Alberta Journal of Educational Research, Vol. 64.1, Spring 2018, 70-87

Deciding Whether to Respond: An Analysis of Nonresponse on Ontario's Grade 9 Assessment of Mathematics

Ruth A. Childs, Orlena Broomes, Monique B. Herbert

Ontario Institute for Studies in Education, University of Toronto

This study investigates nonresponse on Ontario's Grade 9 Assessment of Mathematics—in particular, whether or not students responded to all multiple-choice or all open-response items in two test booklets. Whether students responded to all items of one type (multiple-choice or open-response) by booklet (for the first or second day of testing) was modeled, with and without proportion correct scores by item type as covariates, using latent class analysis. Both a 3-class model without the covariates and a 4-class model with the covariates but without direct effects distinguished among students who responded to all items, students who left both multiple-choice and open-response items blank, and students who left only open-response items blank. The results suggest that deciding to respond to all open-response items is distinct from deciding to respond to all multiple-choice items. Attitudes toward mathematics were also more related to the decision to respond to all open-response items than to the decision to respond to all multiple-choice items.

Cette étude porte sur l'absence de réponse au test de mathématiques pour la 9e année en Ontario —nous cherchions notamment à savoir si les élèves avaient répondu à toutes les questions à choix multiples ou bien à toutes les questions ouvertes dans deux livrets d'examen. Une analyse de structure latente a permis la modélisation du comportement des élèves, à savoir s'ils avaient répondu à tous les items d'un type (questions à choix multiples ou questions ouvertes) dans un livret (lors du premier ou deuxième jour des tests) avec et sans des scores reflétant la proportion de bonnes réponses par type d'items comme covariables. Un modèle de classe 3 sans les covariables ainsi qu'un modèle de classe 4 avec les covariables mais sans effets directs ont tous les deux fait la distinction entre les élèves qui avaient répondu à tous les items, les élèves qui n'avaient ni répondu à certaines questions à choix multiples ni à certaines questions ouvertes et les élèves qui n'avaient pas répondu à certaines questions seulement dans le cas des questions ouvertes. Les attitudes face aux mathématiques ont également joué un plus grand rôle dans la décision de répondre à toutes les questions ouvertes que dans celle de répondre à toutes les questions à choix multiples.

On most large-scale assessments, the majority of the students respond to all the items. However, even on assessments with high stakes, some students leave items blank. This is especially puzzling for multiple-choice items, which require very little time or effort to record a response and, unless the test scoring includes a "penalty for guessing" (i.e., subtracts points for each incorrect response), providing any response, even one chosen at random, can only increase a student's score. For open-response items, deciding whether to leave items blank may be more complicated, because writing, drawing, or typing a response takes time and effort that could be spent on other items.

Correlates of Nonresponse

Numerous studies have sought insight into nonresponse by investigating differences among groups of students. For example, results from the Second International Mathematics Study (SIMS) revealed large differences between countries, suggesting cultural differences in the acceptability of guessing (Schmidt, Wolfe, & Kifer, 1992). Other studies have compared results by students' race/ethnicity (we use the term race/ethnicity not to suggest that these constructs are interchangeable, but because much research, especially in the United States, categorizes students by a combination of race and ethnicity) or gender. For example, in a study of high school students in the United States taking the ACT Assessment, Zhu and Thompson (1995) found that White students were most likely to respond to all items on the multiple-choice ACT assessment, followed by Asian students and Hispanic students, while African American students had the highest nonresponse rates. Koretz, Lewis, Skewes-Cox, and Burstein (1993) found that Hispanic and African American students participating in the National Assessment of Educational Progress (NAEP) responded to fewer open-response mathematics items than White students. DeMars (1998) found that boys were less likely to respond to open-response items than girls in low-stakes situations. Analysing data from 20 years of the state tests for students in Grades 3, 7, and 11 in Iowa, von Schrader and Ansley (2006) found larger, but still small, differences in nonresponse rates by gender for the younger students, with girls leaving more items blank on mathematics tests and boys leaving more blank on reading and vocabulary tests. More recently, Brown, Dai, and Svetina (2014) studied nonresponse across both multiple-choice and open-response items on the 2009 National Assessment of Educational Progress Math Assessments and found that 30% of students left one or more items blank, but that none of the demographic variables, including gender and race/ethnicity, predicted even 0.5% of the variance in nonresponse.

Some studies of nonresponse have considered its relationship to students' overall performance or motivation. For example, Zhu and Thompson (1995) found that students with lower performance (a few items at the end of each part were excluded from these performance estimates to minimize the effect of speed on the scores) left more items blank on each of the parts of the ACT Assessment, but that the correlations between performance and rate of nonresponse were weak, suggesting that students were not deciding to not respond simply because they found items difficult. In a study of 12th grade students in Australia, Matters and Burnett (2003) compared students who answered all items on the Queensland Core Skills Test with those who left three or more blank and found that the latter group had lower academic self-concept, lower motivation to achieve, and a lower estimate of their own ability, suggesting that noncognitive factors may affect students' decisions.

Deciding Whether to Respond

Although some nonresponse may be attributable to low motivation, studies by Wise and colleagues (e.g., Wise, 2015; Wise & Smith, 2011; Wise, Pastor, & Kong, 2009) show that many students with low motivation do respond, even on low-stakes tests, albeit noneffortfully. When tests are administered by computer, some of these noneffortful responses can be identified because students respond in less time than it would take to read the item.

That even students who do not read the items nevertheless respond suggests that decisions about whether or not to respond are not necessarily made one item at a time. For these students, at least, the decision seems to be "Will I respond to all the items"? rather than a series of item-by-item decisions of "Will I respond to this item"? Could the same be true for at least some students when they are responding effortfully? Certainly, a single decision is more consistent than item-by-item decisions with the weak relationships between performance and nonresponse found by Zhu and Thompson (1995) in their analysis of the high-stakes ACT Assessment.

Does this mean that students ask themselves either a single "Will I respond to all the items"? or a series of "Will I respond to this item"? More likely, the questions are considered in sequence: Only if the students answers "no" or "it depends ..." to "Will I respond to all the items"? does the subsequent question, "Will I respond to this item"? become relevant. In other words, all students answer the first question, but only some answer the second.

The question of whether or not students intend to respond to all the items may not be particularly relevant for students with sufficient time, motivation, and the knowledge and skills required by the test items. For most students, however, assuming that the test is not too easy, an insufficiency of time, motivation, knowledge, or skills will make this question very relevant. For those students, the presence of any nonresponse implies a response of "no" to "Will I respond to all the items"?; the absence of nonresponse implies a response of "yes." That is, for all but the most able, motivated, and quick students, it may be possible to infer students' decisions from the presence or absence of nonresponse.

This Study

To better understand students' decisions about whether or not to respond to all items, we analysed data from a large-scale mathematics assessment administered to Grade 9 students in Ontario. Specifically, we used latent class analysis (LCA) to investigate four decisions by students: their decisions about "Will I respond to all the items"? for multiple-choice items and for open-response items in each of two booklets. Although LCA has been used to analyse students' correct responses to large-scale assessments (e.g., Oliveri, Ercikan, & Zumbo, 2013; Scarpati, Wells, Lewis, & Jirka, 2011), to our knowledge no studies have used LCA to analyse nonresponse. In this analysis, we use LCA to group students by their patterns of nonresponse decisions, with proportion correct scores as covariates, before relating these groups to the students' attitudes toward math.

Methods

Instrument and Participants

The data analysed in this study are students' responses to two English-language versions of Ontario's January 2006 Grade 9 Assessment of Mathematics. Students in Grade 9 took either an Academic or Applied mathematics course; these are roughly equivalent to university preparation and technical streams in other jurisdictions, the latter also being less academically rigorous. The Academic and Applied versions of the test were developed to different specifications based on the curricula for the two courses, but shared the same structure: 24 multiple-choice (MC) items and 12 open-response (OR) items administered in two test booklets of 50 min each. Booklet 1 contained 15 MC items and one task with a set of 4 OR items. Booklet 2 contained 9 MC items and two tasks, each with 4 OR items. MC items were machine scored as correct or incorrect; OR items were scored by trained raters using a 4-level rubric specific to each question. A student questionnaire was also administered. French-language versions of the test were also developed and administered but are not included in these analyses because insufficient numbers of students took those versions.

Both versions of the test were administered in 50 min blocks on two consecutive days during the regularly-scheduled mathematics course time by the teachers who taught those courses. The test administration script included reminders to "show all your work" and "answer the questions completely," but did not instruct teachers to tell students to answer all of the items or to leave no blanks (EQAO, 2005). The test has moderate stakes for the students: Both their parents and the school receive their test scores and many schools mark part of the test and count it toward the mathematics course grade.

Students who answered none of the items correctly or did not receive the questionnaire were excluded from these analyses. Students who received testing accommodations or special provisions were also excluded because some of the accommodations, such as prompting students to return their attention to the assessment or scribing students' responses, may affect students' decisions about whether or not to leave items blank. These exclusions together reduced the number of students for the Applied version from 20,556 to 16,258 and the number for the Academic version from 42,821 to 41,045. From each of these versions, random samples of 10,000 students were selected for the LCA analyses (a sample size of 10,000 was chosen because smaller samples did not include all 16 combinations of the four binary nonresponse indicators).

Four binary indicators of the students' decisions (whether the student left blank any MC items or any OR items in Booklet 1 or Booklet 2) were created. Because some studies have found a relationship between performance and nonresponse, proportions of attempted MC and OR items answered correctly were calculated to serve as covariates in the LCA. We also included four attitude variables. These are based on responses to four background questions: (1) Dislike math (responded *Disagree* or *Strongly Disagree* versus *Agree* or *Strongly Agree* to the statement "I like mathematics"), (2) Believe not good at math (*Disagree* or *Strongly Disagree* versus *Agree* or *Strongly Agree* to "I am good at mathematics"), (3) Believe will not need math for a job (*Disagree* or *Strongly Disagree* versus *Agree* or *Strongly Agree* to "I need to keep taking mathematics for the kind of job I want after I leave school"), and (4) Find math boring (*Agree* or *Strongly Agree* versus *Disagree* or *Strongly Disagree* to "Mathematics is boring"). Because whether students had positive or negative attitudes toward mathematics was of greater interest than the degree of positivity or negativity, we dichotomized the responses to each question. This also allowed us to more clearly see the distributions across classes of students with negative or positive attitudes toward mathematics.

Analyses

LCA involves estimating parameters for two equations. The first equation expresses the probability of an examinee being in a particular latent class and the second equation is the probability of an examinee who is in a class having a value of 1 on a binary latent class indicator. Following Muthén (2004), the two equations can be written as follows. The first relates a latent class to the covariates:

$$P\left(c_{ik} = 1 \middle| x_{i1}, x_{i2}, \dots, x_{iQ}\right) = \frac{e^{\alpha_{c_k} + \gamma_{c_k 1} x_{i1} + \gamma_{c_k 2} x_{i2} + \dots + \gamma_{c_k Q} x_{iQ}}}{\sum_{k=1}^{K} \left(e^{\alpha_{c_k} + \gamma_{c_k 1} x_{i1} + \gamma_{c_k 2} x_{i2} + \dots + \gamma_{c_k Q} x_{iQ}} \right)} ,$$

where

$x_{i1}, x_{i2}, \dots, x_{iQ}$ are examinee i's scores on the q covariates,

α_{c_k} is the logit intercept for the k^{th} class, and

$\gamma_{c_k 1}, \gamma_{c_k 2}, \dots, \gamma_{c_k Q}$ are the logit slopes of covariates $x_{i1}, x_{i2}, \dots, x_{iQ}$ for the k^{th} class.

The logit slopes and intercepts for the last class in a model are always standardized to 0.

The second equation relates each of the binary latent class indicators (the outcome variables in these analyses) to the latent categorical variable and covariates. It is calculated for each class:

$$P\left(u_{ij} = 1 \middle| c_{ik}, x_{i1}, x_{i2}, \dots, x_{iQ}\right) = 1 - \frac{1}{1 + e^{-(\tau - (\kappa_{c_k j1} x_{i1} + \kappa_{c_k j2} x_{i2} + \dots + \kappa_{c_k jQ} x_{iQ}))}} ,$$

where

u_{ij} is the j^{th} binary latent class indicator for examinee i,

τ_{c_k} is, for the k^{th} class, the threshold on the underlying continuous variable $u_{ij}^* = (\kappa_{c_k j1} x_{i1} + \kappa_{c_k j2} x_{i2} + \dots + \kappa_{c_k jQ} x_{iQ})$ above which $u_{ij} = 1$, and

$\kappa_{c_k j1}, \kappa_{c_k j2}, \dots \kappa_{c_k jQ}$ are the logit slopes of covariates $x_{i1}, x_{i2}, \dots, x_{iQ}$ for that class.

If there are no covariates in the model or no direct effects between the covariates and the binary latent class indicators, then $u_{ij}^* = 0$.

In this study, models with 2, 3, and 4 latent classes were fitted to the data in three ways: (a) including only binary indicators of nonresponse (responded to all items versus did not respond to one or more items) by item type (multiple-choice or open-response) and booklet (for the first or second day of testing), (b) adding proportion correct scores by item type as covariates, and (c) adding direct effects between the covariates and the binary indicators. Each model was fitted using 400 random starting values, to ensure that the obtained parameter estimates were stable. Comparisons among models were based on the Bayesian Information Criterion (BIC), which indicates the relative fit of the models to the data; the BIC was chosen as it has been found to be more accurate than other fit indices when the sample size is large (Nylund, Asparouhov, & Muthén, 2007). Entropy, which indicates the certainty with which individuals can be placed in classes (Muthén, 2004), was also considered.

SPSS 23 was used to compute the descriptive statistics. Mplus 7.2 (Muthén & Muthén, 2014) was used to perform the LCA.

Results and Discussion

Table 1 shows the sixteen possible combinations of the four binary indicators and how those patterns relate to students' performance. For the Applied version of the test (taken by students in the Applied mathematics course), students who answered all the items (the first row of numbers in the table) had the highest average reported score (a student's reported score can have values between 0.1 and 4.9 and is based on an item response theory calibration with three-parameter logistic and generalized partial credit models; although the Applied and Academic

Table 1

Performance by Response Patterns

				Applied Math Course				Academic Math Course			
Responded to all MC items in Booklet 1?	Responded to all MC items in Booklet 2?	Responded to all OR items in Booklet 1?	Responded to all OR items in Booklet 2?	Percentage (N = 10,000)	Reported Score M (SD)	MC Proportion Correct of Attempted M (SD)	OR Proportion Correct of Attempted M (SD)	Percentage (N = 10,000)	Reported Score M (SD)	MC Proportion Correct of Attempted M (SD)	OR Proportion Correct of Attempted M (SD)
Yes	Yes	Yes	Yes	46.76%	3.09 (0.73)	.68 (.17)	.60 (.20)	47.99%	3.48 (0.52)	.67 (.17)	.69 (.19)
Yes	Yes	Yes	No	30.85%	2.73 (0.66)	.64 (.16)	.59 (.19)	12.62%	2.95 (0.58)	.57 (.16)	.54 (.18)
Yes	Yes	No	Yes	3.33%	2.01 (0.67)	.54 (.16)	.38 (.19)	12.18%	3.20 (0.47)	.62 (.15)	.64 (.18)
Yes	Yes	No	No	11.77%	1.54 (0.62)	.50 (.15)	.39 (.19)	19.47%	2.53 (0.71)	.54 (.15)	.52 (.18)
Yes	No	Yes	Yes	0.50%	2.94 (0.68)	.66 (.17)	.60 (.20)	0.44%	3.34 (0.45)	.67 (.16)	.64 (.18)
Yes	No	Yes	No	0.82%	2.45 (0.59)	.60 (.14)	.57 (.19)	0.35%	2.85 (0.60)	.58 (.17)	.55 (.18)
Yes	No	No	Yes	0.08%	1.95 (0.65)	.59 (.16)	.30 (.12)	0.25%	3.11 (0.43)	.59 (.15)	.61 (.17)
Yes	No	No	No	0.50%	1.60 (0.81)	.55 (.16)	.41 (.21)	0.72%	2.61 (0.75)	.54 (.16)	.49 (.18)
No	Yes	Yes	Yes	1.25%	2.97 (0.65)	.67 (.15)	.60 (.19)	1.71%	3.37 (0.45)	.67 (.17)	.69 (.17)
No	Yes	Yes	No	1.80%	2.55 (0.68)	.63 (.16)	.56 (.19)	0.70%	2.89 (0.63)	.60 (.16)	.54 (.18)
No	Yes	No	Yes	0.31%	1.82 (0.63)	.53 (.14)	.35 (.18)	0.96%	3.22 (0.44)	.66 (.15)	.65 (.18)
No	Yes	No	No	1.23%	1.54 (0.70)	.52 (.17)	.41 (.20)	1.81%	2.45 (0.74)	.57 (.17)	.53 (.18)
No	No	Yes	Yes	0.11%	2.77 (1.06)	.62 (.25)	.59 (.27)	0.10%	3.16 (0.43)	.61 (.14)	.63 (.22)
No	No	Yes	No	0.36%	2.51 (0.61)	.67 (.17)	.61 (.18)	0.10%	2.74 (0.67)	.59 (.17)	.58 (.10)
No	No	No	Yes	0.04%	1.85 (0.57)	.55 (.12)	.33 (.20)	0.11%	3.23 (0.24)	.65 (.16)	.72 (.13)
No	No	No	No	0.29%	1.46 (0.75)	.59 (.16)	.45 (.21)	0.49%	2.17 (0.78)	.56 (.15)	.53 (.16)
All				100.00%	2.70 (0.87)	.63 (.17)	.56 (.21)	100.00%	3.15 (0.70)	.62 (.17)	.62 (.20)

Note. MC = multiple-choice; OR = open-response. Reported score is out of 4.

versions both report scores with the same range, the versions are based on different test specifications and the scales are not equated), followed by students who answered all the OR items, but not all the MC items. For the Academic version, students who answered all the items also received the highest average score, but after that the most important determinant was whether students had answered all the OR items in Booklet 2 (those who left one or more OR item in Booklet 2 blank performed worse).

Table 1 also shows that students taking the Applied version were a little less likely to answer all the items (46.8% answered all the items) than those taking the Academic version (48.0%). The percentages answering all the MC items were similar across versions (92.7% vs. 92.3%). The largest difference was for those students who answered all the MC items and all the OR items in Booklet 1, but left one or more OR item blank in Booklet 2 (30.9% vs. 12.6%).

Overall, the students taking the Applied version have more negative attitudes toward mathematics than those taking the Academic version: A higher percentage responded that they dislike math (40.1% vs. 24.2%), believe they are not good at math (29.7% vs. 17.4%), believe they will not need more math coursework for their future career (22.5% vs. 11.8%), and find math boring (45.1% vs. 32.4%).

Nine latent class models were fit to the data for the Applied and Academic versions separately. For the first set of three models, those without covariates, only the 2-class and 3-class models were identified. Table 2 provides the fit indices. For both versions, the 3-class model compared to the 2-class model has higher relative entropy, indicating greater classification certainty. In addition, the BIC is smaller for the 3-class model, indicating the best fit within this set of models. This model is shown in Figure 1.

For the second set of models, those with covariates, but without direct effects, the 4-class model has the highest entropy and smallest BIC. This model is shown in Figure 2. For the third set of models, those with direct effects with the covariates, entropy was not as high as for the models without direct effects and BIC levels were similar, suggesting that the effects of students' performance on decisions about nonresponse were more parsimoniously modeled as acting through the latent classes instead of also acting directly on the decisions. Consequently, in the interpretation that follows, we will focus on the 3-class model with binary indicators but no covariates and the 4-class model with binary indicators and covariates, but no direct effects.

The relationships of the patterns of nonresponse decisions to their most probable classes from these models are shown in Table 3. The 3-class model with binary indicators but no covariates has the same class structure for the Applied and Academic versions. As Table 3 shows, Class 1 is the 51.8% of the students taking the Applied version and the 62.3% of those taking the Academic version who answered all of the items or left one or more items blank in only one of the following: MC items in Booklet 1, MC items in Booklet 2, or OR items in Booklet 1. For simplicity, we will refer to this class as *responders*. Class 2 is characterized by leaving OR items blank in Booklet 2, perhaps because of fatigue (recall that there were 8 OR items in Booklet 2 and only 4 OR items in Booklet 1): 43.4% of students taking the Applied version and 32.4% of Academic are in this class. We will refer to this class as *Booklet 2 OR nonresponders*. Class 3, the smallest class (4.7% taking the Applied version and 5.2% taking the Academic version), is students we might characterize as *habitual nonresponders*: these students left MC items blank in both booklets, or left MC items blank in one booklet and OR items blank in one or both booklets.

For students taking the Academic version, the 4-class model with both binary indicators and covariates but without direct effects yielded similar class assignments to the 3-class model

Table 2

LCA Fit Indices

Model	Number of Classes	Loglikelihood	Number of parameters	BIC	Entropy
Applied Math Course					
Binary indicators, no covariates	2	-14331.22	9	28745.32	0.481
	3	-14284.44 [a]	14	28697.83	0.651
	4	Model not identified			
Binary indicators and covariates, no direct effects	2	-13595.58	11	27292.47	0.766
	3	-13479.36	18	27124.50	0.719
	4	-13408.07	25	27046.40	0.789
Binary indicators and covariates, with direct effects	2	-13476.10	15	27090.35	0.497
	3	-13401.22	22	27005.06	0.557
	4	-13351.68	29	26970.45	0.603
Academic Math Course					
Binary indicators, no covariates	2	-15574.95	9	31232.79	0.570
	3	-15519.15	14	31167.25	0.718
	4	Model not identified			
Binary indicators and covariates, no direct effects	2	-14828.84	11	29758.99	0.690
	3	-14722.47	18	29610.72	0.646
	4	-14637.01 [a]	25	29504.28	0.716
Binary indicators and covariates, with direct effects	2	-14769.62	15	29677.40	0.535
	3	-14673.21	22	29549.04	0.582
	4	-14613.85	29	29494.81	0.629

Note. [a] extreme values of some logit thresholds of the binary indicators on the covariates had to be set at 15 or -15 ; [b] degrees of freedom cannot be computed (this also precludes computation of a *p* value)

without covariates, described above. Only three classes appear in Table 3 because Class 4 was not most probable for any of the patterns; however, when the proportions correct which were the covariates in the model were also considered in assigning individual students to classes, some of those in Class 1 who in addition to answering all items in at least three of the four item-booklet combinations also had very high proportion correct scores were assigned to Class 4, so we might label Class 4, *high performing responders*. Class assignments differed for three of the more unusual patterns; each of these had less than 1% of the students. The interpretation of the classes remains the same.

For students taking the Applied version, however, the analogous 4-class model had somewhat different class assignments. Essentially the roles of nonresponse to OR items in Booklets 1 and 2 have been reversed, with Class 1 still *responders* and Class 3 still *habitual nonresponders*, but Class 2 now *Booklet 1 OR nonresponders*, instead of *Booklet 2 OR nonresponders*. Class 4 does not appear in Table 3 because it was not most probable for any of

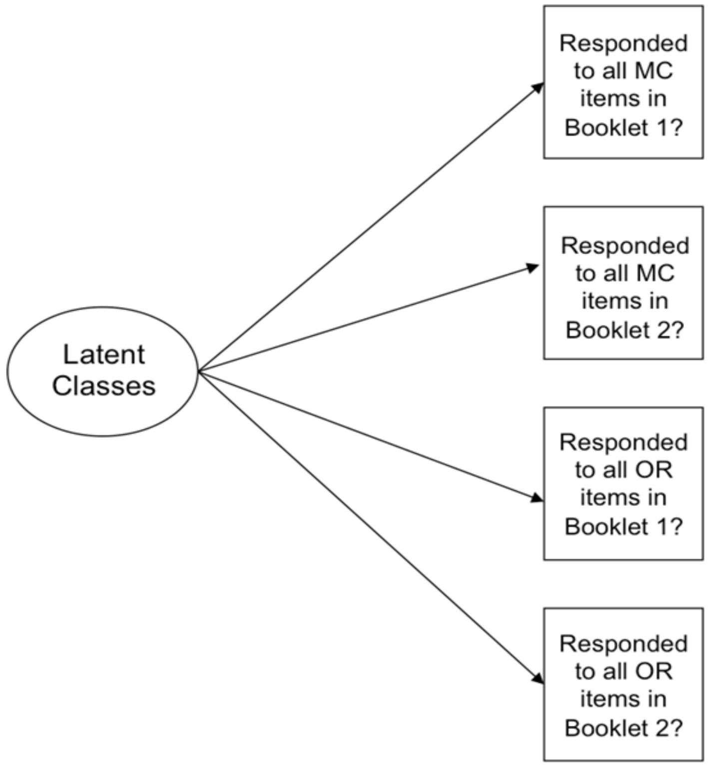


Figure 1. Model with Four Binary Latent Class Indicators and No Covariates.

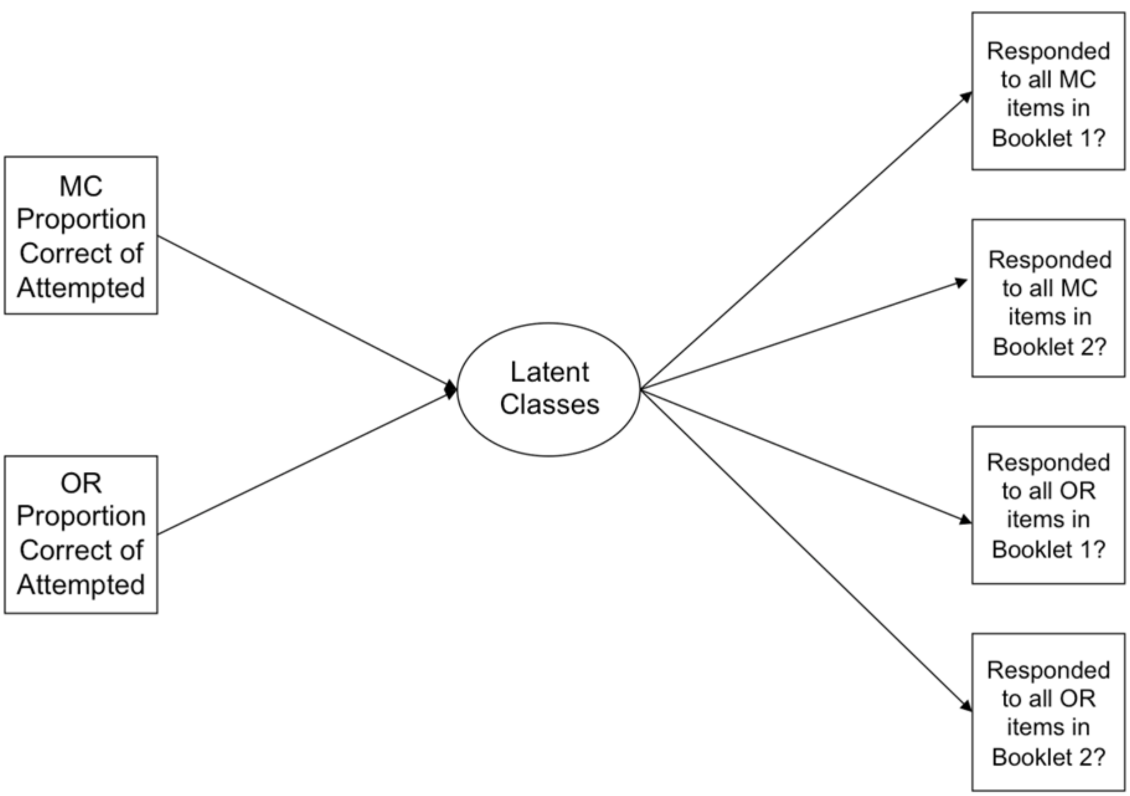


Figure 2. Model with Four Binary Latent Class Indicators and Two Covariates, No Direct Effects.

Table 3

Most Probable Class for Each Response Pattern

						Binary indicators, no covariates 3-class model			Binary indicators and covariates, no direct effects 4-class model [a]			
Responded to all MC items in Booklet 1?	Responded to all MC items in Booklet 2?	Responded to all OR items in Booklet 1?	Responded to all OR items in Booklet 2?	Percentage of Applied Math Course	Percentage of Academic Math Course	Applied and Academic: Most Probable Class	Applied: Probability of Most Probable Class	Academic: Probability of Most Probable Class	Applied: Most Probable Class	Applied: Probability of Most Probable Class	Academic: Most Probable Class	Academic: Probability of Most Probable Class
Yes	Yes	Yes	Yes	46.76%	47.99%	1 (Responders)	.99	.98	1 (Responders)	1.00	1 (Responders)	.68
Yes	Yes	Yes	No	30.85%	12.62%	2 (Booklet 2 OR Nonresponders)	.52	.72	1 (Responders)	.96	2 (Booklet 1 OR Nonresponders)	.93
Yes	Yes	No	Yes	3.33%	12.18%	1 (Responders)	.93	.85	2 (Booklet 2 OR Nonresponders)	.99	1 (Responders)	.96
Yes	Yes	No	No	11.77%	19.47%	2 (Booklet 2 OR Nonresponders)	.83	.89	2 (Booklet 2 OR Nonresponders)	.83	2 (Booklet 1 OR Nonresponders)	1.00
Yes	No	Yes	Yes	0.50%	0.44%	1 (Responders)	.87	.88	1 (Responders)	1.00	1 (Responders)	.89
Yes	No	Yes	No	0.82%	0.35%	2 (Booklet 2 OR Nonresponders)	.46	.56	3 (Habitual Nonresponders)	.67	2 (Booklet 1 OR Nonresponders)	1.00
Yes	No	No	Yes	0.08%	0.25%	3 (Habitual Nonresponders)	.64	.53	2 (Booklet 2 OR Nonresponders)	1.00	1 (Responders)	1.00
Yes	No	No	No	0.50%	0.72%	3 (Habitual Nonresponders)	.51	.51	2 (Booklet 2 OR Nonresponders)	.72	2 (Booklet 1 OR Nonresponders)	1.00
No	Yes	Yes	Yes	1.25%	1.71%	1 (Responders)	.74	.89	1 (Responders)	1.00	1 (Responders)	.68
No	Yes	Yes	No	1.80%	0.70%	3 (Habitual Nonresponders)	.85	.85	3 (Habitual Nonresponders)	.69	2 (Booklet 1 OR Nonresponders)	.60
No	Yes	No	Yes	0.31%	0.96%	3 (Habitual Nonresponders)	.80	.64	2 (Booklet 2 OR Nonresponders)	1.00	3 (Habitual Nonresponders)	.52
No	Yes	No	No	1.23%	1.81%	3 (Habitual Nonresponders)	.99	.98	2 (Booklet 2 OR Nonresponders)	.79	3 (Habitual Nonresponders)	.73
No	No	Yes	Yes	0.11%	0.10%	3 (Habitual Nonresponders)	.90	.79	3 (Habitual Nonresponders)	1.00	3 (Habitual Nonresponders)	.90
No	No	Yes	No	0.36%	0.10%	3 (Habitual Nonresponders)	.99	.99	3 (Habitual Nonresponders)	1.00	3 (Habitual Nonresponders)	1.00
No	No	No	Yes	0.04%	0.11%	3 (Habitual Nonresponders)	.99	.98	2 (Booklet 2 OR Nonresponders) or 3 (Habitual Nonresponders)	.50/.50	3 (Habitual Nonresponders)	1.00
No	No	No	No	0.29%	0.49%	3 (Habitual Nonresponders)	1.00	1.00	3 (Habitual Nonresponders)	.97	3 (Habitual Nonresponders)	1.00

Note. [a] Class 4 was not the most probable for any pattern.

the patterns, but when the covariates are considered, students who had the following combination of pattern and covariates were assigned to Class 4: left some OR items blank, had a low proportion correct score for the MC items, and performed better on the OR items than on the MC items (we will refer to this class as *better OR than MC*).

To illustrate, let us consider the specifics of the 4-class model for the Applied version. Recall that the model involves two equations. Applying the first equation, for this 4-class model with covariates but no direct effects, the probability that examinee (student) i is in Class 1 is (using the estimates computed by Mplus):

$$P\text{ (examinee } i \text{ is a member of Class 1} \mid \text{MCPROP}_i, \text{ORPROP}_i) =$$

$$\frac{e^{(-3.036\,+\,10.844(\text{MCPROP}_i)-\,6.468(\text{ORPROP}_i))}}{e^{(-3.036\,+\,10.844(\text{MCPROP}_i)-\,6.468(\text{ORPROP}_i))}+e^{(-1.222\,+\,12.134(\text{MCPROP}_i)-\,5.659(\text{ORPROP}_i))}+e^{(0.928+11.082(\text{MCPROP}_i)-\,12.290(\text{ORPROP}_i))}+e^{(0\,+\,0(\text{MCPROP}_i)+\,0(\text{ORPROP}_i))}}$$

This is shown graphically in Figure 3, which also shows the relationships for Classes 2, 3, and 4. As these plots show, the proportion correct scores are related to class probability for three of the four classes. Class 1, which we labeled *responders*, is most probable for students with high proportion correct scores on both MC and OR items; that is, the strongest students are in this class. Class 2, which we labeled *Booklet 1 OR nonresponders*, is more probable for students who perform poorly on the OR items, regardless of their performance on the MC items. Class 3, the *habitual nonresponders*, does not include very many students and so is not very probable for any combination of covariate levels. Class 4 is most probable for students who performed poorly on the MC items, but well on the OR items. When examining these graphs, it is important to remember that the distribution of students across possible combinations of the covariates is not uniform, as Figure 4 shows. In other words, the probability of Class 4 may be highest for students who performed well on the OR items and poorly on the MC items, but there are very few students with this pattern of performance.

Using the second equation and the coefficients estimated by Mplus, the probability that examinee (student) i has left blank any MC items in Booklet 1 (that is, has a value of 1 on the first binary indicator), if that examinee is a member of Class 1 is:

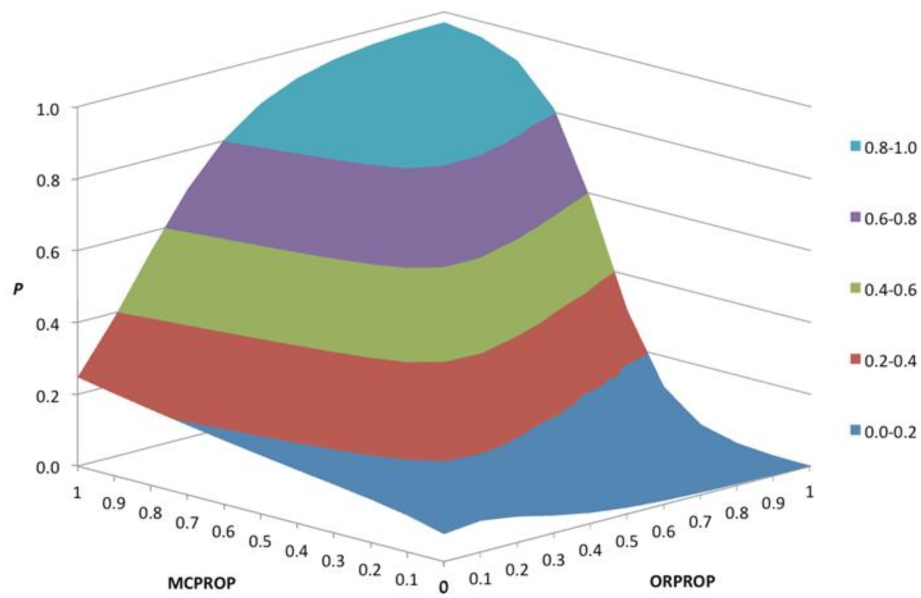
P (examinee i has not answered any MC items in Booklet 1 | member of Class 1, MCPROP_i, ORPROP_i) =

$$1-\frac{1}{1+e^{-(3.698-(0(\text{MCPROP}_i)+0(\text{ORPROP}_i)))}}=.024$$

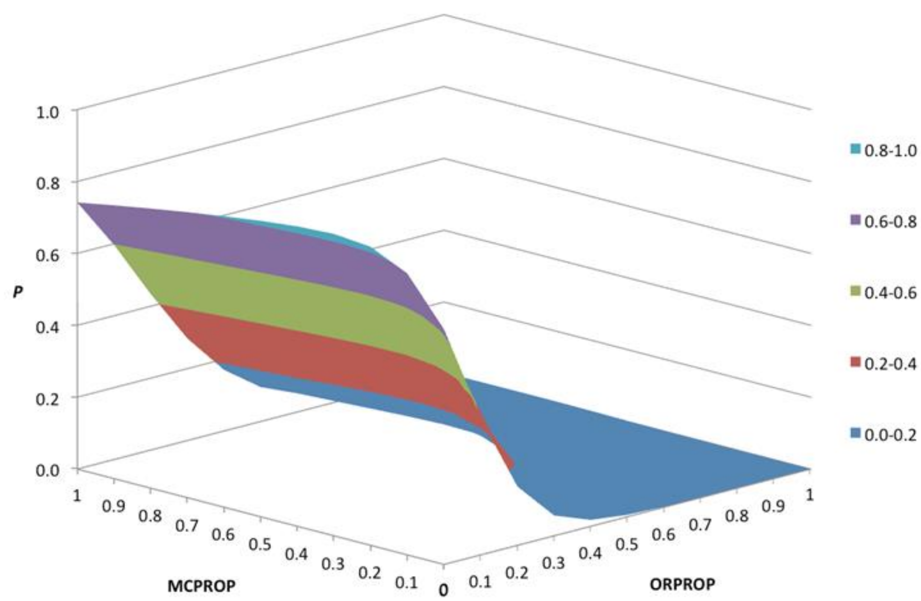
This is shown graphically in Figure 5, which also shows the relationships for Classes 2, 3, and 4 and the other binary latent class indicators. Not surprisingly, the probability of any nonresponse is lowest for Class 1, the responders. The probability of not responding to MC items in either book is highest for Class 3, the habitual nonresponders.

Table 4 summarizes the attitude question responses for the students who are most likely to be assigned to a class based on their values on the binary indicators. The most striking difference in this table is for Classes 2 and 4 for the Academic version. Recall that Class 4 is the subset of responders who also performed very well on both the MC and OR items. As Table 4 shows, those who agreed with the negative statement (that is, they reported that they dislike math) are very unlikely to be in Class 4, compared to those who disagreed. For the Academic version, Class 2 is Booklet 2 OR nonresponders; those with negative attitudes are more likely to

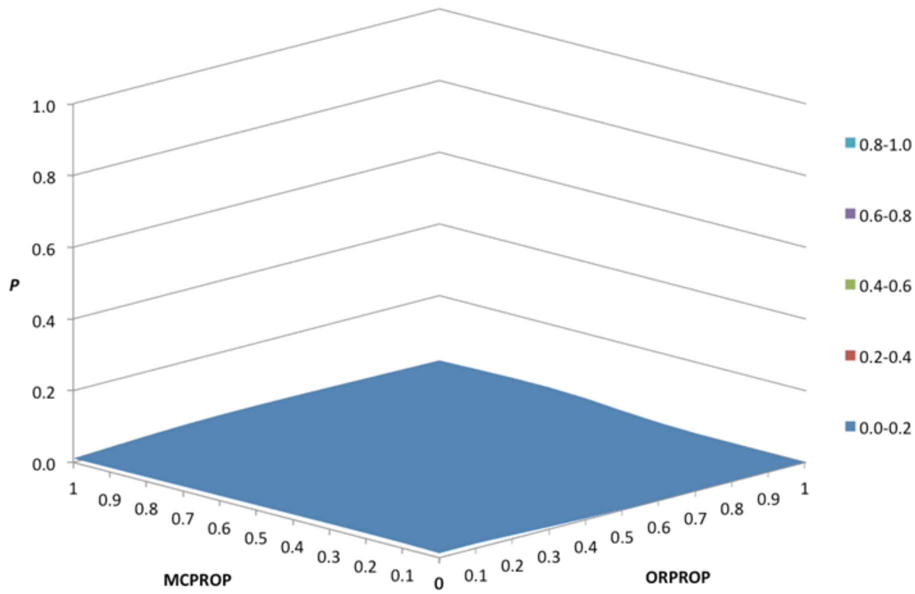
Class 1 (Responders)



Class 2 (Booklet 1 OR Nonresponders)



Class 3 (Habitual Nonresponders)



Class 4 (Better OR than MC)

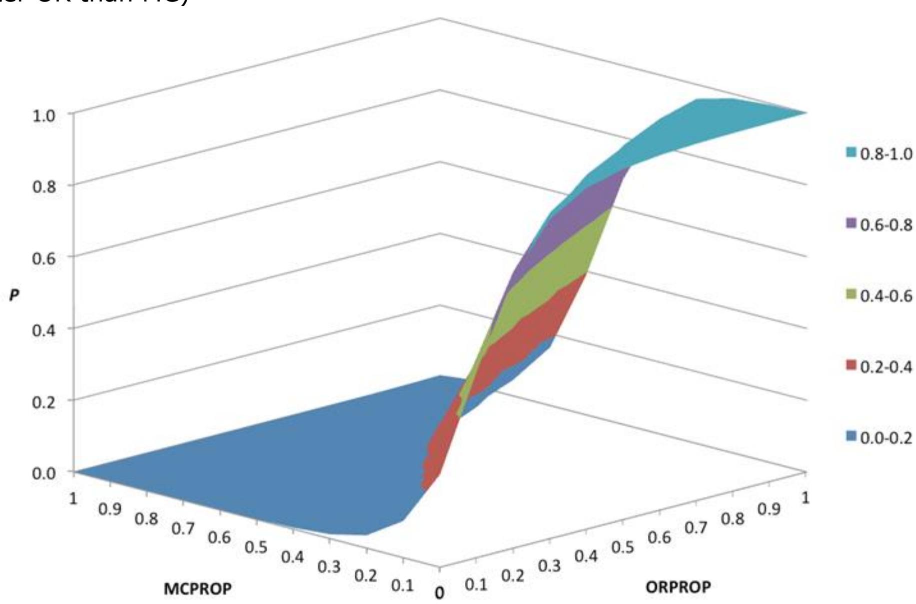


Figure 3 (continued). Class probabilities in relation to the covariates (MC proportion correct of attempted and OR proportion correct of attempted) for the 4-class model with binary indicators and covariates, but no direct effects: Applied version.

be in Class 2 than those with positive attitudes. For the Applied version, those with negative attitudes are less likely than those with positive attitudes to be in Class 1, responders, and more likely to be in Class 2, Booklet 1 OR nonresponders. These results are not surprising, as the OR items require more effort. Table 4 also reports the percentage of students in each class.

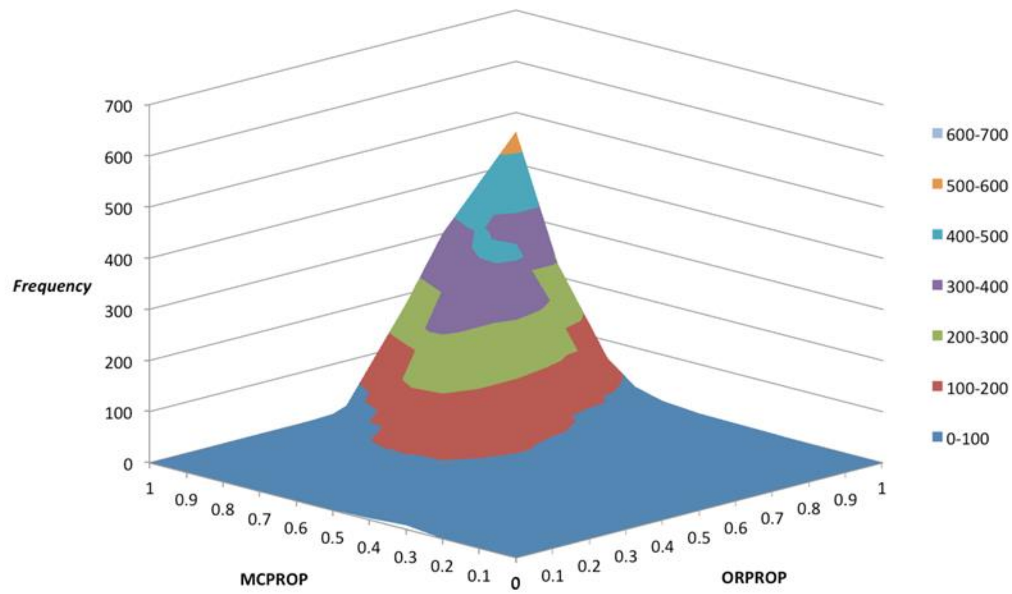


Figure 4. Frequencies across the Covariates (MC proportion correct of attempted and OR proportion correct of attempted)

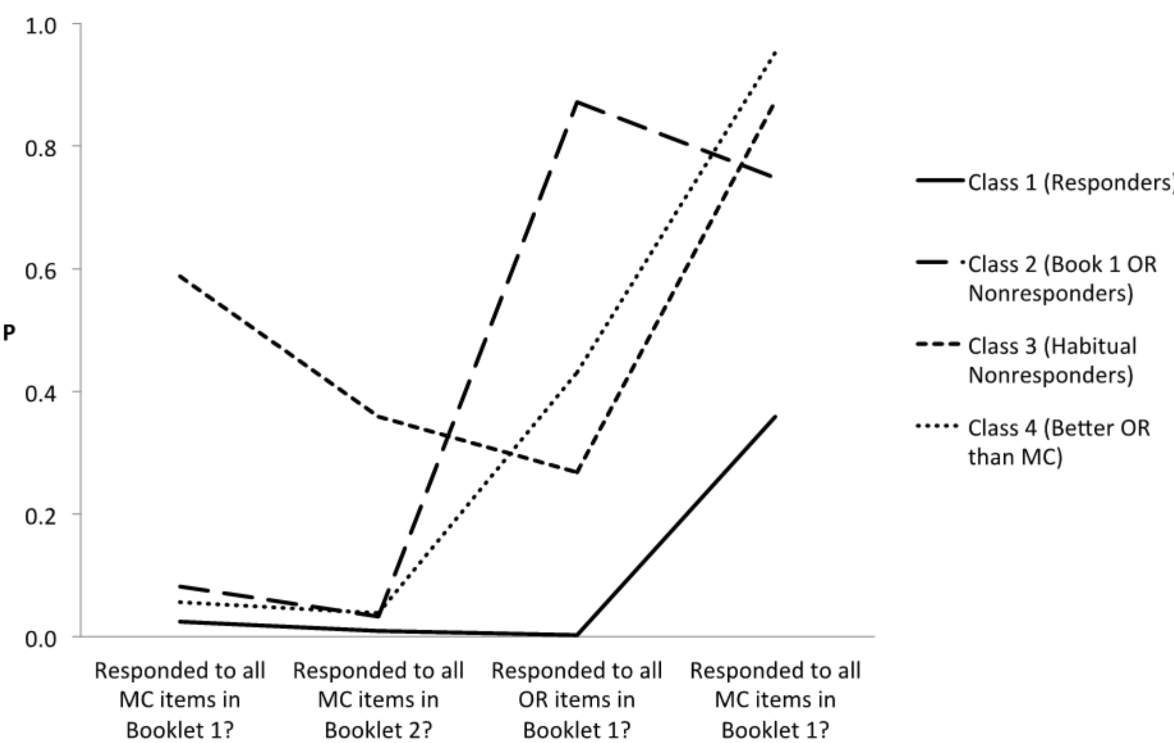


Figure 5. Item profiles for 4-class model with binary indicators and covariates, but no direct effects: Applied version.

Table 4

Latent Class Assignments in Relation to Attitudes toward Math (4-class Model with Binary Indicators and Covariates, but No Direct Effects)

		Applied Math Course					Academic Math Course				
Attitude		Percentage of Students in Applied Math Course	Class 1 (Responders)	Class 2 (Booklet 2 OR Nonresponders)	Class 3 (Habitual Nonresponders)	Class 4 (High Performing Responders)	Percentage of Students in Academic Math Course	Class 1 (Responders)	Class 2 (Booklet 1 OR Nonresponders)	Class 3 (Habitual Nonresponders)	Class 4 (Better OR than MC)
Dislike math	No	59.7%	82.5%	11.9%	2.6%	3.0%	75.7%	47.6%	29.5%	2.8%	20.1%
	Yes	40.0%	72.8%	19.4%	3.0%	4.8%	24.1%	45.5%	44.9%	3.2%	6.5%
Believe not good at math	No	69.9%	83.4%	11.6%	2.4%	2.6%	82.3%	48.5%	28.8%	2.7%	20.0%
	Yes	29.6%	67.3%	22.8%	3.5%	6.4%	17.3%	40.3%	54.1%	3.6%	2.0%
Believe will not need math for a job	No	77.1%	80.2%	13.9%	2.6%	3.3%	87.9%	47.1%	31.8%	3.0%	18.2%
	Yes	22.3%	73.3%	18.3%	3.2%	5.2%	11.7%	46.9%	43.6%	2.1%	7.3%
Find math boring	No	54.6%	81.2%	12.9%	2.8%	3.1%	67.4%	47.1%	30.1%	2.8%	19.9%
	Yes	44.9%	75.6%	17.3%	2.7%	4.4%	32.3%	46.9%	39.6%	2.9%	10.6%
All		100.0%	78.6%	14.9%	2.8%	3.7%	100.0%	47.1%	33.2%	2.9%	16.8%

Discussion and Conclusion

Although the 3-class model with only the binary indicators of nonresponse is much simpler, comparing it with the 4-class models that include proportion correct of attempted as covariates demonstrates the additional insights available in the more complicated models. In the 3-class model, the latent class structures are the same for the students taking the Applied mathematics course in Grade 9 (and so the Applied version of the test) and those taking the Academic course, even though the former had more negative attitudes toward math and were less likely to respond to all items, especially OR items. When proportion correct is included in the model and a fourth class is added, students whose only nonresponse was in the OR items in Booklet 2 were included in Class 1 for the Applied version, but in Class 2 for the Academic version. This seems to suggest that leaving one or more OR items blank in Booklet 2, which is the end of the test, is more similar to full response for students taking the Applied version than for those taking the Academic version. One possible explanation is that these students—and there were many more of them taking the Applied than the Academic version—did not decide to leave items blank, but ran out of time.

Negative attitudes about mathematics were related to nonresponse to OR items (as seen by the greater likelihood of being in Class 2 for either version of the test), less than nonresponse to MC items, leading us to hypothesize that deciding not to answer all OR items may be due to poor self-efficacy and/or a lack of motivation.

As we argued in the introduction, findings from studies of non-effortful responding and the weak relationships observed between performance and the number of items left blank suggest that focusing on how students answered the question, "Will I respond to all the items?" inferred from their response patterns, may provide insights into nonresponse. The use of LCA to explore which patterns of decisions, across two item types and two booklets, are most similar suggest that not responding to all OR items is indeed different from not responding to all MC items and that not responding to all the OR items at the end of a test may be less a decision than a matter of running out of time, especially for students taking the Applied mathematics course. Determining whether students are indeed making such overall decisions or deciding item-by-item whether to respond would require asking students to describe their decisions as they progress through a test, which is beyond the scope of this study. However, the results of this study are consistent with overall decisions.

Should educators encourage students to respond to all items? In other words, is answering "yes" to "Will I respond to all the items"? better than answering "no" or even "it depends...." In an early study of nonresponse, Sherriffs and Boomer (1954) cited an educator who "says he doubts that school people will ever be willing to encourage pupils to guess at every item … feel[ing] this to be condoning carelessness and loose thinking" (p. 89). That educator would likely be surprised that so many students seem to approach taking a test as though they have already committed to responding to all the questions. As researchers and educators, we find ourselves torn between the knowledge that responding to every item may increase a student's score and uncertainty that maximizing scores by guessing is a value schools should promote. Our hope is that studies such as this one will encourage more discussion of the meaning of nonresponse and whether it should be viewed as something to be discouraged.

Finally, these analyses illustrate LCA's potential as an approach for investigating nonresponse. Approaches such as LCA that can suggest, based on analyses of large datasets, what similarities and differences in patterns of nonresponse decisions matter, may help us infer how these decisions are made. Such analyses have the potential to go beyond who does not respond, as identified in analyses that rely on comparisons of nonresponse among groups based on demographic variables, to suggest why test takers might not respond, based on investigations of latent classes.

Acknowledgements

Thank you to Ontario's Education Quality and Accountability Office for providing the data used in this research. The opinions presented in this paper are solely those of the authors and do not necessarily reflect the opinions of EQAO. Funding for this study was provided by the Social Sciences and Humanities Research Council of Canada.

References

Brown, N. J. S., Dai, S., & Svetina, D. (2014, April). *Predictors of omitted responses on the 2009 NAEP Mathematics Assessment*. Poster presented at the annual meeting of the American Educational Research Association, Philadelphia. Retrieved from http://ceep.indiana.edu/ImplicationsFromNAEP/Brown%20et%20al%202014%20AERA%20Predictors%20of%20omitted%20responses.pdf

DeMars, C. E. (1998). Gender differences in mathematics and science on a high school proficiency exam: The role of response format. *Applied Measurement in Education, 11*, 279-299. doi:10.1207/s15324818ame1103_4

Education Quality and Accountability Office. (2005). *Guide for administering the Grade 9 Assessment of Mathematics, winter 2006* (Document 258007). Retrieved from http://govdocs.ourontario.ca/node/23466

Koretz, D., Lewis, E., Skewes-Cox, T., & Burstein, L. (1993). *Omitted and not-reached items in mathematics in the 1990 National Assessment of Educational Progress* (CRE Technical Report 347). Retrieved from Center for Research on Evaluation, Standards, and Student Testing (CRESST) website: http://cresst.org/publications/cresst-publication-2729

Matters, G. & Burnett, P. C. (2003). Psychological predictors of the propensity to omit short-response items on a high-stakes achievement test. *Educational and Psychological Measurement, 63*, 239-256. doi:10.1177/0013164402250988

Muthén, B. O. (2004). *Mplus statistical analysis with latent variables: Technical appendices*. Retrieved from *Mplus* website: http://www.statmodel.com/download/techappen.pdf

Muthén, L. K., & Muthén, B. O. (2007). *Mplus: Statistical analysis with latent variables* [computer software]. Los Angeles, CA: Authors.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, 14, 535-569. doi:10.1080/10705510701575396

Oliveri, M. E., Ercikan, K., & Zumbo, B. (2013). Analysis of sources of latent class differential item functioning in international assessments. *International Journal of Testing, 13*, 272-293. doi:10.1080/15305058.2012.738266

Scarpati, S. E., Wells, C. S., Lewis, C., & Jirka, S. (2011). Accommodations and item-level analyses using mixture differential item functioning models. *Journal of Special Education, 45*, 54-62. doi:10.1177/0022466909350224

Schmidt, W. H., Wolfe, R. G., & Kifer, E. (1992). The identification and description of student growth in mathematics achievement. In L. Burstein (Ed.), *The IEA Study of Mathematics III: Student growth and classroom processes* (pp. 59-99). Oxford, England: Pergamon.

Sherriffs, A. C., & Boomer, D. S. (1954). Who is penalized by the penalty for guessing? *Journal of Educational Psychology, 45*, 81-90.
von Schrader, S., & Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980-2000. *Applied Measurement in Education, 19*, 41-65. doi:10.1207/s15324818ame1901_3
Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 2*8, 237-252. doi:10.1080/08957347.2015.1042155
Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education, 22*, 185-205. doi:10.1080/08957340902754650
Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, and C. W. Buckendal (Eds.), *High-stakes testing in education: Science and practice in K-12 settings* (pp. 139-153). Washington, DC: American Psychological Association.
Zhu, D., & Thompson, T. D. (1995, April). *Gender and ethnic differences in tendencies to omit responses on multiple-choice tests using number-right scoring*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. Retrieved from https://files.eric.ed.gov/fulltext/ED382689.pdf

Ruth A. Childs is a Professor and the Ontario Research Chair in Postsecondary Policy and Measurement in the Department of Leadership, Higher and Adult Education at the Ontario Institute for Studies in Education, University of Toronto, Toronto, ON, Canada. Her research investigates the design and equity of large-scale assessments, admissions processes, and other evaluation systems. Correspondence should be addressed to: Ruth Childs, Ph.D., Ontario Institute for Studies in Education, University of Toronto, 252 Bloor Street West, Toronto, Ontario, Canada M5S 1V6. E-mail: ruth.childs@utoronto.ca

Orlena Broomes is a Senior Business and Information Analyst at the Ontario Ministry of Education. Her research interests include melding sociology and psychometric perspectives to examine the validity and reliability of standardized, large-scale testing and aligning research, policy and practice particularly related to equity in education.

Monique B. Herbert is an Assistant Lecturer in the Department of Psychology at York University, where she teaches research methods and statistics. Her research interests are in the area of educational measurement and outcomes, with a focus on instrument design, program evaluation, missing data patterns and the development of quantitative reasoning skills at the postsecondary level.