METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Collapsing or Not? A Practical Guide to Handling Sparse Responses for Polytomous Items

Yale Quan [1], Chun Wang [1]

[1] *College of Education, University of Washington, Seattle, WA, USA.*

Methodology, 2025, Vol. 21(1), 46–73, https://doi.org/10.5964/meth.14303

Received: 2024-04-02 • **Accepted:** 2024-12-12 • **Published (VoR):** 2025-03-31

Handling Editor: Belén Fernández-Castilla, Universidad Nacional de Educación a Distancia, Madrid, Spain

Corresponding Author: Yale Quan, 312E Miller Hall, 2012 Skagit Ln., College of Education, University of Washington, Seattle, WA 98105, USA. E-mail: yalequan@uw.edu

Abstract

In ordinal data analysis, category collapse is the process of combining adjacent response options to create fewer response categories than were originally measured. When collapsing response categories, researchers need to be aware of inducing data-model misfit and of obtaining biased parameter estimates. Through mathematical derivation we show that category collapse induces data-model misfit when using Generalized Partial Credit IRT model (GPCM) generated data. This data-model misfit is not present when using Graded Response IRT model (GRM) generated data. Using simulation studies, we found that category collapse can indicate better data-model fit in GRM- and GPCM-generated data. In the case of GPCM data, this result is spurious and can lead practitioners to draw conclusions from models that do not fit the data well. Recovered GPCM IRT item parameters were also significantly biased. Recommendations for practitioners who wish to collapse categories are provided.

Keywords

category collapse, item response theory, generalized partial credit model, parameter recovery, data-model fit

Instruments that use ordinal response data (e.g., Likert scale data, or partial credit scoring) are widely used in educational research studies. When using ordinal response data, research practitioners often need to handle potential response sparseness. Response sparseness occurs when a response option (or score) is rarely, if ever, endorsed (or earned). The common practice for addressing response sparseness is to combine adjacent response categories to either reduce the number of response options or dichotomize a polytomous response item to obtain a significant number of responses per category

(e.g., Outpatient and Ambulatory Surgery CAHPS, 2017), to improve data model fit (e.g., National Center for Education Statistics, 2008), or to reshape the data to meet specific modeling needs (e.g., Harpe, 2015). Collapsing response categories to create a dichotomous response item from what was originally a polytomous response item can introduce significant data-model misfit (Jansen & Roskam, 1986) along with a loss of power, reduced scale reliability, and spurious statistical significance when determining measurement invariance (Altman & Royston, 2006; MacCallum et al., 2002; Rutkowski et al., 2019). When collapsing categories to improve polytomous data-model fit, observed improved model fit may be spurious and only applicable to the collected sample (Rutkowski et al., 2019).

Utilizing a category collapse approach is also commonly observed in item response theory (IRT) analyses of ordinal response data to address disordered thresholds (e.g., Kim et al., 2010; Linden et al., 2020; Matovu, 2019; Smith et al., 2003). When modeling ordinal response data with an adjacent-categories IRT model, threshold order indicates how well an instrument functions. Disordered thresholds may indicate sparse response data in one (or more) of the response categories (Wind, 2023). Threshold order depends on the number of participants responding to an item, and collapsing the central response option to address threshold ordering (e.g., Rost & Von Davier, 1995) can complicate the understanding of the original scale by mixing respondents whose central response reflects their latent trait with other respondents (Wetzel & Carstensen, 2014). Recent research suggests that when using a Partial Credit IRT model (PCM), disordered thresholds should be addressed by collapsing response options closest to the disordered threshold category. When the central response option was affected, symmetrically collapsing response categories, collapsing from both ends of the response continuum simultaneously, best addressed the disordered thresholds. (Tsai et al., 2024).

The action of collapsing categories changes the observed item response matrix, and in-turn manipulates the conditional probabilities of a respondent selecting a response option or earning a particular score (Tsai et al., 2024). This action violates the joining assumption (Jansen & Roskam, 1986) needed to fit logistic-adjacent models, like the Partial Credit Model (PCM). Previously, Harel (2014) and subsequently Harel and Steele (2018) provided an in-depth discussion on this topic when fitting a PCM.

In their 2014 paper, Harel demonstrated through mathematical derivation, that fitting a PCM to data with collapsed categories would result in deliberate model misspecification. Harel proposed three category collapse rules based on: (1) item response function, (2) maximum information, and (3) integrated information. Simulation studies were performed to test the proposed rules and the effect of category collapse on person parameter recovery, and study how category collapse influenced threshold parameter recovery. However, none of the proposed rules could be utilized to make a consistent decision on when category collapse is appropriate. Harel and Steele (2018) extended this area of research by proposing an information matrix test (IMT) to assess the degree

of data-model misspecification introduced by collapsing the central response category. They concluded that the direction the central category was collapsed influenced PCM data-model fit significantly, and this misspecification was detectable using the proposed Information Matrix test.

Although Harel and many other researchers have concluded that utilizing collapsed categories negatively impacts statistical analyses, there exists a significant amount of research utilizing data with collapsed categories. MacCallum et al. (2002) found that 11.50% of articles published between 1998 and 2000 in the *Journal of Personality and Social Psychology*, *Journal of Consulting and Clinical Psychology*, and *Journal of Counseling Psychology* contained at least one instance of dichotomization through category collapse. Additionally, only 20% of these studies provided justification for dichotomizing responses. This finding illustrates the need for a comprehensive guide on when category collapse is justified and proper guidelines for implementing category collapse.

There are two main open questions concerning category collapse that this paper aims to address. Firstly, despite the significant research performed on collapsing response categories to address threshold order and data-model fit, there is no clear conclusion on whether category collapse introduces bias into polytomous IRT model item parameter estimates. If biased item parameter estimates are obtained, it is also unclear to what extent the ability estimates may be impacted. Therefore, the question remains: by collapsing response categories to address threshold order and/or data model fit, are we inadvertently biasing item and person parameter estimates?

Secondly, the majority of previous research concerning category collapse focused on the Partial Credit IRT model. In practice, the primary two polytomous IRT models used to model education data are the Generalized Partial Credit Model (GPCM; Muraki, 1992) and the Graded Response Model (GRM; Samejima, 1968). However, to date, there has not been research focused on examining and comparing what effect category collapse has on these data generating models. Therefore, the question remains about if these results also hold for GRM and GPCM data.

Thus, the purpose of this study is to extend the literature on category collapse by investigating, through rigorous Monte Carlo simulations, if category collapse can be justified when utilizing the Graded Response and Generalized Partial Credit Item Response Theory models. We extend the literature by examining if sample size, the direction of collapse, and the number of responses within a collapsed category have an influence on Graded Response Model and Generalized Partial Credit model parameter estimation and data-model fit. By performing this in-depth investigation, we can develop recommendations for educational research practitioners who want to collapse adjacent response options while introducing the least amount of bias in their results.

We begin by presenting relevant IRT methodology along with discussing statistical methodology used for item parameter estimation. The second section presents our simulation study and results. Lastly, we present an empirical example of category collapse and

discuss how our simulation study results are supported by this study. We conclude with recommendations for research practitioners based on our study.

Methodology

Item Response Theory Methodology

Two of the most popular unidimensional polytomous Item Response Theory models used to analyze ordinal response data are the Graded Response and Generalized Partial Credit models. While these two models can be used to analyze ordinal response data, education research practitioners should be aware of the fundamental differences between these two models.

The Graded Response Model

The Graded Response model is used to model the probability of a respondent endorsing a particular response. For example, consider an instrument composed of items that have Likert scale responses coded as $k = 0, \quad ..., \quad K$ where zero is a low endorsement of the latent trait and K is high endorsement. Then the probability of an examinee with ability level θ endorsing Category k on Item j is defined as:

$$P_{jk}(x = k|\theta) = P^*_{jk}(\theta) - P^*_{jk+1}(\theta) = P(x \geq k) - P(x \geq k+1) \tag{1}$$

where

$$P^*_{jk} = \frac{1}{1 + e^{-(d_{jk} + a_j\theta)}} \quad (k = 1, \quad ..., \quad K)$$

with item-specific discrimination Parameter a_j, and K many threshold Parameters $d_1, d_2, ..., d_K$. Furthermore, we define $P^*_{j0} \equiv 1$ and $P^*_{jK} \equiv 0$.

The Generalized Partial Credit Model

In contrast, the Generalized Partial Credit model is used to model the probability of an examinee earning the k^{th} score out of a possible score of K. Respondents who receive maximum credit on an item earn a score of K, and this score decreases to 0 as respondents make mistakes. A fundamental assumption of the Generalized Partial Credit model is that potential item scores are ordered such that a respondent who earns a score of k has correctly satisfied the requirements to earn a score of $k - 1$. But the level of difficulty moving from score of $k - 1$ to k may not necessarily be higher than the level of difficulty stepping up from $k - 2$ to $k - 1$, hence the threshold parameters need not to be strictly ordered. In contrast, in GRM, since it is a "difference model," all threshold parameters

need to be strictly ordered. Then, the probability of an examinee with ability level θ earning the k^{th} score on Item j is:

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=0}^{k} - d_{jv} + a_j\theta\right]}{\sum_{c=0}^{K}\exp\left[\sum_{v=0}^{c} - d_{jv} + a_j\theta\right]} \tag{2}$$

with item specific discrimination parameter a_j, and $K_j - 1$ threshold parameters $d_1, d_2, \ldots, d_{K_j-1}$. Note that the Index c is not associated with IRT parameters and is only used to index the summations. We also define $d_{j0} \equiv 0$ and $\sum_{k=0}^{K} P_{jk}(\theta) \equiv 1$.

The Influence of Category Collapse on GRM and GPCM Parameter Estimation

Comparing Equations 1 and 2 we can see that, while both the GRM and GPCM are used to model polytomous item response data, they have significantly different functional forms. The GRM belongs to the "difference" model family while the GPCM belongs to the "divide-by-total" family of IRT models (García-Pérez, 2017; Thissen & Steinberg, 1986). This suggests that category collapse may influence each model differently. In this section, we highlight how threshold parameter estimation might be influenced by category collapse.

In Appendix A we derive the complete-data log-likelihood function and partial derivatives for GRM and GPCM respectively. From this derivation we can conclude that if a GRM is fit to the observed item response, the direction of collapse does not influence threshold parameter estimation. Collapsing the highest two parameters involves removing the threshold parameter corresponding to the highest response option. Similarly, when collapsing into a lower response category, the threshold parameter for the higher response option is removed. After removing the collapsed threshold, a GRM can still be used to estimate threshold parameters. The derivation is based on theory, and it alone will not provide a practical guide as to when collapse is needed. Instead, simulation studies must be used to determine thresholds for sparseness based on empirical evidence.

In contrast to GRM, using collapsed categories with GPCM changes the underlying threshold parameter estimation equation. Threshold parameters in GRM models the probability of responding to Category k or above. The link between non-adjacent categories allows for removal of a threshold parameter without changing the underlying IRT model. In contrast, threshold parameters in GPCM relate to two adjacent response categories with no connection to other response categories. Therefore, collapsing adjacent response categories by removing a threshold parameter influences the estimation of all other response categories. This collapse process changes the underlying IRT model. Prior research has found similar conclusions (e.g., Harel, 2014; Harel & Steele, 2018). However, these prior studies did not examine if the data-model misfit significantly biases

parameter estimation. The degree to which parameter estimates are influenced, can only be studied through simulation.

Testing Data-Model Fit

Here, we are interested in using the Generalized $S-X^2$ test for polytomous item response data (Kang & Chen, 2008) to validate the conclusions presented in prior category collapse research. That: (1) analyzing collapsed GPCM data by fitting a GPCM results in significant data-model misfit percentage, and (2) analyzing collapsed GRM data by fitting a GRM results in a nominal data-model misfit percentage. If Conclusion (1) holds, we would observe a large percentage of GPCM simulations using collapsed data to be flagged for data-model misfit as compared to baseline data-model fit percentages. If Conclusion (2) holds, we would observe a similar percentage of data-model misfit for all GRM simulations. A brief outline of the $S-X^2$ test is provided below.

The generalized $S-X^2$ test is used to determine how well a proposed IRT model matches the observed item responses. The $S-X^2$ test statistic follows an asymptotic Chi-Square distribution, and its corresponding p-value can be interpreted similar to traditional Chi-Square goodness of fit tests. The $S-X^2$ test statistic is calculated as:

$$S-X^2 = \sum_{m=K_j}^{F-K_j} N_m \sum_{k=0}^{K_j} \frac{\left(O_{ikm}-E_{ikm}\right)^2}{E_{ikm}} \tag{3}$$

where k is the item response category/score, Kj is the highest possible score/response category of Item j, F is a perfect test score, and N_m is the number of examines in Group m. Note that the outer summation starts at $m = Kj$ because groups with extreme test scores (e.g., all correct or all incorrect) will have an expected proportion of zero. The conditional expected and observed category proportions, E_{ikm} and O_{ikm} respectively, along with its degrees of freedom are calculated using methodology outlined in Kang and Chen (2008).

Simulation Study Methodology

For the simulation study we began by generating two datasets, one using a GRM data generating model and the other using a GPCM data generating model. Each having 12 items with 5 response Categories (0, 1, 2, 3, 4). We induced sparseness in response Category 3 by manipulating the data generating thresholds for the d_3 parameter until we achieved an endorsement rate of approximately 5% for Items 7–9, and approximately 2.5% for Item 10–12. No other response categories, aside from Category 3, contained sparse endorsement rates. Data were generated using the MIRT package (Chalmers, 2012) within R Statistical Software v4.3.2 (R Core Team, 2023). Data generating parameters are provided in Appendix B.

Two endorsement thresholds were used to determine when category collapse is appropriate: (1) When a response category is endorsed by 5% or less of respondents, and (2) When a response category is endorsed by 2.5% or less of the respondents. Using the 2.5% endorsement threshold Items 10–12 would have Category 3 collapsed into an adjacent category, and using the 5% endorsement threshold would result in Items 7–12 having Category 3 collapsed into an adjacent category.

This collapse process resulted in 6 unique datasets for each data-generating process is outlined in Table 1 below. The *Baseline* dataset which is the original data that does not contain any collapsed categories, (2a & 2b) The *Collapsed-Up* dataset where responses for Category 3 were recoded into responses for Category 4, and (3a & 3b) The *Collapsed-Down* dataset where responses for Category 3 were recoded into responses for Category 2. Datasets 2a and 3a result from using a 5% endorsement threshold, and Datasets 2b and 3b are from a 2.5% endorsement threshold. Each of these five datasets were generated using one of six sample sizes: 150, 250, 500, 1000, 1500, and 2000.

After generating the analysis datasets, a Generalized Partial Credit or Graded Response IRT model was fitted to the data depending on which data generating model was used. Parameter estimation was obtained using standard EM algorithm with fixed quadrature within the MIRT package. To determine how closely recovered item parameters from collapsed datasets match parameters recovered from the baseline dataset we calculated the relative bias (Hoogland & Boomsma, 1998) of estimated item parameters. Using the discrimination parameter as an example we denote a_j as the simulated true discrimination parameter and $\overline{\widehat{a}_j}$ as the mean of the recovered discrimination parameters for the j^{th} item over all simulations. Then the relative bias of the discrimination parameter for the j^{th} item is calculated as:

$$B\left(\overline{\widehat{a}_j}\right) = \frac{\overline{\widehat{a}_j} - a_j}{a_j} \tag{4}$$

Relative bias values more extreme than $|0.05|$ indicate that the item parameter of interest was not well recovered.

Table 1

Data Generation Process

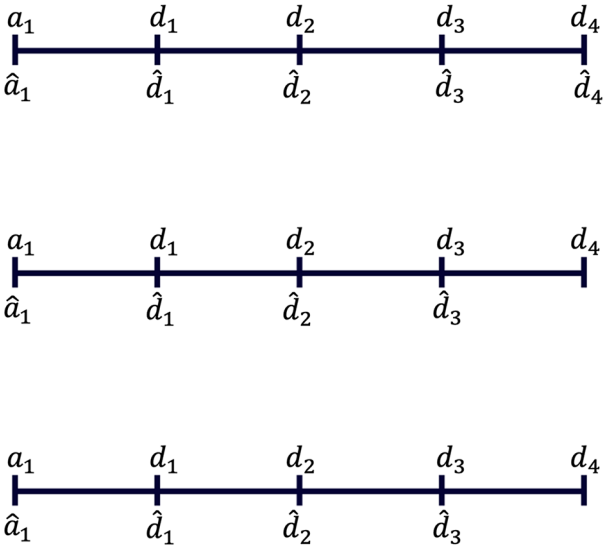
	Sparseness Condition					
Simulation Condition	**2.5%**			**5%**		
Collapse Direction	Baseline	Collapsed Up	Collapsed Down	Baseline	Collapsed Up	Collapsed Down
Sample Size	150, 250, 500, 1000, 1500, 2000	150, 250, 500, 1000, 1500, 2000	150, 250, 500, 1000, 1500, 2000	150, 250, 500, 1000, 1500, 2000	150, 250, 500, 1000, 1500, 2000	150, 250, 500, 1000, 1500, 2000
Dataset Labeling	1a	2a	3a	1b	2b	3b

Methodology
2025, Vol. 21(1), 46–73
https://doi.org/10.5964/meth.14303

When collapsing categories, the number of threshold parameters changes. Figure 1 displays how threshold parameters change when collapsing categories. The parameters on the top of the number line denote the simulated true data generating parameters and the parameters on the bottom of the number line denote the recovered item parameters. When the data does not contain any collapsed categories (Figure 1a) we can directly compare recovered parameters which are denoted as $\widehat{a}$, $\widehat{d}_1$, $\widehat{d}_2$, $\widehat{d}_3$ and $\widehat{d}_4$. When collapsing response Category 3 up into Category 4 (Figure 1b) we are essentially removing the d_4 parameter and directly comparing the remaining item parameters. When collapsing response Category 3 down into Category 2 (Figure 1c) we are removing the d_3 parameter.

Figure 1

Parameter Comparisons



Note. Figure 1a (top) represents the baseline condition. Figure 1b (middle) represents collapsing Category 3 up. Figure 1c (bottom) represents collapsing Category 3 down.

To determine how closely the recovered person parameters matched the simulated true data generating person parameters, we calculated the simulation average bias and Root Mean Squared Error (RMSE) of recovered person parameters. RMSE is calculated using Equation 5 where $\widehat{\theta}_i$ is the recovered person parameter of Person i and θ_i is the simulated true data generating person parameter for Person i.

$$RMSE(\widehat{\theta}) = \sqrt{\frac{\sum_{i=1}^{N}(\widehat{\theta}_i - \theta_i)^2}{N}} \quad (5)$$

To understand if category collapse results in detectable model misfit we calculated the percentage of items that were flagged for model misspecification across all simulations. In summary, our simulation study is a $6 \times 2 \times 3 \times 2$ completely crossed design with (a) 6 sample size levels (150, 250, 500, 1000, 1500, 2000), (b) 2 levels of response sparseness for Category 3 (5%, 2.5%), (c) 3 collapse directions (baseline, collapsed-up, collapsed-down), and (d) 2 IRT models (GRM and GPCM).

Simulation Study Results

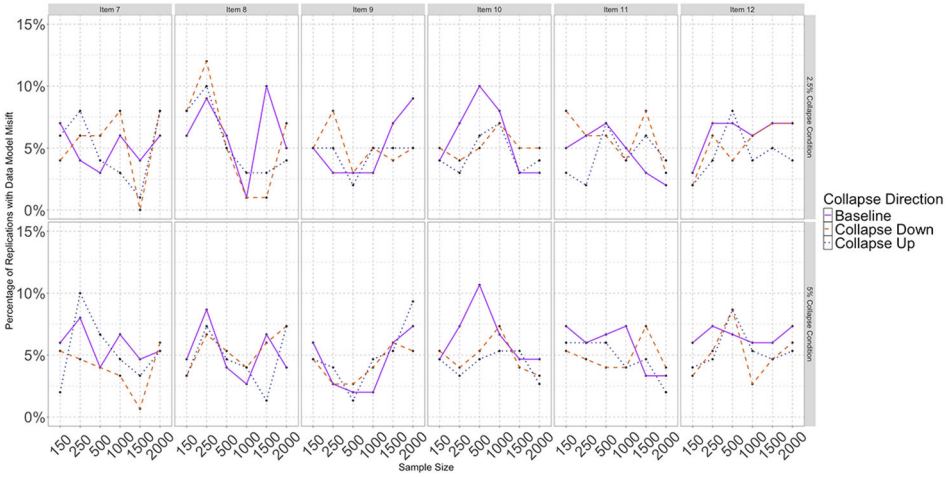
Category Collapse and Data-Model Fit

Figures 2 and 3 below display the percentage of simulations (out of 500) that an item containing collapsed categories was flagged for data-model misfit using an alpha level of $\alpha = 0.05$. When examining GRM data-model fit (Figure 2), we observed slightly inflated Type I errors for Items 7–12 in the GRM baseline dataset. This suggests that the $S - X^2$ test is sensitive to sparseness in GRM data. Kang and Chen (2008) also observed inflated Type 1 errors when using the $S - X^2$ with sparse response frequencies. Kang and Chen concluded that despite the inflated Type-I, the $S - X^2$ test can still be used for determining GRM data-model fit. When fitting a GRM to data containing collapsed categories, the data-model fit improves for items containing collapsed categories. In almost all cases, the percentage of flagged simulations decreased after category collapse.

Conversely, when using GPCM with collapsed categories (Figure 3) the percentage of simulations flagged for data-model misfit increased. This result confirms research performed in Harel (2014) by demonstrating that collapsing categories with GPCM data introduces undesirable data-model misfit. Additionally, unique to our study, we can see from Figure 3 that the power to detect this misfit is only present at larger sample sizes. At smaller sample sizes, collapsing GPCM response categories does not seem to worsen the data-model fit.

Figure 2

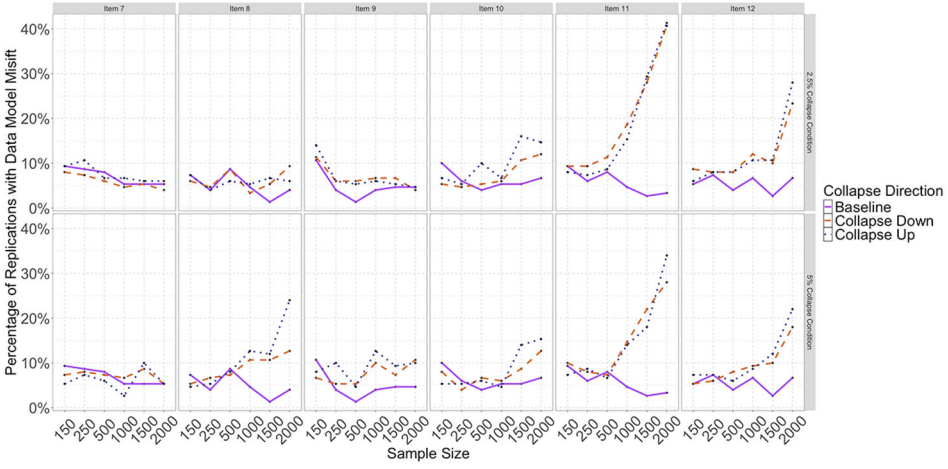
Percentage of Flagged Simulations: Items 7–12. GRM Data



Note. When using a 2.5% endorsement sparseness collapse rule Items 10–12 contained collapsed categories and using a 5% endorsement sparseness collapse rule Items 7–12 contained collapsed categories.

Figure 3

Percentage of Flagged Simulations: Items 7–12. GPCM Data



Note. When using a 2.5% endorsement sparseness collapse rule Items 10–12 contained collapsed categories and using a 5% endorsement sparseness collapse rule Items 7–12 contained collapsed categories.

Category Collapse and Person Parameter Recovery

The person parameter (θ) was well recovered when fitting a graded response model for all sample size, collapse direction, and sparseness conditions. The average bias of recovered person parameters over 500 simulations was very close to zero in all situations. Additionally, collapsing response Category 3 resulted in a decrease of the RMSE, suggesting that person parameters were better recovered under the collapsed conditions. These results are presented in Figures 4 and 5. This finding supports the research performed in Jiang (2018). When adjacent response categories are collapsed due to response sparseness person parameter recovery was not affected. These results support that as much as 5% of the total number of responses can be collapsed without influencing GRM person parameter recovery with a sample size as small as 150.

Similar results were observed when examining person parameter recovery when using the GPCM with collapsed categories (Figures 6 and 7). The average bias was very close to zero for all simulation conditions. Standard errors were consistently around 0.30 for all simulation conditions. RMSE of recovered person parameters was smallest when no responses were collapsed. Both collapse conditions resulted in similar RMSE values. However, the increase in RMSE was marginal (> .10). These results suggest that despite the intentional model-misspecification induced by using GPCM with collapsed categories, person parameter recovery is not affected.

Figure 4

Average Bias of Recovered GRM Person Parameters

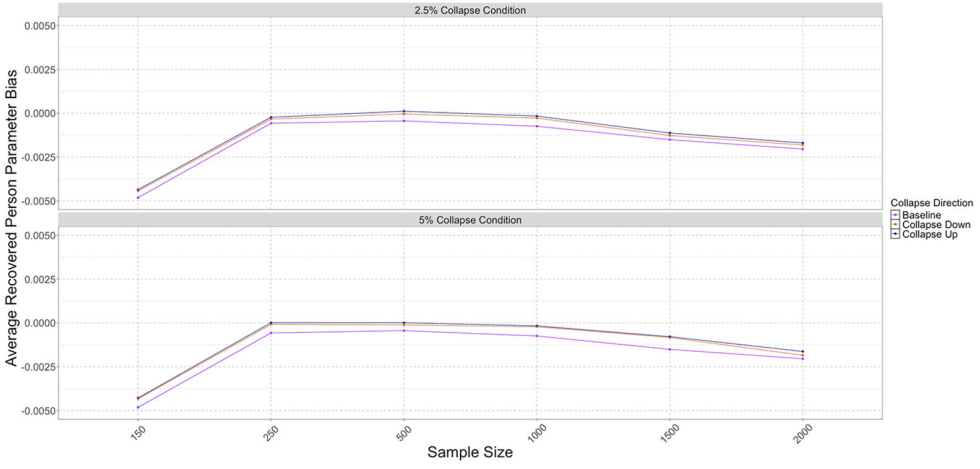


Figure 5

RMSE of Recovered GRM Person Parameters

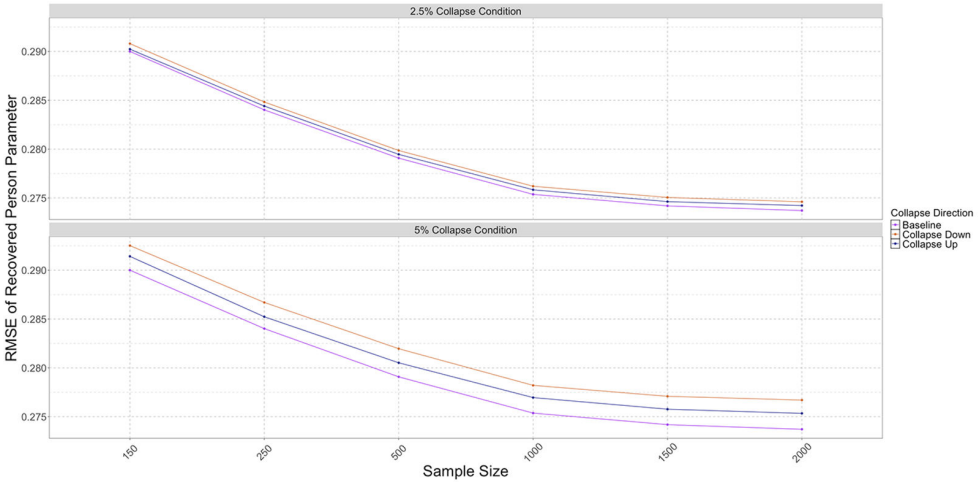


Figure 6

Average Bias of Recovered GPCM Person Parameters

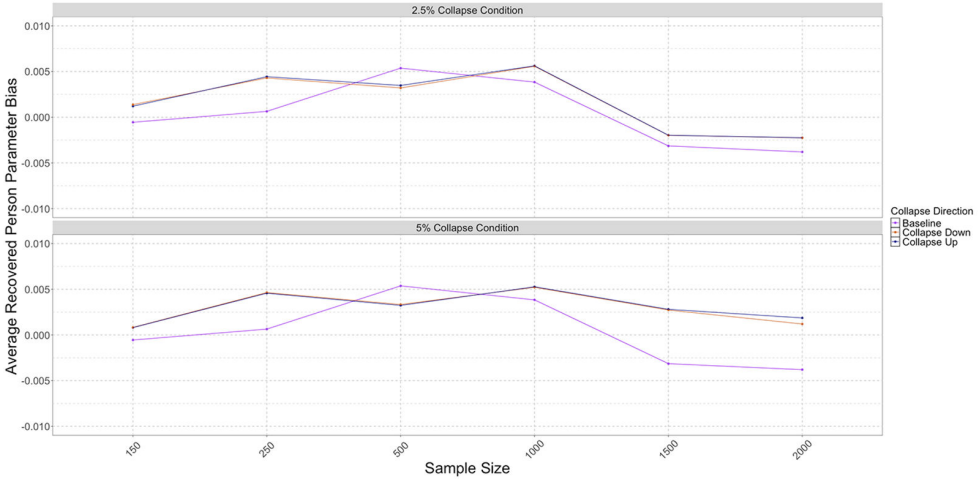
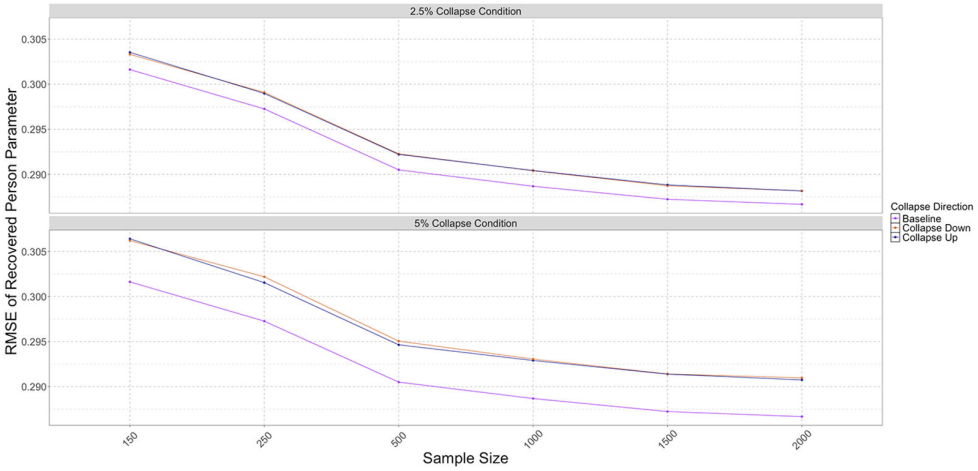


Figure 7

RMSE of Recovered GPCM Person Parameters

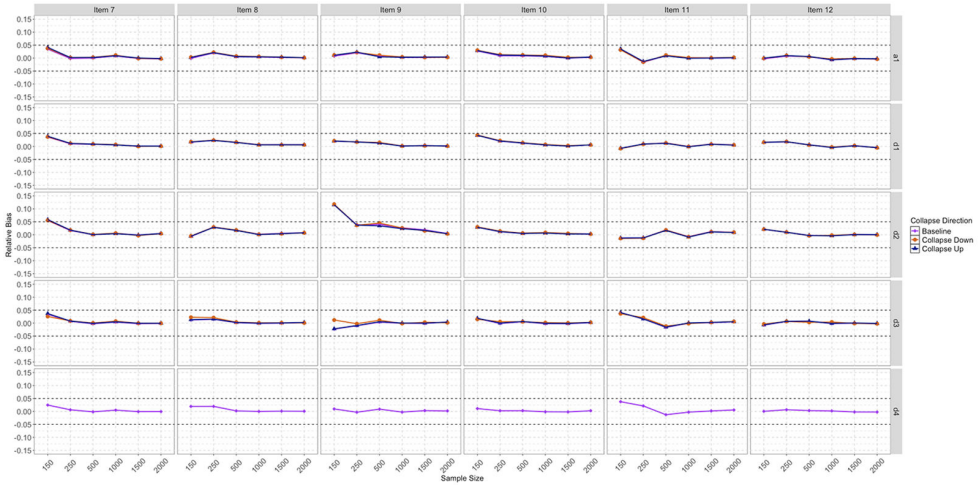


Category Collapse and Item Parameter Recovery

When fitting a Graded Response model using a 5% collapse rule Items 7–12 contain collapsed categories. The slope parameter (a_1) was well recovered for all items containing collapsed categories. All the threshold parameters (d_1, d_2, d_3) were also well recovered from items containing collapsed categories. Using a 2.5% collapse rule Items 10–12 contained collapsed categories. Similar to the 5% category collapse results the slope and threshold parameters were well recovered from all items containing collapsed categories. In general, item parameters were best recovered when response Category 3 was collapsed up into response Category 4. These results are displayed in Figures 8 and 9 respectively.

Figure 8

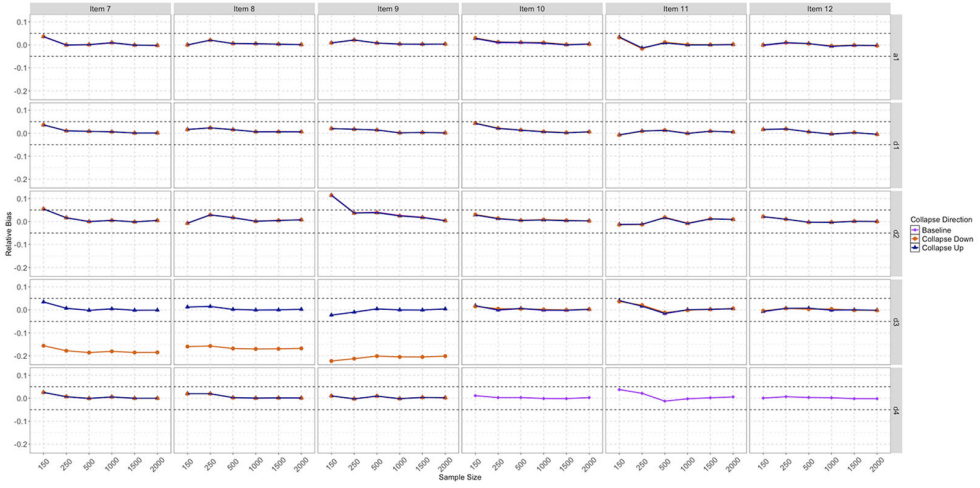
Relative Bias of Recovered GRM Item Parameters 5% Collapse Condition



Note. When using a 5% endorsement sparseness collapse rule Items 7–12 contained collapsed categories. The dashed horizontal lines indicate the |0.05| threshold for extreme relative bias.

Figure 9

Relative Bias of Recovered GRM Item Parameters 2.5% Collapse Condition



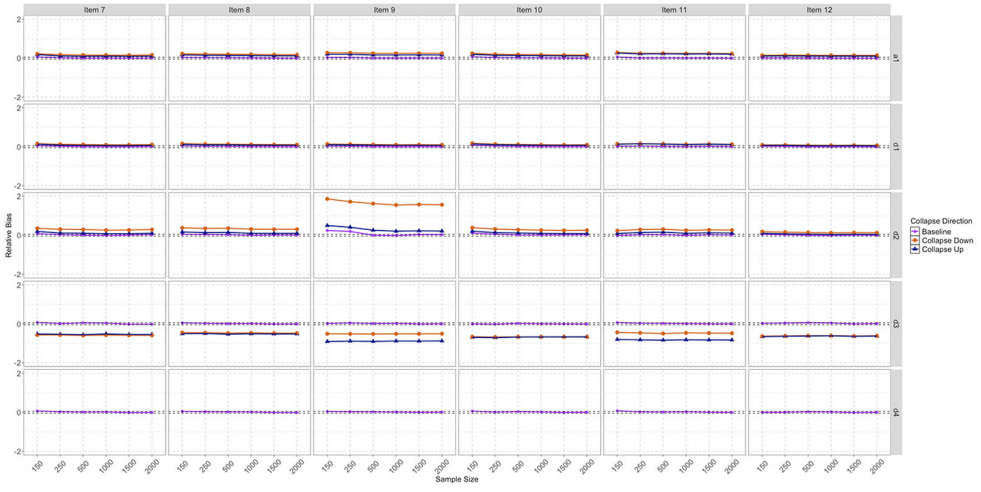
Note. When using a 2.5% endorsement sparseness collapse rule Items 10–12 contained collapsed categories. The dashed horizontal lines indicate the |0.05| threshold for extreme relative bias.

Figures 10 and 11 display the relative bias of recovered item parameters when using a 5% and 2.5% collapse rule respectively for response Category 3 with GPCM. Using a 5% collapse rule, where Items 7–12 have collapsed categories, we can see from Figure 10 that when using data with collapsed categories item parameters are not well recovered. The d_3 parameter displays significant bias regardless of collapse direction, however collapsing up seems to introduce the least amount of bias into the parameter estimate. In addition, the d_2 and the a_1 parameters display significant bias when using collapsed down data. This suggests that using GPCM with collapsed categories not only biases the parameters related to the target response category, but also influences other response categories.

Using a 2.5% collapse threshold (Figure 11) resulted in lower (but still significant) relative bias values. From Figure 11 we can also compare parameter estimates between items without collapsed categories (Items 7–9) and items with collapsed categories (Items 10–12). Using collapsed categories significantly increases the relative bias in item parameter estimates. All item parameters display significant relative bias values regardless of collapse direction. Similar to the 5% collapse condition, collapsing response Category 3 up induces lower (but still significant) relative bias values.

Figure 10

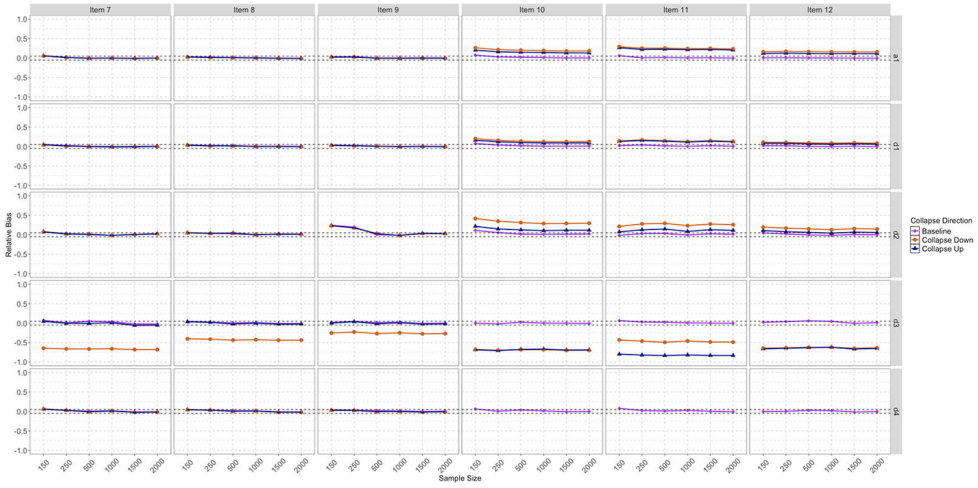
Relative Bias of Recovered GPCM Item Parameters 5% Collapse Condition



Note. When using a 5% endorsement sparseness collapse rule Items 7–12 contained collapsed categories. The dashed horizontal lines indicate the |0.05| threshold for extreme relative bias.

Figure 11

Relative Bias of Recovered GPCM Item Parameters 2.5% Collapse Condition



Note. When using a 2.5% endorsement sparseness collapse rule Items 10–12 contained collapsed categories. The dashed horizontal lines indicate the |0.05| threshold for extreme relative bias.

Empirical Example of Category Collapse

In this section we analyzed two-cohort repeated measure data from the Alzheimer's Disease Neuroimage Initiative (ADNI). 1,656 ADNI participants were recruited from 57 sites in the United States and Canada. Participants were between the ages of 55 and 90. Participants responded to series of initial tests that were repeated at intervals over subsequent years. We used cognitive battery data from 2 ADNI cohorts: ADNI1 and ADNI2/ADNI GO. Specifically, we focus on the memory section (ADNI-MEM) of the cognitive battery.

To address MCMC convergence errors, cognitive battery polytomous item response categories were combined if less than 20 individuals endorsed a particular category. Dichotomous response items were dropped if less than 20 individuals endorsed an option. ADNI1 collapsed categories were used for any shared items with ADNI2/ADNI GO. Table 2 below shows how the items measuring memory were recoded (Gibbons et al., 2012; Wang et al., 2023).

Methodology
2025, Vol. 21(1), 46–73
https://doi.org/10.5964/meth.14303

Table 2

Recoded Response Categories ADNI-MEM in Gibbons et al.

	Recoded Score									
ADNI-MEM Measure	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
RAVLT Trial 1	0–2[a]	3	4	5	6	7	8	9	10	11–15[a]
RAVLT Trial 2	0–2[a]	3	4	5	6	7	8	9	10	11–15[a]
RAVLT Trial 3	0–2[a]	3	4	5–6[a]	7–8[a]	9	10	11	12	13–14[a]
RAVLT Trial 4	0–3[a]	4	5–6[a]	7–8[a]	9	10	11	12	13	14–15[a]
RAVLT Trial 5	0–3[a]	4	5	6–7[a]	8–9[a]	10–11[a]	12	13	14	15
Interference	0–1[a]	2	3	4	5	6	8	7	8–15[a]	
Immediate recall	0	1–2[a]	3–4[a]	5–6[a]	7	8	9	10–11[a]	12–13[a]	14–15[a]
30-minute delay	0	1–2[a]	3–4[a]	5–6[a]	7	8	9	10–11[a]	12–13[a]	14–15[a]
Recognition	0	1	2–3[a]	4–5[a]	6–7[a]	8–9[a]	10–11[a]	12–13[a]	14	15
ADAS Cog – Trial 1	0–1[a]	2	3	4	5	6	7	8–10[a]		
ADAS Cog – Trial 2	0–2[a]	3	4	5	6	7	8	9	10	
ADAS Cog – Trial 3	0–2[a]	3	4	5	6	7	8	9	10	
Recall	0	1	2	3	4	5	6	7	8	9–10[a]
Recognition present	0–3[a]	4	5	6	7	8	9	10	11	12
Recognition absent	0–4[a]	5–6[a]	7	8	9	10	11	12		
Logical Memory – Immediate	0–1[a]	2–3[a]	4–5[a]	6–7[a]	8–9[a]	10–12[a]	13–14[a]	15–16[a]	17–18[a]	19–25[a]
Logical Memory – Delay	1	1–2[a]	3–4[a]	5–8[a]	9–11[a]	12	13	14–15[a]	16–17[a]	18–25[a]
MMSE Ball Recall	2	1								
MMSE Flag Recall	2	1								
MMSE Tree Recall	2	1								

Note. Cell values indicate recoded score values.

[a] indicates that adjacent response categories were combined.

A 2-Parameter Logistic (2PL) model was fit to items with binary responses and a Graded Response Model (GRM) was fit to items with polytomous responses. GRM was selected for polytomous responses due to the Likert-type data collected. This fitting was performed twice: once on the original dataset and once on the collapsed dataset. Person parameters were extracted from each model fitting. Bias, Root Mean Square Error (RMSE), and average Standard Error (SE) was calculated with the original dataset serving as the "true" person parameters.

From Table 3 we can see that collapsing response categories did not have a negative effect on person parameter recovery. The bias between person parameter estimates from the original dataset (without collapsed categories) to person parameter estimates using collapsed data was very close to zero for all four datasets. The RMSE was also very close to zero indicating that the person parameter was well recovered after collapsing response categories. The average standard error of person parameter estimates increased slightly after collapsing response categories, but the increase was not significant.

Table 3

ADNI-MEM Person Parameter Recovery Summary Statistics

Data	Bias	RMSE	Average SE Change
ADNI1 Baseline	0.001	0.057	0.007
ADNI1 Follow up	0.001	0.063	0.008
ADNI2 Baseline	0.004	0.054	0.005
ADNI2 Follow up	0.004	0.057	0.007

Note. Table values indicate the Bias, RMSE, and Average SE Change between the original dataset and the dataset containing collapsed categories.

Practical Implications for Researchers

This study examines the impact category collapse has on IRT parameter recovery and IRT data-model fit. Our study expands the literature by exploring parameter recovery and data-model fit for the Generalized Partial Credit IRT model along with the Graded Response IRT model.

In practice, since the true data generating model is unknown, the candidate IRT models (e.g., GRM or GPCM) are often selected based on item types. For instance, data collected from a Likert-type item would be appropriate for GRM, while scores from a constructed response item would be appropriate for GPCM. We provide the following example for researchers to consider when deciding which candidate IRT model to use.

Consider an assessment scored from 0 to 4 where students are asked to solve a constructed response math item. When scoring student work, students who earn a 3 have also successfully completed enough work to earn a score of 1 and 2. This assumption leads us to recommend the GPCM as a candidate IRT model. In contrast, consider a

Likert-style survey item with responses "Disagree," "Neutral," and "Agree." This would lead us to recommend the GRM as a candidate IRT model.

We caution researchers against applying both GRM and GPCM to their data and selecting the one with the best-fitting results because we need to interpret model parameters that align with the specific item types. Researchers can use a statistical test such as the $S - X^2$ test to confirm if their proposed IRT model fits their data prior to data collapse. We strongly recommend that researchers confirm that their observed item response data fit their proposed IRT model prior to category collapse.

In sum, we provide the following recommendations to practitioners: Firstly, if the observed item response data comes from Likert scale items, research practitioners can fit a GRM to a collapsed dataset. There was no significant data-model misfit introduced for any items containing collapsed categories. Secondly, if GRM is used, practitioners may use either the 2.5% or 5% endorsement threshold when deciding to collapse adjacent response categories. Practitioners should collapse the targeted response category down into the next lowest adjacent category. This combination resulted in the least bias in recovered IRT person and item parameters.

Thirdly, if the observed item response data best fits a GPCM we caution against fitting a GPCM to a collapsed dataset. This process introduced significant data-model misfit for larger sample sizes along with substantial relative bias in recovered person and item parameters. Lastly, if researchers wish to collapse categories when using GPCM we recommend that researchers collapse the target response category up into the next highest response category. This process still produces significant relative bias values in recovered item and person parameters, but the bias is lower when compared to collapsing down.

Conclusion

The purpose of this study is to extend the literature on category collapse by investigating, through rigorous Monte Carlo simulations, if category collapse can be justified when utilizing the Graded Response (GRM) and Generalized Partial Credit (GPCM) Item Response Theory (IRT) models. From our extensive simulation study, we concluded that when using the Graded Responses model adjacent response categories can be combined without biasing IRT parameter estimation. In contrast, when using a Generalized Partial Credit model with data containing collapsed categories IRT parameters were not well recovered and significant data-model misfit was introduced into the analysis.

Funding: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D200015 and R305D240021 to University of Washington.

Acknowledgments: The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Competing Interests: The authors have declared that no competing interests exist.

References

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *BMJ, 332*, Article 1080. https://doi.org/10.1136/bmj.332.7549.1080

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

García-Pérez, M. A. (2017). An analysis of (dis)ordered categories, thresholds, and crossings in difference and divide-by-total IRT Models for ordered responses. *Spanish Journal of Psychology, 20*, Article E10. https://doi.org/10.1017/sjp.2017.11

Gibbons, L. E., Carle, A. C., Mackin, R. S., Harvey, D., Mukherjee, S., Insel, P., Curtis, S. M., Mungas, D., & Crane, P. K. (2012). A composite score for executive functioning, validated in Alzheimer's Disease Neuroimaging Initiative (ADNI) participants with baseline mild cognitive impairment. *Brain Imaging and Behavior, 6*(4), 517–527. https://doi.org/10.1007/s11682-012-9176-1

Harel, D. (2014). *The effect of model misspecification for polytomous logistic adjacent category item response theory models* [Doctor of Philosophy Dissertation, McGill University]. https://escholarship.mcgill.ca/concern/theses/jh343w167

Harel, D., & Steele, R. J. (2018). An information matrix test for the collapsing of categories under the partial credit model. *Journal of Educational and Behavioral Statistics, 43*(6), 721–750. https://doi.org/10.3102/1076998618787478

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Currents in Pharmacy Teaching & Learning, 7*(6), 836–850. https://doi.org/10.1016/j.cptl.2015.08.001

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329–367. https://doi.org/10.1177/0049124198026003003

Jansen, P. G. W., & Roskam, E. E. (1986). Latent trait models and dichotomization of graded responses. *Psychometrika, 51*(1), 69–91. https://doi.org/10.1007/BF02294001

Jiang, S. (2018). *The different effects of collapsing categories on graded response model and generalized partial credit model.* American Educational Research Association Annual Conference.

Kang, T., & Chen, T. T. (2008). Performance of the Generalized $S\text{-}X^2$ Item Fit Index for polytomous IRT models. *Journal of Educational Measurement, 45*(4), 391–406. https://doi.org/10.1111/j.1745-3984.2008.00071.x

Kim, D.-H., Wang, C., & Ng, K.-M. (2010). A Rasch Rating Scale modeling of the Schutte Self-Report Emotional Intelligence Scale in a sample of international students. *Assessment, 17*(4), 484–496. https://doi.org/10.1177/1073191110376593

Linden, K., Berg, M., Sparud-Lundin, C., Adolfsson, A., & Melin, J. (2020). Initial validation of the Diabetes and Breastfeeding Management Questionnaire (DBM-Q). *International Journal of Environmental Research and Public Health, 17*(9), Article 3044. https://doi.org/10.3390/ijerph17093044

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*(1), 19–40. https://doi.org/10.1037/1082-989X.7.1.19

Matovu, M. (2019). A validation of the Assessment Practices Inventory Modified (APIM) Scale using Rasch measurement analysis. *Interdisciplinary Journal of Education, 2*(2), 116–135. https://doi.org/10.53449/ije.v2i2.85

Muraki, E. (1992). A Generalized Partial Credit Model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

National Center for Education Statistics. (2008). *NAEP analysis and scaling—Combining the categories of constructed-response items*. National Center for Education Statistics. https://nces.ed.gov/nationsreportcard/tdw/analysis/2000_2001/scaling_checks_compar_poor_comb.aspx

Outpatient and Ambulatory Surgery CAHPS. (2017). *Options for reporting "About You" response data to HOPDs and ASCs when there are fewer than 11 responses in one or more categories*. OAS CAHPS. https://oascahps.org/General-Information/Announcements/EntryId/55/Options-for-Reporting-About-You-Response-Data-to-HOPDs-and-ASCs-when-there-are-Fewer-than-11-Responses-in-One-or-More-Categories

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Rost, J., & Von Davier, M. (1995). Mixture distribution Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models* (pp. 257–268). Springer. https://doi.org/10.1007/978-1-4612-4230-7_14

Rutkowski, L., Svetina, D., & Liaw, Y.-L. (2019). Collapsing categorical variables and measurement invariance. *Structural Equation Modeling, 26*(5), 790–802. https://doi.org/10.1080/10705511.2018.1547640

Samejima, F. (1968). Estimation of latent ability using a response pattern of graded scores. *ETS Research Bulletin Series, 1968*(1), i–169. https://doi.org/10.1002/j.2333-8504.1968.tb00153.x

Smith, E. V., Wakely, M. B., De Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement, 63*(3), 369–391. https://doi.org/10.1177/0013164403063003002

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567–577. https://doi.org/10.1007/BF02295596

Tsai, C.-L., Wind, S., & Estrada, S. (2024). Exploring the effects of collapsing rating scale categories in polytomous item response theory analyses: An illustration and simulation study. *Measurement: Interdisciplinary Research and Perspectives, 23*(1), 66–89. https://doi.org/10.1080/15366367.2023.2288791

Wang, C., Zhu, R., Crane, P. K., Choi, S.-E., Jones, R. N., & Tommet, D. (2023). Using Bayesian item response theory for multicohort repeated measure design to estimate individual latent change scores. *Psychological Methods*. https://doi.org/10.1037/met0000635

Wetzel, E., & Carstensen, C. H. (2014). Reversed thresholds in partial credit models: A reason for collapsing categories? *Assessment, 21*(6), 765–774. https://doi.org/10.1177/1073191114530775

Wind, S. A. (2023). Exploring rating scale functioning for survey research. SAGE Publications.

Appendices

Appendix A: GRM and GPCM Parameter Estimation GRM Threshold Parameters

We begin by deriving the GRM complete-data log-likelihood function and respective partial derivatives. Let X_i be the observed item response vector for person i, with $i = 1...N$ and $j = 1...n$ denoting the sample size and test length respectively. The joint maximum likelihood is:

$$L(x_1, x_2, ..., x_N|a, d, \theta) = \prod_{j=1}^{n}\left[\prod_{k=0}^{K-1} P_{jk}\left(x_{ij} = k\middle|a_j, d_k, \theta_i\right)^{I_{(x_{ij}=k)}}\right] \tag{A1}$$

For notational simplicity let $P(x_i|\theta_i, \epsilon) = \prod_{k=0}^{K-1} P_{jk}\left(x_{ij} = k \mid a_j, d_k, \theta_i\right)^{I_{(x_{ij}=k)}}$, with ϵ being the $n \times 2$ matrix of item parameters. The marginal likelihood of X is:

$$L = \prod_{i=1}^{N}\int P(x_i|\theta_i, \epsilon)\cdot g(\theta|\tau) \quad d\theta \tag{A2}$$

Where $g(\theta|\tau)$ is the normal density. We further simplify the notation by letting $P(x_i, \epsilon) = \prod_{j=1}^{n}\int P(x_i|\theta, \epsilon)\cdot g(\theta|\tau) \quad d\theta$. Then, to finally find the MMLE of d_k we take the respective partial derivative of the log likelihood.

$$\frac{\partial}{\partial d_k}(\log L) = \frac{\partial}{\partial d_k}\log(P(x_i, \epsilon)) \tag{A3}$$

$$= \sum_{i=1}^{N}\frac{1}{P(x_i, \epsilon)}\cdot\int\left(\frac{\partial}{\partial d_k}P(x_i|\theta, \epsilon)\right)g(\theta|\tau) \quad d\theta \tag{A4}$$

$$= \sum_{i=1}^{N}\frac{1}{P(x_i, \epsilon)}\cdot\int\left(\frac{\partial}{\partial d_k}\left[\log P(x_i|\theta, \epsilon)\right]\right)P(x_i|\theta, \epsilon)\cdot g(\theta|\tau) \quad d\theta \tag{A5}$$

$$= \sum_{i=1}^{N} \int \frac{\partial}{\partial d_k}\left[\log P(x_i|\theta,\epsilon)\right]\left[\frac{P(x_i|\theta,\epsilon)\cdot g(\theta|\ \tau)}{P(x_i,\epsilon)}\right]\ d\theta \tag{A6}$$

Note that using Bayes Theorem we have $\frac{P(x_i|\theta,\epsilon)\cdot g(\theta|\ \tau)}{P(x_i,\epsilon)} = P(\theta|x_i,\tau,\ \epsilon)$.

$$\frac{\partial}{\partial d_k}(\log L) = \sum_{i=1}^{N} \int \frac{\partial}{\partial d_k}\left[\log P(x_i|\theta,\epsilon)\right] P(\theta|x_i,\tau,\epsilon)\ \ d\theta \tag{A7}$$

We now focus on solving $\frac{\partial}{\partial d_k}\left[\log P(x_i|\theta,\epsilon)\right]$.

$$\frac{\partial}{\partial d_k}\left[\log P(x_i|\theta,\epsilon)\right] = \frac{\partial}{\partial d_k}\left[\log \prod_{k=0}^{K-1} P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right)^{I_{(x_{ij}=k)}}\right] \tag{A8}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot \frac{\partial}{\partial_{d_k}}\left[\log P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right)\right] \tag{A9}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot \frac{1}{P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right)}\cdot \frac{\partial}{\partial d_k} P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right) \tag{A10}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot \frac{1}{P_K^*(\theta) - P_{K+1}^*\theta}\cdot \left[P_k^*(\theta)\cdot\left(1-P_k^*(\theta)\right)\right] \tag{A11}$$

After substituting Equation A11 into Equation A7 we obtain:

$$\frac{\partial}{\partial d_k}(\log L) = \sum_{i=1}^{N} \int \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot \frac{1}{P_K^*(\theta) - P_{K+1}^*\theta}\left[P_k^*(\theta)\cdot Q_k^*(\theta)\right]\cdot P(\theta|x_i,\tau,\epsilon)\ \ d\theta \tag{A12}$$

Where $Q_k^*(\theta) = \ 1 - P_k^*(\theta)$.

From Equation A12 we can see that threshold estimation only involves the two adjacent response categories, k and $k+1$. Collapsing the highest two parameters involves removing the threshold parameter corresponding to the highest response option. Similarly, when collapsing into a lower response category, the threshold parameter for the higher response option is removed. After removing the collapsed threshold, a GRM can still be used to estimate threshold parameters because category collapse has not altered the order of the remaining threshold parameters.

GRM Discrimination Parameter

Continuing the derivation presented in the previous section, we now focus on finding the MMLE of the item specific discrimination parameter a_j. To do so we take the respective partial derivative of the log likelihood.

$$\frac{\partial}{\partial a_j}(\log L) = \frac{\partial}{\partial a_{kj}}\log(P(x_i,\epsilon)) \tag{A13}$$

Following a similar derivation as the threshold parameters we obtain

$$\frac{\partial}{\partial a_j}(\log L) = \sum_{i=1}^{N} \int \frac{\partial}{\partial a_j}[\log P(x_i|\theta,\epsilon)] P(\theta|x_i,\tau,\epsilon) \quad d\theta \tag{A14}$$

We now focus on solving $\frac{\partial}{\partial a_j}[\log P(x_i|\theta,\epsilon)]$.

$$\frac{\partial}{\partial a_j}[\log P(x_i|\theta,\epsilon)] = \frac{\partial}{\partial a_j}\left[\log \prod_{k=0}^{K-1} P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right)^{I_{(x_{ij}=k)}}\right] \tag{A15}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \frac{\partial}{\partial_{a_j}}[\log P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right)] \tag{A16}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \frac{1}{P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right)} \cdot \frac{\partial}{\partial a_j} \quad P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right) \tag{A17}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \frac{1}{P_K^*(\theta) - P_{K+1}^*(\theta)} \cdot \left[\theta P_k^*(\theta) Q_k^*(\theta) - \theta P_{k+1}^*(\theta) Q_{k+1}^*(\theta)\right] \tag{A18}$$

Where $Q_k^*(\theta) = \quad 1 - P_k^*(\theta)$. Substituting equation A18 into A14 we obtain:

$$\frac{\partial}{\partial d_k}(\log L) = \sum_{i=1}^{N} \int \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \frac{\left[\theta P_k^*(\theta) Q_k^*(\theta) - \theta P_{k+1}^*(\theta) Q_{k+1}^*(\theta)\right]}{P_K^*(\theta) - P_{K+1}^*(\theta)} \cdot P(\theta|x_i,\tau,\epsilon) d\theta \tag{A19}$$

As in the case of the threshold parameter estimation, the estimation of the GRM discrimination parameter also only involves the two adjacent response categories, k and $k + 1$.Therefore, in theory, category collapse will not influence the estimation of item specific GRM discrimination parameters.

GPCM Threshold Parameters

Let X_i be the observed item response vector for person i, with N and n denoting the sample size and test length respectively. Let P_{ijk} be the probability that student i is assigned score k on Item j, with all items having a maximum score of K. The joint maximum likelihood is:

$$L(x_1, x_2, \ldots, x_N|a,d,\theta) = \prod_{j=1}^{n} \left[\prod_{k=0}^{K} P_{ijk}\left(x_{ij}=k\middle|a_j,d_k,\theta_i\right)^{I_{(x_{ij}=k)}}\right] \tag{A20}$$

With

$$P_{ijk} = \frac{\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]}{\sum_{k=0}^{K} \exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]} \tag{A21}$$

Similar to the GRM derivation, we write the first derivative of the log-likelihood as:

$$\frac{\partial}{\partial d_{jm}}(\log L) = \sum_{i=1}^{N} \int \frac{\partial}{\partial d_m}[\log P(x_i|\theta,\epsilon)] P(\theta|x_i,\tau,\epsilon) \quad d\theta \tag{A22}$$

Where $(x_i|\theta_i,\epsilon) = \prod_{k=0}^{K-1} P_{jk}\left(x_{ij}=k \mid a_j, d_k, \theta_j\right)^{I_{(x_{ij}=k)}}$. Focusing on $\frac{\partial}{\partial d_k}[\log P(x_i|\theta,\epsilon)]$:

$$\frac{\partial}{\partial d_k}[\log P(x_i|\theta,\epsilon)] = \frac{\partial}{\partial d_m}\left[\log \prod_{k=0}^{K-1} P_{jk}\left(x_{ij}=k\middle|a_j,d_k,\theta_j\right)^{I_{(x_{ij}=k)}}\right] \tag{A23}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \frac{\partial}{\partial d_m}\log\left[\frac{\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]}\right] \tag{A24}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \frac{\partial}{\partial d_m}\left[\sum_{m=0}^{k} a_j(\theta_i - d_m) - \log\left(\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]\right)\right] \tag{A25}$$

When calculating the first derivative we consider 2 cases: when an examinee is assigned a score of 0 (i.e., $k=0$) and when the assigned score is above zero (i.e., $k=1,\quad,2\ldots K$).

When $k=0$ we have:

$$\frac{\partial}{\partial d_k}\log\left(P(x_i|\theta,\epsilon)\right) = \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot\left(0 - \frac{1}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]} \times -a_j\sum_{k=t}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]\right) \tag{A26}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot\left(a_j\frac{\sum_{k=t}^{K}\exp\left[\sum_{m=0}^{k} a_j(\theta_i - d_m)\right]}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]}\right) \tag{A27}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot\left(a_j\frac{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j(\theta_i - d_m)\right]}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]} - \frac{\sum_{k=0}^{t-1}\exp\left[\sum_{m=0}^{k} a_j(\theta_i - d_m)\right]}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]}\right) \tag{A28}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot a_j\left(-\sum_{k=0}^{t-1} p_{jk}(\theta)\right) \tag{A29}$$

Similarly, when $k=1,2,\ldots,K$ we have:

$$\frac{\partial}{\partial d_t}\log\left(P(x_i|\theta,\epsilon)\right) = \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot\left(-a_j - \frac{1}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]} \times -a_j\sum_{k=t}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]\right) \tag{A30}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot\left(-a_j + a_j\frac{\sum_{k=t}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]}\right) \tag{A31}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k}\cdot\left(-a_j + a_j\frac{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j(\theta_i - d_m)\right]}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]} - \frac{\sum_{k=0}^{t-1}\exp\left[\sum_{m=0}^{k} a_j(\theta_i - d_m)\right]}{\sum_{k=0}^{K}\exp\left[\sum_{m=0}^{k} a_j\left(\theta_i - d_{jm}\right)\right]}\right) \tag{A32}$$

$$= \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \left(- \sum_{k=0}^{t-1} p_{jk}(\theta) \right) \tag{A33}$$

Combining the results, we obtain the first derivative of the log-likelihood with respect to the threshold parameter of interest as:

$$\frac{\partial}{\partial d_k} \log\left(P(x_i|\theta,\epsilon)\right) = \begin{cases} \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot a_j \left(- \sum_{k=0}^{K-1} p_{jk}(\theta) \right), & k = 0 \\ \sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \left(- \sum_{k=0}^{K-1} p_{jk}(\theta) \right), & k = 1, 2, \ldots, K \end{cases} \tag{A34}$$

After substituting Equation A36 into Equation A7 we obtain two cases for the derivative:

$$\frac{\partial}{\partial d_k}(\log L) = \begin{cases} \sum_{i=1}^{N} \int \left(\sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot a_j \left(- \sum_{k=0}^{K-1} p_{jk}(\theta) \right) \right) \times P(\theta|x_i,\tau,\epsilon) \quad d\theta, & k = 0 \\ \sum_{i=1}^{N} \int \left(\sum_{k=0}^{K-1} I_{x_{ij}=k} \cdot \left(- \sum_{k=0}^{K-1} p_{jk}(\theta) \right) \right) \times P(\theta|x_i,\tau,\epsilon) \quad d\theta, & k = 1, 2, \ldots, K \end{cases} \tag{A35}$$

In contrast to the GRM threshold parameter estimation, we can see that the estimation of GPCM threshold parameters relies on all response categories. Since category collapse involves the removal of a threshold parameter, estimated GPCM threshold parameters using collapsed category data will not be the same as the ones estimated from the original dataset. This introduces data-mode misfit since the resulting dataset after category collapse no longer follows a GPCM item response function.

Appendix B: Data Generating Parameters

Graded Response Model Data Generating Parameters

Item	a	d1	d2	d3	d4
Item 1	1.9682867	3.210208	0.6041567	-1.7253799	-2.78973
Item 2	1.2967538	2.341573	1.261702	-0.2455449	-2.013333
Item 3	3.6240119	6.163613	2.7843951	-1.0279311	-5.787697
Item 4	1.6383015	2.669297	0.6504664	-1.0579291	-2.658706
Item 5	2.6412785	4.14342	0.4016618	-0.9414887	-2.823617
Item 6	1.8363191	3.078195	1.4678899	-1.7952482	-2.747437
Item 7	2.7586827	3.981743	1.7643366	-3.0573067	-3.746977
Item 8	1.6792375	2.140485	0.9788973	-2.3030821	-2.773269
Item 9	1.2427276	2.300461	0.3158956	-1.1612637	-1.459518
Item 10	2.592369	4.816314	1.7304576	-3.0712945	-3.356455
Item 11	0.8634693	1.088787	0.6129386	-0.9727626	-1.102283
Item 12	1.8793678	3.697137	1.7929021	-2.6441636	-2.907275

Generalized Partial Credit Model Data Generating Parameters

Item	a	d1	d2	d3	d4
Item 1	1.9682867	3.210208	0.6041567	-1.7253799	-2.78973
Item 2	1.2967538	2.341573	1.261702	-0.2455449	-2.013333
Item 3	3.6240119	6.163613	2.7843951	-1.0279311	-5.787697
Item 4	1.6383015	2.669297	0.6504664	-1.0579291	-2.658706
Item 5	2.6412785	4.14342	0.4016618	-0.9414887	-2.823617
Item 6	1.8363191	3.078195	1.4678899	-1.7952482	-2.747437
Item 7	2.7586827	3.981743	1.7643366	-1.2641629	-3.746977
Item 8	1.6792375	2.140485	0.9788973	-1.5978024	-2.773269
Item 9	1.2427276	2.300461	0.3158956	-1.0867	-1.459518
Item 10	2.592369	4.816314	1.7304576	-1.6714152	-3.356455
Item 11	0.8634693	1.088787	0.6129386	-2.1384461	-1.102283
Item 12	1.8793678	3.697137	1.7929021	-1.685686	-2.907275