Supplementary Material: Using Latent Class Analysis to Justify a Latent Continuum in Item Development

Jay Verkuilen*[1], Sydne T. McCluskey[1,2], Magdalen Beiting-Parrish[3], Aleksandra Kazakova[1], and Howard T. Everson[1,4]

[1]CUNY Graduate Center

[2]Imagine Learning

[3]Federation of American Scientists

[4]SRI International

Author Note

All correspondence regarding this article should be addressed to Jay Verkuilen, Ph.D. Program in Educational Psychology, City University of New York Graduate Center, 365 Fifth Avenue, New York NY 10016. jverkuilen@gc.cuny.edu

Supplementary Material: Using Latent Class Analysis to Justify a Latent Continuum in Item Development

Data Generation for Examples 1 and 2

Table 1 reports parameter value settings for example 1, the mean and standard deviation for the model parameters across the 100 replications, and the probabilities corresponding to the simulation parameter settings and mean values. Table 2 displays the same information for example 2. All independent variables are generated via a multivariate normal joint distribution in Mplus Monte Carlo simulations, which are then categorized to create categorical independent variables (the latent classes) for latent class analysis per automatic Mplus procedure. The dependent variables were then generated using maximum likelihood estimation of a logistic model using the values of the thresholds set in the simulation (see Tables 1 and 2). The simulations took less than one minute on a laptop computer with an 11th Generation Intel Core i7-1165G7 at 2.80GHz, 2803 Mhz with 4 Cores, 8 Logical Processor on a Windows 11 operating system with 16GB of RAM.

The simulation and data analysis Mplus code for these examples is provided in the Example 1 and Example 2 folders in the Supplementary Materials. Each folder includes a file which runs the simulation, a file which fits an LCA model to the first simulated dataset, and a file which fits a 3pl model to the first simulated dataset.

IRT Output for Examples 1 and 2

This section presents the parameter estimates from the 3pl IRT models fit in examples 1 and 2 in the main paper.

Simulation Studies

To assess how our proposed contrasts work in practice, we conducted three simulation studies schematically based on the data from Example 2 in the main paper. In real practice, we would suggest that novel items be considered with well-studied anchor items that can improve identification.

The simulation studies were carried out using Mplus version 8.7 (Muthén & Muthén, 1998-2021) to observe the behavior of IRT and LCA analyses under different circumstances. In all three simulation studies, 100 replications of 10 binary items were generated with sample sizes of 5000. Code for all three simulations is provided below, along with graphical and numerical results of the studies and subsequent analyses.

Code for each simulation study can be found in the Supplemental Materials in the Sim Study 1, Sim Study 2, and Sim Study 3 folders, respectively. Each folder includes three Mplus code files: one that simulates the data, one that fits LCA to the first simulated dataset, and one that fits IRT to the first simulated dataset.

Sim Study 1

In the first study, data were generated according to the unidimensional 3PL IRT model with discrimination parameters between 0.3 and 3, guessing parameters all around 0.25, and first thresholds between -2 and 2. Code for the simulation is available in the supplemental materials. The goal of this study was to observe the behavior of the 3PL and 3-class LCA analyses under "typical" circumstances - that is, examinees are ordered along a latent continuum of ability and no examinee groups perform unexpectedly on any items. No problems with convergence of the Monte Carlo simulation were encountered.

A 3PL IRT model was fit to the data. Parameter estimates from this model (displayed in Table 5) were within normal ranges and fit was good (mean $LR\chi^2(993) = 986.914, p = .55$). The data were analyzed with a 3-class LCA model and fit was also good (mean $LR\chi^2(991) = 1013.302, p = .25$). The 3PL model is preferred, based on information criteria (mean AIC = 53817.203 and 53826.615, mean BIC = 54012.719 and 54035.165 for the 3PL and LCA, respectively). Parameter estimations for the LCA are displayed in Table 6. Class 1 had an overall high class-conditional item difficulty on all items, while class 3 had an overall low class-conditional item difficulty. Class 2 had class conditional item difficulty between classes 1 and 3 on all items. Thus, all items were strictly ordered by class. This shows that the contrasts are good indicators of the presence

of a latent continuum.

Sim Study 2

In the second study, data were generated according to a 3-class LCA where classes are strictly ordered on items 1, 6, 7, 8, 9, and 10 and disordered on items 2, 3, 4, and 5. For the 4 disordered items, the ordering of classes 1 and 2 was reversed; that is, the "middle" class had a lower probability of choosing the correct answer than the "low" class. This was meant to simulate a situation where the group of "middle performance" students had misconceptions which lead them to answer incorrectly on a subset of the items. Again no problems with convergence were encountered with the Monte Carlo simulation.

A 3PL IRT model was fit to the data from simulation 2; fit was poor (mean $LR\chi^2(993) = 2080.5, p < 0.0000$). Items 2, 3, and 7 had very high discriminations, essentially becoming Guttman items (see parameter estimates in Table 7). By contrast, LCA recovered the parameters well; fit was also poor (mean $LR\chi^2(993) = 1098.147, p = .01$), although poor fit is commonly induced in LCA models even with minor differences in observed and expected patterns (see Masyn, 2013 for a thorough discussion of this issue). Results of the LCA analysis are shown in Table 8. The within-class response probabilities for the 6 "normally" functioning items were strictly ordered, as they were simulated. Meanwhile, the conditional probabilities for the 4 disordered items correctly assigned a lower probability of correct response to the "middle" performing group than the "high" performing group. The LCA is preferred based on information criteria (mean AIC = 61404.888 and 60364.867, mean BIC = 61600.404 and 60573.417 for the 3PL and LCA, respectively).

Class	Parameter	Value	Value Prob.	Mean	Mean Prob.	Std.Dev.
1	η_1	0.35	0.41	0.34	0.42	0.12
1	η_2	2.97	0.05	3.97	0.02	3.21
1	η_3	0.69	0.33	0.71	0.33	0.16
1	η_4	1.74	0.15	1.74	0.15	0.20
1	η_5	2.73	0.06	2.80	0.06	0.45
1	η_6	0.79	0.31	0.77	0.32	0.17
1	η_7	0.19	0.45	0.19	0.45	0.14
1	η_8	2.55	0.07	2.84	0.06	1.80
1	η_9	0.11	0.47	0.62	0.35	0.15
1	η_{10}	2.08	0.11	2.13	0.11	0.29
1	γ_1	0.00	0.50	0.02	0.50	0.19
2	η_1	0.14	0.47	0.15	0.46	0.17
2	η_2	-0.40	0.60	-0.40	0.60	0.26
2	η_3	-0.60	0.65	-0.63	0.65	0.20
2	η_4	0.87	0.30	0.88	0.29	0.20
2	η_5	0.53	0.37	0.52	0.37	0.22
2	η_6	-0.57	0.64	-0.57	0.64	0.20
2	η_7	-0.47	0.62	-0.51	0.62	0.18
2	η_8	0.23	0.44	0.22	0.45	0.25
2	η_9	0.43	0.39	0.12	0.47	0.16
2	η_{10}	0.33	0.42	0.31	0.42	0.22
2	γ_2	0.00	0.50	0.01	0.50	0.14
3	η_1	-1.81	0.86	-1.84	0.86	0.22
3	η_2	-2.87	0.95	-3.02	0.95	0.47
3	η_3	-2.33	0.91	-2.36	0.91	0.27
3	η_4	-0.02	0.50	-0.01	0.50	0.13
3	η_5	-1.48	0.81	-1.52	0.82	0.18
3	η_6	-1.26	0.78	-1.25	0.78	0.14
3	η_7	-2.14	0.89	-2.18	0.90	0.26
3	η_8	-2.12	0.89	-2.14	0.89	0.28
3	η_9	-0.85	0.70	-0.86	0.70	0.14
3	η_{10}	-1.88	0.87	-1.92	0.87	0.26

Table 1

Simulation threshold value settings (Value), probabilities corresponding to the value settings (Value Prob.), means of parameter values across 100 replications (Mean), corresponding probabilities (Mean Prob.), and standard deviations of parameters across 100 replications for example 1. ***Is there a line for*** η_{10} ***missing in Class 1?***

Class	parameter	Value	Value Prob.	Mean	Mean Prob.	Std.Dev.
1	η_1	0.09	0.48	0.25	0.44	1.51
1	η_2	0.25	0.44	0.24	0.44	0.21
1	η_3	1.19	0.23	1.21	0.23	0.18
1	η_4	0.45	0.39	0.62	0.35	1.48
1	η_5	1.20	0.23	1.21	0.23	0.18
1	η_6	0.95	0.28	0.95	0.28	0.21
1	η_7	0.62	0.35	0.61	0.35	0.21
1	η_8	1.15	0.24	1.19	0.23	0.47
1	η_9	1.75	0.15	1.82	0.14	0.33
1	η_{10}	-1.65	0.84	-1.68	0.84	0.23
1	γ_1	0.00	0.50	0.03	0.49	0.43
2	η_1	-1.55	0.82	-1.71	0.85	1.42
2	η_2	2.00	0.12	4.71	0.01	5.38
2	η_3	1.55	0.18	1.61	0.17	0.43
2	η_4	-1.25	0.78	-1.26	0.78	0.50
2	η_5	1.75	0.15	1.80	0.14	0.39
2	η_6	2.65	0.07	3.50	0.03	3.24
2	η_7	-0.20	0.55	-0.05	0.51	1.56
2	η_8	0.65	0.34	0.68	0.34	0.27
2	η_9	0.55	0.37	0.57	0.36	0.32
2	η_{10}	-0.45	0.61	-0.47	0.62	0.41
2	γ_2	0.00	0.50	0.06	0.48	0.62
3	η_1	-1.35	0.79	-1.36	0.80	0.24
3	η_2	-0.05	0.51	-0.39	0.60	2.01
3	η_3	0.80	0.31	0.78	0.31	0.20
3	η_4	-2.55	0.93	-3.09	0.96	2.20
3	η_5	1.70	0.15	1.70	0.16	0.24
3	η_6	1.82	0.14	1.83	0.14	0.28
3	η_7	-2.25	0.90	-3.94	0.98	4.15
3	η_8	-0.12	0.53	-0.15	0.54	0.23
3	η_9	-0.45	0.61	-0.48	0.62	0.25
3	η_{10}	0.51	0.38	0.54	0.37	0.24

Table 2

Simulation threshold value settings (Value), probabilities corresponding to the value settings (Value Prob.), means of parameter values across 100 replications (Mean), corresponding probabilities (Mean Prob.), and standard deviations of parameters across 100 replications for example 2.

Parameter	Estimate	S.E.
Q1 Discrimination	2.27	0.80
Q2 Discrimination	3.32	0.40
Q3 Discrimination	1.88	0.30
Q4 Discrimination	0.89	0.12
Q5 Discrimination	2.58	0.28
Q6 Discrimination	0.95	0.11
Q7 Discrimination	1.28	0.28
Q8 Discrimination	2.52	0.27
Q9 Discrimination	1.43	0.49
Q10 Discrimination	2.53	0.36
Q1 Difficulty	0.69	0.14
Q2 Difficulty	-0.03	0.05
Q3 Difficulty	-0.20	0.14
Q4 Difficulty	1.40	0.18
Q5 Difficulty	0.36	0.05
Q6 Difficulty	-0.15	0.09
Q7 Difficulty	-0.17	0.31
Q8 Difficulty	0.20	0.05
Q9 Difficulty	0.91	0.18
Q10 Difficulty	0.32	0.07
Q1 Guessing	0.42	0.05
Q2 Guessing	0.05	0.01
Q3 Guessing	0.17	0.06
Q4 Guessing	0.08	0.02
Q5 Guessing	0.04	0.01
Q6 Guessing	0.11	0.02
Q7 Guessing	0.24	0.11
Q8 Guessing	0.05	0.01
Q9 Guessing	0.32	0.06
Q10 Guessing	0.09	0.03

Table 3

Results of 3PL analysis for example 1.

Parameter	Estimate	S.E.
Q1 Discrimination	0.62	0.13
Q2 Discrimination	2183.24	0.40
Q3 Discrimination	0.65	2.01
Q4 Discrimination	1.60	0.31
Q5 Discrimination	3.33	3.70
Q6 Discrimination	-1.06	0.29
Q7 Discrimination	3.38	1.87
Q8 Discrimination	0.79	0.14
Q9 Discrimination	1.29	0.20
Q10 Discrimination	-1.10	0.16
Q1 Difficulty	-1.16	0.24
Q2 Difficulty	1.79	0.00
Q3 Difficulty	6.21	13.07
Q4 Difficulty	-0.45	0.16
Q5 Difficulty	3.41	2.80
Q6 Difficulty	-2.68	0.54
Q7 Difficulty	0.32	0.17
Q8 Difficulty	1.38	0.26
Q9 Difficulty	0.90	0.13
Q10 Difficulty	0.28	0.11
Q1 Guessing	0.17	0.02
Q2 Guessing	0.36	0.02
Q3 Guessing	0.21	0.06
Q4 Guessing	0.19	0.07
Q5 Guessing	0.17	0.01
Q6 Guessing	0.07	0.02
Q7 Guessing	0.35	0.07
Q8 Guessing	0.13	0.03
Q9 Guessing	0.09	0.02
Q10 Guessing	0.14	0.03

Table 4

Results of 3PL analysis for example 2.

Parameter	Estimate	Std.Dev.
Q1 Discrimination	2.47	0.27
Q2 Discrimination	2.51	0.27
Q3 Discrimination	2.23	0.21
Q4 Discrimination	1.66	0.17
Q5 Discrimination	2.34	0.21
Q6 Discrimination	2.95	0.29
Q7 Discrimination	1.23	0.16
Q8 Discrimination	0.46	0.12
Q9 Discrimination	2.62	0.21
Q10 Discrimination	2.70	0.24
Q1 Difficulty	-0.70	0.11
Q2 Difficulty	0.71	0.03
Q3 Difficulty	0.51	0.04
Q4 Difficulty	-1.01	0.17
Q5 Difficulty	-0.09	0.07
Q6 Difficulty	-0.44	0.06
Q7 Difficulty	0.96	0.07
Q8 Difficulty	-1.71	0.57
Q9 Difficulty	-0.36	0.05
Q10 Difficulty	-0.54	0.07
Q1 Guessing	0.22	0.06
Q2 Guessing	0.12	0.01
Q3 Guessing	0.18	0.02
Q4 Guessing	0.22	0.08
Q5 Guessing	0.17	0.03
Q6 Guessing	0.19	0.04
Q7 Guessing	0.15	0.03
Q8 Guessing	0.22	0.09
Q9 Guessing	0.16	0.03
Q10 Guessing	0.17	0.04

Table 5

Results of 3PL analysis on simulation 1 data. Parameter estimates are the average over all datasets created in simulation 1.

Class	Parameter	Estimate	Std.Dev.
Class 1	Q1	0.99	0.00
	Q2	0.75	0.02
	Q3	0.83	0.02
	Q4	0.98	0.01
	Q5	0.95	0.01
	Q6	0.99	0.00
	Q7	0.64	0.02
	Q8	0.83	0.01
	Q9	0.98	0.01
	Q10	0.99	0.00
Class 2	Q1	0.87	0.02
	Q2	0.25	0.02
	Q3	0.38	0.02
	Q4	0.87	0.01
	Q5	0.62	0.02
	Q6	0.80	0.02
	Q7	0.35	0.02
	Q8	0.75	0.01
	Q9	0.75	0.02
	Q10	0.83	0.02
Class 3	Q1	0.44	0.02
	Q2	0.13	0.01
	Q3	0.21	0.01
	Q4	0.56	0.02
	Q5	0.24	0.01
	Q6	0.30	0.02
	Q7	0.22	0.01
	Q8	0.65	0.02
	Q9	0.27	0.02
	Q10	0.33	0.02

Table 6

Results of 3-class LCA on simulation 1 data. Parameter estimates are the average over all datasets created in simulation 1.

Parameter	Estimate	Std.Dev.
Q1 Discrimination	2.27	1.16
Q2 Discrimination	4.68	2.33
Q3 Discrimination	3.07	0.99
Q4 Discrimination	3.80	1.56
Q5 Discrimination	2.81	0.93
Q6 Discrimination	2.93	1.72
Q7 Discrimination	3.30	1.66
Q8 Discrimination	2.59	1.49
Q9 Discrimination	2.12	1.21
Q10 Discrimination	2.31	1.19
Q1 Difficulty	0.42	0.55
Q2 Difficulty	0.38	0.17
Q3 Difficulty	0.41	0.15
Q4 Difficulty	0.38	0.17
Q5 Difficulty	0.48	0.16
Q6 Difficulty	0.37	0.57
Q7 Difficulty	0.44	0.34
Q8 Difficulty	0.38	0.60
Q9 Difficulty	0.40	0.73
Q10 Difficulty	0.43	0.56
Q1 Guessing	0.36	0.17
Q2 Guessing	0.10	0.09
Q3 Guessing	0.13	0.08
Q4 Guessing	0.11	0.09
Q5 Guessing	0.10	0.08
Q6 Guessing	0.33	0.19
Q7 Guessing	0.41	0.11
Q8 Guessing	0.33	0.19
Q9 Guessing	0.34	0.21
Q10 Guessing	0.34	0.18

Table 7

Results of 3PL analysis on simulation 2 data. Parameter estimates are the average over all datasets created in simulation 2.

Class	Parameter	Estimate	Std.Dev.
Class 1	Q1	0.84	0.01
	Q2	0.89	0.01
	Q3	0.86	0.01
	Q4	0.88	0.01
	Q5	0.81	0.01
	Q6	0.86	0.01
	Q7	0.90	0.01
	Q8	0.84	0.01
	Q9	0.82	0.01
	Q10	0.84	0.01
Class 2	Q1	0.61	0.02
	Q2	0.02	0.01
	Q3	0.12	0.01
	Q4	0.08	0.01
	Q5	0.06	0.01
	Q6	0.63	0.02
	Q7	0.58	0.01
	Q8	0.63	0.01
	Q9	0.66	0.02
	Q10	0.61	0.01
Class 3	Q1	0.39	0.02
	Q2	0.41	0.02
	Q3	0.37	0.01
	Q4	0.39	0.02
	Q5	0.37	0.02
	Q6	0.34	0.02
	Q7	0.42	0.02
	Q8	0.35	0.02
	Q9	0.38	0.01
	Q10	0.38	0.02

Table 8

Results of 3-class LCA on simulation 2 data. Parameter estimates are the average over all datasets created in simulation 2.

Sim Study 3

In the third study, data were generated according to a 3-class LCA where classes were weakly ordered on many items, strictly ordered on a few, and disordered on few. This was meant to mimic a situation where separation between classes is sometimes poor and examinees in the "middle performance" group are strongly attracted to an incorrect answer

choice on some items. Again, no problems with convergence were encountered with the Monte Carlo simulation.

Fit of the 3pl model was poor (mean $LR\chi^2(993) = 1189.491, p < 0.0000$), and the discrimination parameter for item 2 is a Heywood case. See parameter estimates for the 3PL model fit to to simulation 3 date in Table 9. By contrast, LCA once again recovered the parameters and fit was good (mean $LR\chi^2(991) = 972.312, p = .66$). The conditional probabilities were strictly ordered on items 4, 7, 8 and 9; weakly ordered on items 1, 3, and 5; and disordered on items 2, 6, and 10 (see Table 10 for model results). These are exactly the patterns created in the simulation. The LCA was again preferred based on information criteria (mean AIC = 58420.089 and 58185.549, mean BIC = 58615.604 and 58394.099 for the 3PL and LCA, respectively).

Supplemental Examples

The data used in supplementary examples 1 and 2 are simulated to closely adhere to the characteristics of items from the pilot study that motivated this research. The pilot study was of a mathematics assessment aimed at students in remedial college mathematics courses. The novel items were designed to test conceptual understanding of algebra and differed from those typically encountered by the examinees. They were administered to examinees at a large public urban community college in the Northeastern United States via a pencil and paper format. The examinee group whose responses the simulation is based on were ethnically and linguistically diverse, generally of low socioeconomic status as indicated by financial aid qualification, and in their early 20s. Both groups of examinees comprised approximately 700 students and has quite modest amounts of missingness. Therefore both simulations were rounded to N = 700. We consider two examples:

- Supplementary Example 1 is based on early prototype items that were administered exclusively in remedial algebra classes. Mplus code for this example is available in the "Supplementary Example 1" folder in the Supplementary Materials. This folder includes the simulation file as well as an LCA analysis file. Readers can optionally

remove the comment on line 14 of the LCA analysis file to fit the model with Bayesian estimation (with default priors) instead of the default maximum likelihood estimation.

- Supplementary Example 2 is based on a larger set of revised items administered to students in a broader range of course levels, from remedial algebra to more advanced classes such as calculus and linear algebra. This study also included some anchors taken from released NAEP mathematics items and made use of a four form matrix sampling design to alleviate examinee burden. Mplus code for this example is available in the "Supplementary Example 2" folder in the Supplementary Materials. This folder includes the simulation file, a 3pl IRT analysis file, an LCA analysis with maxmum likelihood estimation file, and an LCA analysis file with Bayesian estimation and calculation of the contrasts.

Supplementary Example 1: Absence of a Latent Continuum

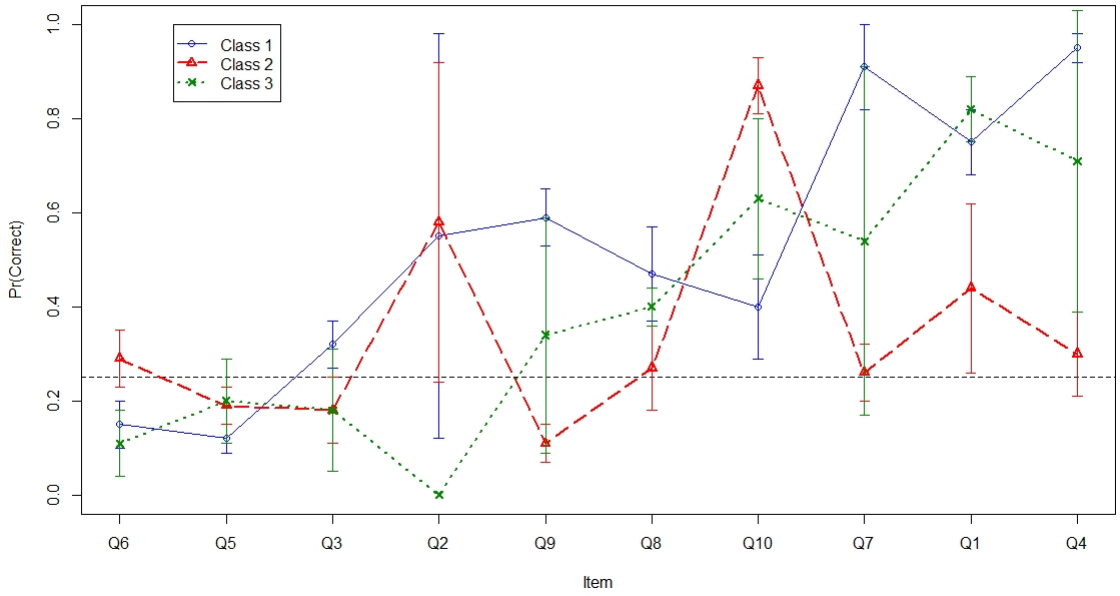
As previously mentioned, this example is based on early pilot data administered to a fairly homogeneous sample, at least in terms of ability. Results from preliminary CTT and Mokken analyses are presented in Table 1 in the main paper; based on the extremely low value of .10 for Cronbach's alpha, it is clear a simple total score does not adequately summarize the data. Furthermore, none of the Loevinger's H item coefficients approach the suggested lower boundary of 0.3; consequently neither does the scale H coefficient. The items have insufficient homogeneity to form a scale. Based on these findings, it would not be appropriate to proceed with CTT or parametric unidimensional IRT analyses for these data, although we did check to be sure and IRT models were similarly poor. Some of this may be due to range restriction.

Because the sum score does not provide an adequate summary of the data, we instead considered LCA. We fit LCA models with $K = 1, \ldots, 5$ classes; fit indices are presented in Table 11. $K = 1$ is a model that presumes independence and is a benchmark.

Models $K = 5$ seemed to be too complicated for the data to support and encountered convergence problems. The 2-, 3-, and 4-class models all fit the data reasonably well. We selected the 3-class model based on a balance of penalized information criteria, the VLMR-LRT p-value, the Bayes Factor, and $cm\hat{P}_k$. Regarding information criteria, BIC and AWE favor the 2-class model strongly to the 3- and 4-class models, while AIC just barely favors the 3-class over the 4-class solution and strongly over the 2-class solution. The Vuong-Lo-Mendell-Rubin Likelihood Ratio Test (VLMR-LRT) compares the fit of the K-class estimated model to that of the $K - 1$-class model; its p-value indicates the probability the data were generated by the simpler model. The 3-class model is the first with a large p-value ($p = .47$), thus the 2-class model is the simplest model not rejected by the test. The Bayes Factor (BF) compared the fit of the estimated K-class model to that of the $K + 1$-class model; higher values indicate greater evidence in favor of the simpler model. The 3-class model is the simplest for which there is strong evidence based on the BF. The Approximate Model Correct Probability ($cm\hat{P}_k$) gives the estimated probability each model in a set is correct (assuming one of the models *is* correct). $cm\hat{P}_k$ strongly favors the 2-class model, with the 3-class model a distant second. Entropy is a measure of classification precision, where 0 indicates classification is no better than chance and 1 indicates perfect classification. Although it is not recommended to use entropy for model selection, extremely low values can indicate classes are poorly separated and therefore a simpler model may be preferred.

The latent class analysis revealed three groups that we characterize in the following way: class 1 (39%): Students who likely have both good procedural skills and emergent conceptual understanding. Class 2 (23%): Students whose answers to most of the items are indistinguishable from random guessing, likely because of low knowledge and/or lack of motivation. Class 3 (39%): Students who likely have some good procedural skills but limited conceptual understanding. From examining the profile plot for these data (see Figure 1), we see the classes do not have a stable ordering. Class 2 has the worst

Figure 1
Profile plot for 3-class LCA for Example 1.



performance overall, but the best performance on the most difficult item (6) and outperforms class 3 on the four most difficult items. Students in class 2 answer correctly at rates consistent with chance on questions 6, 3, 2, 8, and 4; the correct response rate is only greater than .5 on two items (.58 and .87, on items 2 and 10, respectively). It is noteworthy that, of the 3 questions where class 2's performance exceeds chance, two (10 and 1) are among the easiest questions. Thus, it seems likely the ordering of classes reverses on the most difficult items because classes 1 and 3 were drawn to attractive distractors, while students in class 2 were guessing.

The prototypes were designed to test conceptual (rather than procedural) understanding and are intentionally constructed differently than traditional items to which students were accustomed. However, some are closer to traditional procedural questions. Questions 1, 4, 7, 8, 9, and 10 (the easiest items) are more similar to standard problems and procedures than questions 2, 3, 5, and 6, which use more abstract or non-standard formulations of algebraic ideas. Responses on these more conceptual questions are precisely

what primarily distinguish class 2 from class 3. Class 3 answers significantly worse than chance on questions 2 and 6 because of the presence of attractive distractors that likely tap into student misconceptions related to the misuse of procedures. Classes 1 and 3 are distinguished by improved performance on the items overall as well as fewer misconceptions. Students who passed the course were most likely to be in class 1, then class 3, and least likely to be in class 2. An end-of-course standardized assessment showed a similar outcome. These results shed light on why measures of scalability are so poor and indicate that, while it is not possible to justify a skill continuum for these items, useful information can still be gleaned from examining how students in different latent classes perform on them.

Supplementary Example 2: Item Analysis with Planned Missingness Designs

This example illustrates the value of LCA to manage a matrix sampling design. Matrix sampling, which is a planned missingness design, administers only a randomly selected subset of items to help shorten the assessment while maintaining adequate item exposure. It also helps reduce examinee burden and potential non-response induced by boredom, fatigue or frustration. Examinees asked to respond to too many items often end up begin to perform carelessly, maliciously, or simply drop out entirely. In addition, stakeholders such as schools or classroom teachers may decline to participate if the assessment takes too long. In any case, estimated item parameters cannot be trusted due to the potential for a non-representative sample of items and/or examinees. By introducing missingess by design, we can ensure that missingness is missing at random (MAR) rather than being informatively missing/missing not at random (MNAR). Unfortunately, planned missingness designs require special care when being constructed and require advanced methods to handle the fact that data are missing, e.g., multiple imputation or a latent variable model.

In this example, a total of 34 items were examined; 24 prototype traditional multiple choice items and 10 anchor items taken from released items from a large scale

standardized test that were deemed relevant by curriculum experts. As mentioned previously, recruitment in this case was from the general population of students taking mathematics classes more broadly, not just remedial mathematics. In particular, we wanted to ensure that the items were exposed to a broad variety of students for two reasons: First, we wanted to see if prior course level of the students predicted item performance. Second, we wanted to avoid the restricted ability range shown in Supplementary Example 1. Examinees took one of four test forms, each with twelve prototype items and all ten anchors, which were administered at the end of the assessment.[1] The forms were generated by ordering the items by anticipated difficulty and then assigning odd numbered items to form 1, even items to form 2, first half odd items and second half even items to form 3, and the reverse to form 4. Forms were printed and interleaved to be handed out in classrooms to help ensure that a examinee had an equal chance of getting any form. Examination of the coverage descriptives suggests that the randomization was successful, non-response outside of that implied by matrix sampling was quite minimal, and coverage across course level was also even. Students were assigned to low-, mid-, and high-level based on the mathematics course in which the student took the test. For example, a student taking the test in a calculus course would be assigned to the high-level group, while those taking the test in a college algebra or remedial algebra course would be assigned to the mid- and low-level groups, respectively. Course level was an important piece of validity evidence in our investigation, since, on average, students in higher course levels should have better performance on the test overall as well as on each item. If latent classes are ordered along a continuum, then students in higher-level mathematics courses should also be more likely to be in latent classes ordered higher on the ability continuum.

We considered $K = 1, \ldots, 5$ latent classes. Indicators for the mid- and high-level mathematics courses were predictors of the latent class variable. Examination of fit

[1] It might have been better for these to be interspersed with the prototypes. In addition, this is a very heavy anchor, but the other items were not well-understood so we erred on the side of administering more well-understood ones.

statistics (see Table 12) and theoretical concerns led us to focus on $K = 3$, with an entropy of .85. We interpret the 3 classes based on the model-estimated response probabilities within each class, a subset of which is presented in Figure 2. The classes are ordered in terms of performance; class 1 (53%) has the highest performance overall, while class 3 (13%) has the lowest, and class 2 (34%) performs between classes 1 and 3. The class ordering is also observed in all items, although there are a few items for which some classes are barely distinguishable (see Figure 2, first group of items). Performance of classes 1 and 2 is indistinguishable on Items 7 and 8, among the easiest of the 34 items (this same pattern is observed in several of the easiest items because both classes have almost no incorrect responses). We interpret these results as indicating that the three discrete classes are actually ordered along a latent continuum, an interpretation bolstered by the results of orthogonal polynomial contrasts of within-class response probabilities. This is bolstered by the interpretation of the course-level variable. Class membership was strongly related to course level, particularly for class 1 (the highest performers): The odds ratios for enrollment in a high- and mid-level course for a student assigned to class 1 were 7.073 and 17.876, respectively. Membership in class 2 (students with mid-level performance on the test) was also predicted by course level, although not as strongly as for class 1.

To assess the ordering of items over classes, we calculated contrasts for all items. For the first contrast, probability of correct response for students in class 3 was subtracted from probability of correct response for students in class 1. For the second contrast, probability of correct response for students in class 3 was subtracted from the probability for class 2. For the third contrast, probability of correct response for students in class 2 was subtracted from the probability for class 1. We reasoned that, if the classes do form a continuum, all contrasts should be positive. This was, in fact, observed for all items with a few exceptions (all contrasts are presented in the supplemental materials). The contrasts for classes 1 and 3, as well as those for classes 2 and 3, were statistically significantly different from 0 for all items. For classes 1 and 2, performance on many items was quite high, with probabilities of

correct response in excess of .9; with such high performance in both classes, it is not surprising that some items do not distinguish between the classes. Indeed, contrasts classes 1 and 2 were not significantly different from 0 on items 7 and 8; both items are among the easiest with performance in both classes exceeding 0.98. Posterior distributions for contrasts 2 and 3 is graphed in Figure 3, along with within-class probability of correct response for class 3 (equivalent to a contrast bewteen probability of correct response for students in class 3 and 0). This plot presents the probable values for each contrast and demonstrates that probability of correct response increases as class membership moves from students with low to mid to high overall performance. It also shows that, for the mid and high level students (i.e. classes 2 and 1, respectively), the difference in probability of correct response is practically equivalent to 0. Item 8 performs more or less as expected, but does not separate well between students in the mid- and high-ranges of ability. This is very helpful for making the validity argument about the items.

Figure 2

Profile plot with 95% credible intervals for 3-class LCA for Example 2. Items 20, 7, and 8 have atypical performance; items 6, 17, and 2 have typical performance, and items 25, 30, and 28 are anchors. Items are ordered by marginal easiness.

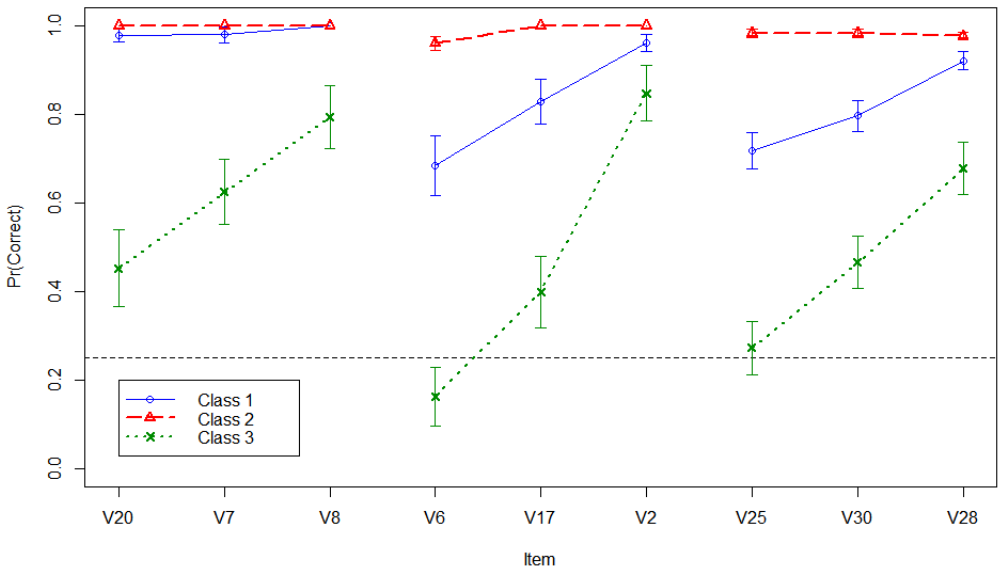
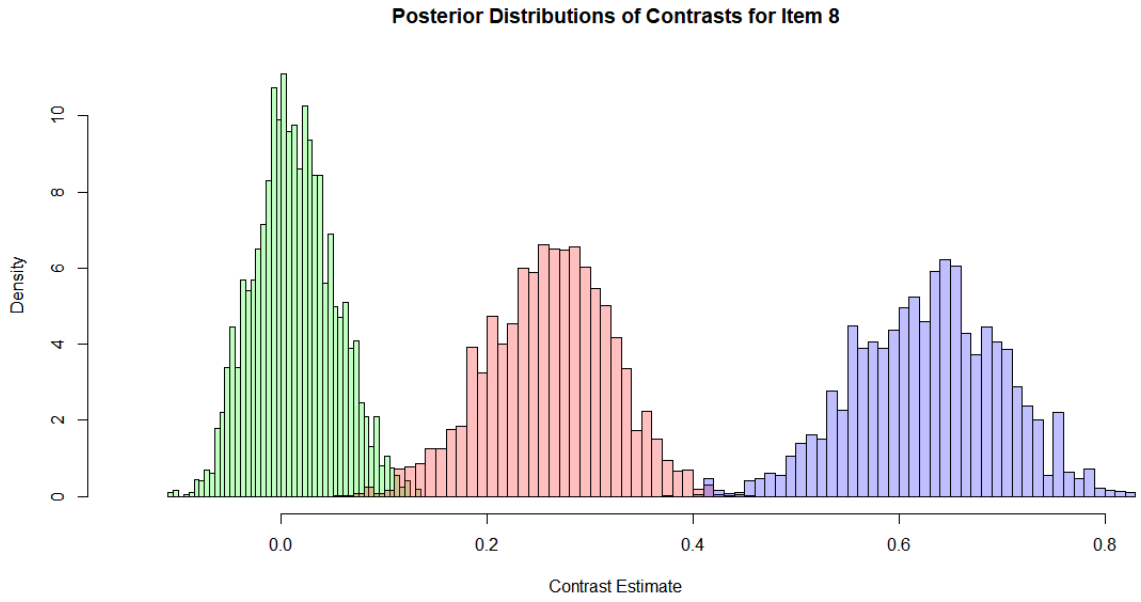


Figure 3

Posterior distributions for contrasts comparing probability of correct response among the classes. Green = class 1 (high)-class2 (mid), orange = class 2- class 3 (low), blue = class 3 - 0



Parameter	Estimate	Std.Dev.
Q1 Discrimination	0.70	0.05
Q2 Discrimination	180.58	451.06
Q3 Discrimination	1.93	4.38
Q4 Discrimination	1.67	0.15
Q5 Discrimination	-0.84	0.51
Q6 Discrimination	-1.97	0.52
Q7 Discrimination	2.05	0.43
Q8 Discrimination	0.76	0.12
Q9 Discrimination	1.12	0.09
Q10 Discrimination	-1.28	0.20
Q1 Difficulty	-1.08	0.09
Q2 Difficulty	2.16	3.43
Q3 Difficulty	3.26	1.10
Q4 Difficulty	-0.53	0.09
Q5 Difficulty	-4.18	1.40
Q6 Difficulty	-2.25	0.20
Q7 Difficulty	0.17	0.11
Q8 Difficulty	1.44	0.14
Q9 Difficulty	0.77	0.06
Q10 Difficulty	0.01	0.16
Q1 Guessing	0.10	0.01
Q2 Guessing	0.28	0.03
Q3 Guessing	0.21	0.02
Q4 Guessing	0.15	0.04
Q5 Guessing	0.13	0.02
Q6 Guessing	0.11	0.01
Q7 Guessing	0.27	0.05
Q8 Guessing	0.13	0.03
Q9 Guessing	0.06	0.02
Q10 Guessing	0.21	0.06

Table 9

Results of 3PL analysis on simulation 3 data. Parameter estimates are the average over all datasets created in simulation 3.

Class	Parameter	Estimate	Std.Dev.
Class 1	Q1	0.79	0.01
	Q2	0.52	0.03
	Q3	0.31	0.02
	Q4	0.93	0.01
	Q5	0.16	0.01
	Q6	0.14	0.01
	Q7	0.91	0.02
	Q8	0.53	0.02
	Q9	0.61	0.02
	Q10	0.37	0.02
Class 2	Q1	0.83	0.02
	Q2	0.02	0.02
	Q3	0.18	0.02
	Q4	0.78	0.02
	Q5	0.15	0.01
	Q6	0.07	0.01
	Q7	0.55	0.03
	Q8	0.34	0.02
	Q9	0.37	0.02
	Q10	0.61	0.02
Class 3	Q1	0.47	0.02
	Q2	0.44	0.02
	Q3	0.23	0.02
	Q4	0.39	0.02
	Q5	0.23	0.01
	Q6	0.28	0.01
	Q7	0.35	0.02
	Q8	0.24	0.01
	Q9	0.15	0.01
	Q10	0.84	0.01

Table 10

Results of 3-class LCA on simulation 3 data. Parameter estimates are the average over all datasets created in simulation 3.

Table 11
Model fit indices for exploratory LCA for example 1. VLMR-LRT: $H_0 : K$ classes; $H_1 : K + 1$ classes

Model	LL	params	BIC	AIC	AWE	$VLMR$	$\hat{BF}_{k,k+1}$	$cm\hat{P}_k$	Entropy
1 class	-4183	10	8432	8386	8453	-	0	0.00	-
2 class	-4082	21	8302	8206	8345	0.00	2	0.68	0.56
3 class	-4055	32	8319	8173	8383	0.47	13	0.29	0.56
4 class	-4044	43	8369	8174	8455	0.16	15	0.02	0.67
5 class	-4035	54	8424	8178	8532	0.29	-	0.00	0.75

Table 12
Model fit indices for Exploratory LCA for example 2. VLMR-LRT $H_0 : K$ classes; $H_1 : K + 1$ classes

Model	LL	params	BIC	AIC	AWE	$VLMR$	$\hat{BF}_{k,k+1}$	$cm\hat{P}_k$	Entropy
1 class	-7324	38	14896	14723	14972	-	0.00	0.00	-
2 class	-5318	71	11101	10778	11242	0.0	0.00	0.00	0.91
3 class	-5133	108	10974	10483	11188	0.0	3039.35	1.00	0.85
4 class	-5092	145	11134	10475	11421	0.6	2726.43	0.00	0.76
5 class	-5050	182	11293	10464	11652	0.7	-	0.00	0.78