	0.027	53.264	35	0.025
24	48.010	30	0.020	48.926	29	0.012
25	52.164	28	0.004	52.719	28	0.003
26	35.552	37	0.537	35.380	36	0.498
27	24.877	34	0.873	25.480	34	0.854
28	41.341	39	0.369	41.409	39	0.366
29	45.368	38	0.192	45.656	38	0.184
30	25.793	37	0.917	25.849	37	0.916
31	43.785	39	0.276	42.947	38	0.268
32	50.447	39	0.104	50.455	39	0.103
33	54.465	39	0.051	54.589	39	0.050
34	66.219	33	0.001	65.253	33	0.001
35	41.877	40	0.389	41.934	40	0.387
36	44.041	39	0.267	44.326	39	0.257
37	36.204	37	0.506	36.698	37	0.483
38	67.240	38	0.002	67.732	38	0.002
39	36.686	35	0.391	36.747	35	0.388
40	49.589	31	0.018	50.805	31	0.014
41	40.908	36	0.264	41.009	36	0.260
42	49.186	33	0.035	47.129	32	0.041
43	43.595	37	0.211	43.688	37	0.209
44	28.296	32	0.655	28.349	32	0.652
45	39.909	38	0.385	39.995	38	0.382
46	52.373	38	0.060	52.914	38	0.055
47	51.558	37	0.056	52.111	37	0.051
48	53.624	39	0.060	53.657	39	0.059
49	31.767	39	0.788	31.896	39	0.783
50	31.324	38	0.770	31.376	38	0.768
51	42.194	37	0.256	42.468	37	0.247
52	37.244	36	0.412	37.311	36	0.409

: Refers to items flagged at the 5% significance level only.

: Refers to items flagged at both 1% and 5% significance levels.

Table 9: *Comparison in $S - X^2$ Results between Two Approaches to Dealing with Score Categories (MIN Value = 5)*

Item	Concurrent			Separate		
	$S - X^2$	*df*	p-val	$S - X^2$	*df*	p-val
1	31.436	37	0.727	32.292	37	0.689
2	45.611	34	0.088	46.590	34	0.074
3	45.397	31	0.046	45.553	31	0.044
4	37.208	33	0.281	37.334	31	0.201
5	29.888	35	0.713	33.079	35	0.561
6	30.443	23	0.137	31.462	22	0.087
7	38.491	28	0.089	40.304	28	0.062
8	37.639	36	0.394	38.613	36	0.352
9	31.209	35	0.652	30.549	34	0.638
10	43.472	34	0.128	42.879	34	0.141
11	26.223	29	0.614	27.098	29	0.566
12	37.439	28	0.109	39.162	28	0.078
13	33.449	33	0.445	33.587	33	0.439
14	31.613	33	0.536	32.033	33	0.515
15	36.278	36	0.456	36.343	35	0.406
16	25.730	31	0.734	25.769	30	0.687
17	31.027	32	0.516	31.209	31	0.456
18	39.374	31	0.144	37.687	31	0.190
19	42.718	34	0.145	43.098	34	0.136
20	37.075	35	0.373	37.832	35	0.341
21	38.610	34	0.269	38.230	34	0.283
22	37.518	30	0.163	35.854	30	0.213
23	45.046	30	0.038	45.875	30	0.032
24	42.312	23	0.008	42.811	23	0.007
25	32.897	21	0.047	37.149	21	0.016
26	30.137	30	0.459	30.497	30	0.440
27	18.032	29	0.944	18.636	29	0.930
28	31.810	34	0.575	33.274	34	0.503
29	41.422	32	0.123	41.881	32	0.113
30	22.295	32	0.899	22.612	32	0.890
31	39.150	33	0.213	39.498	32	0.170
32	45.323	35	0.114	45.741	35	0.106
33	49.891	33	0.030	50.234	32	0.021
34	36.263	25	0.068	36.629	25	0.063
35	37.484	35	0.356	38.145	34	0.286
36	40.410	35	0.244	41.836	35	0.198
37	26.712	32	0.731	27.532	32	0.692
38	61.837	34	0.002	63.182	34	0.002
39	23.838	28	0.690	25.433	28	0.604
40	40.089	23	0.015	41.765	23	0.010
41	40.408	29	0.077	40.570	29	0.075
42	39.058	26	0.048	40.085	26	0.038
43	36.504	29	0.159	36.975	29	0.147
44	25.526	26	0.489	26.972	26	0.411
45	38.210	34	0.284	38.400	34	0.277
46	51.795	34	0.026	51.682	34	0.027
47	41.956	32	0.112	43.186	32	0.090
48	45.227	35	0.115	46.273	35	0.096
49	30.266	35	0.696	30.655	35	0.678
50	26.497	31	0.697	27.292	31	0.657
51	36.038	30	0.207	34.652	30	0.256
52	34.504	31	0.304	34.839	31	0.290

: Refers to items flagged at the 5% significance level only.

: Refers to items flagged at both 1% and 5% significance levels.

Figure 2: Collapsing OE Table using the Mid Procedure (MIN Value = 1) for Item A

(a) After the First Iteration of Collapsing the First and Last Rows

	Correct		Incorrect		
Score k	O_{Ak1}	E_{Ak1}	$N_k - O_{Ak1}$	$N_k - E_{Ak1}$	N_k
4	1	0.4823	2	2.5177	3
5	0	0.4625	2	1.5375	2
6	3	3.4680	9	8.5320	12
7	5	4.9042	9	9.0958	14
:	:	:	:	:	:
41	66	65.9575	0	0.0425	66
42	77	76.9672	0	0.0328	77
43	103	102.9823	0	0.0177	103

(b) After the Second Iteration

	Correct		Incorrect		
Score k	O_{Ak1}	E_{Ak1}	$N_k - O_{Ak1}$	$N_k - E_{Ak1}$	N_k
5	1	0.9448	4	4.0552	5
6	3	3.4680	9	8.5320	12
7	5	4.9042	9	9.0958	14
:	:	:	:	:	:
41	66	65.9575	0	0.0425	66
42	180	179.9495	0	0.0505	180

(c) After the Third Iteration

	Correct		Incorrect		
Score k	O_{Ak1}	E_{Ak1}	$N_k - O_{Ak1}$	$N_k - E_{Ak1}$	N_k
6	4	4.4128	13	12.5872	17
7	5	4.9042	9	9.0958	14
:	:	:	:	:	:
41	246	245.9070	0	0.0930	246

Note. O_{Ak1} and E_{Ak1} refer to the observed and expected numbers of examinees with NC score k who obtain score 1 on item A, respectively.

Figure 3: Collapsing OE Table using the Even Procedure (MIN Value = 1) for Item A

(a) After the First Iteration of Collapsing Rows for Scores 3 and 4

	Correct		Incorrect		
Score k	O_{Ak1}	E_{Ak1}	$N_k - O_{Ak1}$	$N_k - E_{Ak1}$	N_k
4	1	0.4823	2	2.5177	3
5	0	0.4625	2	1.5375	2
6	3	3.4680	9	8.5320	12
:	:	:	:	:	:
30	179	179.5450	2	1.4550	181
31	140	139.0662	0	0.9338	140
32	147	148.1768	2	0.8232	149
33	159	159.2699	1	0.7301	160
:	:	:	:	:	:
44	54	53.9944	0	0.0056	54

(b) After the Second Iteration of Collapsing Rows for Scores 4 and 5

	Correct		Incorrect		
Score k	O_{Ak1}	E_{Ak1}	$N_k - O_{Ak1}$	$N_k - E_{Ak1}$	N_k
5	1	0.9448	4	4.0552	5
6	3	3.4680	9	8.5320	12
:	:	:	:	:	:
30	179	179.5450	2	1.4550	181
31	140	139.0662	0	0.9338	140
32	147	148.1768	2	0.8232	149
33	159	159.2699	1	0.7301	160
:	:	:	:	:	:
44	54	53.9944	0	0.0056	54

Note. O_{Ak1} and E_{Ak1} refer to the observed and expected numbers of examinees with NC score k who obtain score 1 on item A, respectively.

Figure 3: Cont'd

(c) After the Third Iteration of Collapsing Rows for Scores 5 and 6

	Correct		Incorrect		
Score k	O_{Ak1}	E_{Ak1}	$N_k - O_{Ak1}$	$N_k - E_{Ak1}$	N_k
6	4	4.4128	13	12.5872	17
:	:	:	:	:	:
30	179	179.5450	2	1.4550	181
31	140	139.0662	0	0.9338	140
32	147	148.1768	2	0.8232	149
33	159	159.2699	1	0.7301	160
:	:	:	:	:	:
44	54	53.9944	0	0.0056	54

(d) After the Fourth Iteration of Collapsing Rows for Scores 31 and 32

	Correct		Incorrect		
Score k	O_{Ak1}	E_{Ak1}	$N_k - O_{Ak1}$	$N_k - E_{Ak1}$	N_k
6	4	4.4128	13	12.5872	17
:	:	:	:	:	:
30	179	179.5450	2	1.4550	181
32	287	287.2430	2	1.7570	289
33	159	159.2699	1	0.7301	160
:	:	:	:	:	:
44	54	53.9944	0	0.0056	54

Note. O_{Ak1} and E_{Ak1} refer to the observed and expected numbers of examinees with NC score k who obtain score 1 on item A, respectively.

6 References

Bjorner, J. B., Smith, K. J., Orlando, M., Stone, C., Thissen, D., & Sun, X. (2006). *IRTFIT: A macro for item fit and local dependence tests under IRT models* [Computer Program]. Lincoln, RI: QualityMetric Incorporated.

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

Chalmers, P. (2019). mirt: Multidimensional Item Response Theory. R package version 1.30. `https://CRAN.R-project.org/package=mirt`

Cochran, W. G. (1952). The chi-squre test of goodness of fit. *Annals of Mathematical Statistics*, *23*, 315-345.

Chon, K. H., Lee, W., & Dunbar, S. B. (2010). A comparison of item fit statistics for mixed IRT models. *Journal of Educational Measurement*, *47*(3), 318-338.

Kang, T., & Chen, T. T. (2008). Performance of the generalized $S - X^2$ item fit index for polytomous IRT models. *Journal of Educational Measurement*, *45*(4), 391-406.

Kim, H. J., & Lee, W. (2022). Evaluation of factors affecting the performance of the $S - X^2$ item-fit index. *Journal of Educational Measurement*, *59*(1), 105-133.

Kolen, M. K., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer-Verlag.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement*, *8*, 452-461.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement*, *9*, 49-57.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous Item Response Theory models. *Applied Psychological Measurement*, *24*(1), 50-64.

Orlando, M., & Thissen, D. (2003). Further investigation of the performance of $S - X^2$: An item fit index for use with dichotomous item response theory models. *Applied Psychological Measurement*, *27*(4), 289-298.

Yen, W. (1981). Using simulation results to choose a latent trait model. *Applied Psychological Measurement*, *5*, 245-262.

Appendix A Computation of $S - X^2$ Item-Fit Index

According to Orlando and Thissen (2000), the expected number of examinees with number-correct (NC) score k who respond correctly to item i can be approximated using equally-spaced quadrature θ-points from -4.5 and 4.5 as follows:

$$E_{ik1} = N_k \frac{\sum_{q=1}^{q=Q} T_i(\theta_q) S_{k-1}^{*i}(\theta_q) w(\theta_q)}{\sum_{q=1}^{q=Q} S_k(\theta_q) w(\theta_q)}, \tag{A1}$$

where k is the NC score including item i; N_k is the number of examinees with NC score k; θ_q is the q^{th} quadrature point such that $\theta_1 = -4.5$ and $\theta_Q = 4.5$; $w(\theta_q)$ is the probability density for θ_q such that $\sum_{q=1}^{Q} w(\theta_q) = 1$; $T_i(\theta_q)$ is the probability that an examinee with ability θ_q responds correctly to item i; $S_{k-1}^{*i}(\theta_q)$ is the probability that an examinee with ability θ_q obtains NC score $k-1$ without item i; and $S_k(\theta_q)$ is the probability that an examinee with ability θ_q obtains NC score k including item i. This Appendix presents how $S_k(\theta_q)$ and $S_{k-1}^{*i}(\theta_q)$ can be computed and uses an example to demonstrate the procedure.

Suppose that $S_k(\theta_q)$ in the denominator of Equation (A1) is computed first. Since there can be $(n+1)$ possible NC scores for n items and Q possible θ points, the matrix $\mathbf{S}$ can be denoted

$$\underset{((n+1)\times Q)}{\mathbf{S}} = \begin{bmatrix} S_0(\theta_1) & S_0(\theta_2) & \cdots & S_0(\theta_Q) \\ S_1(\theta_1) & S_1(\theta_2) & \cdots & S_1(\theta_Q) \\ & & \ddots & \\ \vdots & \vdots & S_k(\theta_q) & \vdots \\ & & & \ddots \\ S_n(\theta_1) & S_n(\theta_2) & \cdots & S_n(\theta_Q) \end{bmatrix},$$

where the rows represent $(n+1)$ possible NC scores from 0 to n and the columns represent Q possible θ points. Each cell, $S_k(\theta_q)$, represents the probability that examinees with ability θ_q obtain NC score k. In $\mathbf{S}$, each column represents probabilities of obtaining NC scores 0 through n (i.e., score distribution) for given θ_q. The score distribution for given θ_q can be computed using a recursive algorithm suggested by Lord and Wingersky (1984) (i.e., Lord-Wingersky (LW) algorithm).

According to Kolen and Brennan (2014), $f_r(k|\theta_q)$ is defined as the distribution of NC scores over the first r items for examinees of ability θ_q such that $r = 1, ..., n$ and $k = 0, ..., r$. Let $f_1(k = 0|\theta_q)$ and $f_1(k = 1|\theta_q)$ be defined as follows:

$$f_1(k = 0|\theta_q) = 1 - P_1(\theta_q) = Q_1(\theta_q)$$
$$f_1(k = 1|\theta_q) = P_1(\theta_q)$$

where $Q_1(\theta_q)$ and $P_1(\theta_q)$ are the probabilities that examinees with ability θ_q obtain score 0 and 1 on the first item, respectively. Then, for $r \geq 2$, the LW algorithm can be generalized using the recursion formula as follows:

$$\begin{aligned} f_r(k|\theta_q) &= f_{r-1}(k|\theta_q)Q_r(\theta_q), && k = 0 \\ &= f_{r-1}(k|\theta_q)Q_r(\theta_q) + f_{r-1}(k-1|\theta_q)P_r(\theta_q), && 0 < k < r, \\ &= f_{r-1}(k-1|\theta_q)P_r(\theta_q), && k = r \end{aligned} \tag{A2}$$

where $P_r(\theta_q)$ and $Q_r(\theta_q)$ are the probabilities that examinees with ability θ_q respond to the r^{th} item correctly and incorrectly, respectively. In the recursion formula (A2), the maximum possible value for r is n because there are n items. After the LW algorithm is completed for all n items (i.e., $r = n$), $f_n(k = 0|\theta_q)$, $f_n(k = 1|\theta_q)$, ..., $f_n(k = n|\theta_q)$ are the score distribution given θ_q, which are equivalent to $S_0(\theta_q)$, ..., $S_n(\theta_q)$ in the q^{th} column of $\mathbf{S}$.

The recursion formula can be easily expressed in the matrix format as follows:

$$\underset{((n+1)\times n)}{\mathbf{f}}(\theta_q) = \begin{bmatrix} f_1(k=0) & f_2(k=0) & \cdots & f_{n-1}(k=0) & f_n(k=0) \\ f_1(k=1) & f_2(k=1) & \cdots & f_{n-1}(k=1) & f_n(k=1) \\ & f_2(k=2) & \cdots & f_{n-1}(k=2) & f_n(k=2) \\ & & \ddots & \vdots & \vdots \\ & & & f_{n-1}(k=n-1) & f_n(k=n-1) \\ & & & & f_n(k=n) \end{bmatrix},$$

where the columns represent the r^{th} recursion for the first r items for $r = 1, ..., n$, and the rows represent possible NC scores k using the first r items such that $k = 0, ..., r$. The last column (i.e., $r = n$) represents the probabilities of obtaining scores 0 through n considering all n items, $S_0(\theta_q)$, $S_1(\theta_q)$, ..., and $S_n(\theta_q)$. In other words, $S_k(\theta_q)$ in the denominator of Equation (A1) is equivalent to $f_n(k)$ from the n^{th} recursion formula in the last column of $\mathbf{f}$. Note that, for the sake of notational simplicity, the conditional variable θ_q was dropped from all f terms in $\mathbf{f}$. The same procedure can be repeated for all θ quadrature points, and at the end, there are Q score distributions, one for each θ. The final Q score distributions then replace the Q columns of $\mathbf{S}$.

Values for $S^{*i}_{k-1}(\theta)$ in the numerator of Equation (A1) can be obtained in the same manner, but using the rest $(n-1)$ items without item i. Since the possible NC scores for a set of $(n-1)$ items range from 0 to $n-1$, the matrix $\mathbf{S}^{*i}$ can be denoted

$$\underset{(n\times Q)}{\mathbf{S}^{*i}} = \begin{bmatrix} S_0^{*i}(\theta_1) & S_0^{*i}(\theta_2) & \cdots & S_0^{*i}(\theta_Q) \\ S_1^{*i}(\theta_1) & S_1^{*i}(\theta_2) & \cdots & S_1^{*i}(\theta_Q) \\ \vdots & \vdots & \ddots & \vdots \\ S_{n-1}^{*i}(\theta_1) & S_{n-1}^{*i}(\theta_2) & \cdots & S_{n-1}^{*i}(\theta_Q) \end{bmatrix}.$$

In $\mathbf{S}^{*i}$, each column at given θ_q can be obtained using the LW recursive algorithm expressed in the matrix $\mathbf{f}^{*i}(\theta_q)$ which can be denoted

$$\underset{(n\times(n-1))}{\mathbf{f}^{*i}}(\theta_q) = \begin{bmatrix} f_1^{*i}(k=0) & f_2^{*i}(k=0) & \cdots & f_{n-2}^{*i}(k=0) & f_{n-1}^{*i}(k=0) \\ f_1^{*i}(k=1) & f_2^{*i}(k=1) & \cdots & f_{n-2}^{*i}(k=1) & f_{n-1}^{*i}(k=1) \\ & f_2^{*i}(k=2) & \cdots & f_{n-2}^{*i}(k=2) & f_{n-1}^{*i}(k=2) \\ & & \ddots & \vdots & \vdots \\ & & & f_{n-2}^{*i}(k=n-2) & f_{n-1}^{*i}(k=n-2) \\ & & & & f_{n-1}^{*i}(k=n-1) \end{bmatrix}.$$

Note that the conditional notations θ_q are dropped from f^{*i}s in $\mathbf{f}^{*i}$. The last column of the matrix $\mathbf{f}^{*i}(\theta_q)$ becomes the score distribution and corresponds to the q^{th} column of the matrix $\mathbf{S}^{*i}$.

After $\mathbf{S}$ and $\mathbf{S}^{*i}$ for item i are computed, $S_k(\theta_q)$ in the denominator of Equation (A1) for $q = 1, ..., Q$ is substituted by the values in the $(k+1)^{th}$ row of $\mathbf{S}$, and $S_{k-1}^{*i}(\theta_q)$ in the numerator of Equation (A1) for $q = 1, ..., Q$ is substituted by the values in the k^{th} row of $\mathbf{S}^{*i}$.

Example

As an example, consider a subset of three items from the example presented in Kolen and Brennan (2014, p.195). For the three items, Table A1 presents the item discrimination (a), difficulty (b), and pseudo-guessing (c) parameters as well as the probabilities of responding correctly and incorrectly to each item when $\theta = 0$. Suppose that the $S-X^2$ item-fit index is computed for Item 3, and the expected number of examinees is estimated for score category 3 (i.e., E_{331}). According to Equation (A1), E_{331} can be obtained as follows:

$$E_{331} = N_3 \frac{\sum_{q=1}^{q=Q} T_3(\theta_q) S_2^{*3}(\theta_q) w(\theta_q)}{\sum_{q=1}^{q=Q} S_3(\theta_q) w(\theta_q)}. \tag{A3}$$

Since there are only four possible NC scores (i.e., 0, 1, 2, and 3) with three items, $\mathbf{S}$ becomes a $4 \times Q$ matrix as below

Table A1: *Item Parameters and Probabilities of Responding Correctly and Incorrectly to Each Item at* $\theta = 0$

Item Parameter	a_i	b_i	c_i	P_i	Q_i
Item 1	.60	-1.70	.20	.79	.21
Item 2	1.00	.80	.25	.48	.52
Item 3	1.40	1.30	.25	.35	.65

$$\underset{(4\times Q)}{\mathbf{S}} = \begin{bmatrix} S_0(\theta_1) & S_0(\theta_2) & \cdots & S_0(\theta_Q) \\ S_1(\theta_1) & S_1(\theta_2) & \cdots & S_1(\theta_Q) \\ S_2(\theta_1) & S_2(\theta_2) & \cdots & S_2(\theta_Q) \\ S_3(\theta_1) & S_3(\theta_2) & \cdots & S_3(\theta_Q) \end{bmatrix}. \tag{A4}$$

In order to compute $S_k(\theta_q)$ for $k = 0, ..., 3$, $\mathbf{f}(\theta_q)$ is denoted

$$\underset{(4\times 3)}{\mathbf{f}}(\theta_q) = \begin{bmatrix} Q_1 & Q_1Q_2 & Q_1Q_2Q_3 \\ P_1 & Q_1P_2 + P_1Q_2 & Q_1Q_2P_3 + (Q_1P_2 + P_1Q_2)Q_3 \\ & P_1P_2 & (Q_1P_2 + P_1Q_2)P_3 + P_1P_2Q_3 \\ & & P_1P_2P_3 \end{bmatrix},$$

where the columns represent the r^{th} recursion formulas for $r = 1$, 2, and 3, and the rows represent scores k using the first r items. The first column represents the probabilities of obtaining scores 0 or 1 considering only the first item (Item 1) (i.e., $r = 1$); the second column represents the probabilities of obtaining scores 0, 1, or 2 considering the first two items (Items 1 and 2) (i.e., $r = 2$); and the third column represents the probabilities of obtaining scores 0 through 3 considering all three items (i.e., $r = 3$). The column values from top to bottom are equivalent to $S_0(\theta_q)$, $S_1(\theta_q)$, $S_2(\theta_q)$, and $S_3(\theta_q)$, respectively. Note that the Ps and Qs in $\mathbf{f}(\theta_q)$ are all conditional on θ_q.

Let us consider calculating a score distribution for $\theta = 0$ (i.e., $S_0(0)$, $S_1(0)$, $S_2(0)$, and $S_3(0)$). In order to do so, $\mathbf{f}$ should be obtained at $\theta_q = 0$ using the probabilities given in Table A1. At $\theta_q = 0$, $\mathbf{f}$ can be represented as

$$\underset{(4\times 3)}{\mathbf{f}}(\theta_q = 0)$$

$$= \begin{bmatrix} .21 & .21 \times .52 & .21 \times .52 \times .65 \\ .79 & .21 \times .48 + .79 \times .52 & .21 \times .52 \times .35 + (.21 \times .48 + .79 \times .52).65 \\ & .79 \times .48 & (.21 \times .48 + .79 \times .52).35 + .79 \times .48 \times .65 \\ & & .79 \times .48 \times .35 \end{bmatrix}$$

$$= \begin{bmatrix} .21 & .1092 & .07098 \\ .79 & .5116 & .37076 \\ & .3792 & .42554 \\ & & .13272 \end{bmatrix}.$$

Therefore, the probabilities of obtaining scores 0, 1, 2, and 3 at $\theta_q = 0$ are .07098, .37076, .42554, and .13272, respectively. That is $S_0(0) = .07098$, $S_1(0) = .37076$, $S_2(0) = .42554$, and $S_3(0) = .13272$. The score distributions for other θs can be obtained using the same procedure.

Since the $S - X^2$ item-fit index is computed for Item 3, $\mathbf{S}^{*3}$ and $\mathbf{f}^{*3}(\theta_q)$ should be computed using the other two items (Items 1 and 2), and can be denoted

$$\underset{(3\times Q)}{\mathbf{S}^{*3}} = \begin{bmatrix} S_0^{*3}(\theta_1) & S_0^{*3}(\theta_2) & \cdots & S_0^{*3}(\theta_Q) \\ S_1^{*3}(\theta_1) & S_1^{*3}(\theta_2) & \cdots & S_1^{*3}(\theta_Q) \\ S_2^{*3}(\theta_1) & S_2^{*3}(\theta_2) & \cdots & S_2^{*3}(\theta_Q) \end{bmatrix}, \tag{A5}$$

and

$$\underset{(3\times 2)}{\mathbf{f}^{*3}}(\theta_q) = \begin{bmatrix} Q_1(\theta_q) & Q_1(\theta_q)Q_2(\theta_q) \\ P_1(\theta_q) & Q_1(\theta_q)P_2(\theta_q) + P_1(\theta_q)Q_2(\theta_q) \\ & P_1(\theta_q)P_2(\theta_q) \end{bmatrix}.$$

Then, for given θ_q, the second column gives the probabilities of obtaining scores 0, 1, and 2 based on Items 1 and 2, which are equivalent to $S_0^{*3}(\theta_q)$, $S_1^{*3}(\theta_q)$, and $S_3^{*3}(\theta_q)$. For example, when $\theta_q = 0$, $\mathbf{f}^{*3}(\theta_q)$ becomes

$$\underset{(3\times 2)}{\mathbf{f}}(\theta_q = 0) = \begin{bmatrix} .21 & .1092 \\ .79 & .5116 \\ & .3792 \end{bmatrix}.$$

Therefore, based on Item 1 and Item 2 (without Item 3), the probabilities of obtaining scores 0, 1, and 2 at $\theta_q = 0$ are .1092, .5116, and .3792, respectively–i.e., $S_0^{*3}(0) = .1092$, $S_1^{*3}(0) = .5116$, and $S_2^{*3}(0) = .3792$.

After $S_0(\theta_q)$, $S_1(\theta_q)$, $S_2(\theta_q)$, and $S_3(\theta_q)$ are computed for all possible θ_q for $q = 1$, ..., Q, values in the fourth row of $\mathbf{S}$ (i.e., $S_3(\theta_1)$, ..., $S_3(\theta_Q)$) should be used for $S_3(\theta_q)$

in the denominator of Equation (A3). Similarly, after $S_0^{*3}(\theta_q)$, $S_1^{*3}(\theta_q)$, and $S_2^{*3}(\theta_q)$ are computed for all possible θ_q for $q = 1, ..., Q$, values in the third row of $\mathbf{S}^{*3}$ (i.e., $S_2^{*3}(\theta_1)$, ..., $S_2^{*3}(\theta_Q)$) should be used for $S_2^{*3}(\theta_q)$ in the numerator of Equation (A3). The term $T_3(\theta_q)$ in the numerator simply is $P_3(\theta_q)$ for Item 3. Then, E_{331} can be obtained once the densities $w(\theta_q)$ are determined for $q = 1, ..., Q$.