Title: Modeling for Directly Setting Theory-Based Performance Levels

Authors: David Torres Irribarra, Ronli Diakow, Rebecca Freund, and Mark Wilson

Publication Date: 2015

This research was partially supported by Institute of Science grant R305B120017.

Psychological Test and Assessment Modeling, Volume 57, 2015 (3), 396-422

Modeling for Directly Setting Theory-Based Performance Levels

David Torres Irribarra[1], *Ronli Diakow* [2], *Rebecca Freund* [3] & *Mark Wilson*[4]

Abstract

This paper presents the Latent Class Level-PCM as a method for identifying and interpreting latent classes of respondents according to empirically estimated performance levels. The model, which combines elements from latent class models and reparameterized partial credit models for polytomous data, can simultaneously (a) identify empirical boundaries between performance levels and (b) estimate an empirical location of the centroid of each level. This provides more detailed information for establishing performance levels and interpreting student performance in the context of these levels. The paper demonstrates the use of the Latent Class L-PCM on an assessment of student reading proficiency for which there are strong ties between the hypothesized theoretical levels and the polytomously scored assessment data. Graphical methods for evaluating the estimated levels are illustrated.

Keywords: Construct Modeling, Performance Levels, Ordered Latent Class Analysis, Standard Setting, Level Partial Credit Model

[1] *Correspondence concerning this article should be addressed to:* David Torres Irribarra, PhD, Centro de Medición MIDE UC, Pontificia Universidad Católica de Chile, Santiago, Chile; email: davidtorres@uc.cl

[2] New York City Department of Education

[3] University of California, Berkeley

[4] University of California, Berkeley

Introduction

Meaningful interpretation of assessment results is a critical step in a successful educational assessment effort. Without it, the results cannot provide diagnostic information or guide actions for improving student learning. As emphasized by the National Research Council, practitioners of educational assessment should strive to produce meaningful results by designing assessments that coordinate three elements: a cognitive theory of student learning, observations of student performance, and an interpretation of the evidence collected through those observations (Glaser, Chudowsky, & Pellegrino, 2001). This paper focuses on connecting these elements by explicitly examining the relation between the substantive theory of learning used to design an assessment and the mathematical models used to analyze the data collected through that assessment in the context of setting and evaluating performance levels.

This paper expands the work of Diakow, Torres Irribarra, and Wilson (2013), which examined how to trace the interpretation of model-based levels to the substantive theory in the case where (a) the theory specifies multiple ordered levels, (b) the assessment consists of polytomous items that are meant to capture the aforementioned ordered performance levels, and (c) the responses are modeled using a continuous rather than ordinal model. In their work, Diakow et al. (2013) relied on a reparameterization of the Partial Credit Model (PCM; Masters, 1982) to analyze an assessment from the *Striving Readers* curriculum, which was developed according to a learning theory of ordered performance levels. Their approach focused on how the reparameterized model could be used to obtain interpretable level boundaries.

However, under that formulation there is no parameter that explicitly models the location of performance classes on the latent continuum. In this paper, we address this issue through a variation of the original model using latent class analysis. A latent class-based model will simultaneously (a) identify empirical boundaries between performance levels and (b) estimate an empirical location of the locations of each level. This will provide more detailed information for establishing performance levels and interpreting student performance in the context of these levels.

We begin by further introducing the context for this work, standard setting and a motivating empirical example of reading performance. Then, in “The Level Partial Credit Models,” we present the item response models to be used for empirically setting and examining performance levels. “Analysis and Results” contains analysis and results from applying these models to estimate levels on empirical data. We conclude with a discussion of the utility of these methods for setting performance levels in “Discussion.”

Setting performance standards

Assessments are often motivated by substantive theories that describe student performance in terms of a series of ordered levels. One commonly used progression starts at *below basic* and proceeds through *basic* and *proficient* to *advanced.* This and other similar ordered sets of performance categories are common in educational assessments (Perie, 2008), being used for example in tests such as PISA (Programme for International Student Assessment, 2007), TIMSS (Gonzales et al., 2008), and NAEP (Bourque, 2009). In the case of the *Striving Readers* curriculum, the empirical example used in this paper, the levels are labeled *Disengaging, Engaging, Discriminating, Cross-checking*, and *Synthesizing*.

The use of performance levels in the underlying learning theory or in reporting results raises questions about how to conceptualize these levels. Based on the fact that a set of performance levels have been specified, we know that we expect to find a different class of students associated to each performance level, each with a qualitatively different description; however, this still leaves unresolved the issue of how to model these classes. Figure 1 illustrates three ways in which a set of performance levels could be modeled.

We could conceive of the classes as a mixture of continuous distributions (Lubke & Muthén, 2005), as in pane (a) of Figure 1 or we can think of the classes as occupying a single location along the continuum (Formann, 1995), as illustrated in pane (b). Alternatively, we could even drop the assumption of an underlying continuous variable, and simply estimate an ordered set of latent classes (Croon, 1990) as illustrated in pane (c).

However, these models are rarely used by practitioners, who tend to rely more commonly on traditional item response models (such as the Rasch or 2PL models), which usually do not explicitly incorporate the performance levels. These models assume an underlying continuous variable, with proficiency estimates reported along a single continuous scale, as illustrated by the sideways histogram over the latent trait in pane (a) of Figure 2. These models do not incorporate a way to segment the continuous latent variable, hence, it is necessary to conduct an additional procedure to establish a mapping between the continuous results of the mathematical model and the theory-based ordered performance levels.

Practitioners seek to establish a series of cutpoints in order to discretize the results from the item response model, as illustrated in pane (b) of Figure 2. To do so, it is common to rely on a standard-setting procedure (see Cizek, 2001; Cizek, Bunch, & Koons, 2004), which, generally speaking, relies on the input of experts to determine the appropriate location of each of the necessary cutpoints. Multiple standard-setting methods have been proposed, including the Bookmark Method (Lewis, Mitzel, & Green, 1996), the

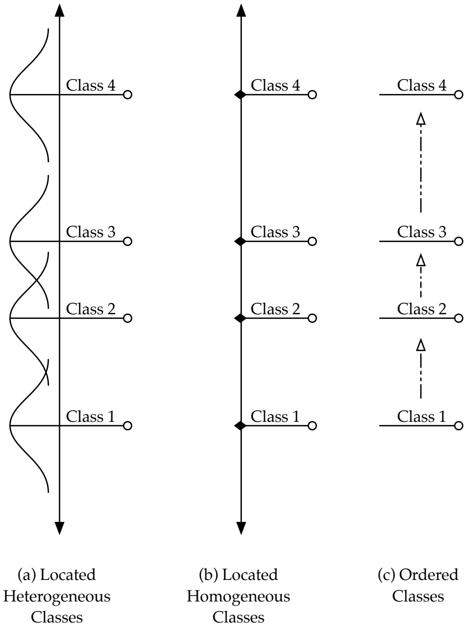


Figure 1: Three alternative ways of conceptualizing levels of performance.

Angoff (Angoff, 1971) and Modified Angoff Methods, and Holistic Methods (Cizek et al., 2004).

In addition to these traditional procedures, the Construct Mapping method (Wilson & Draney, 2002), a blend of holistic methods with the item-mapping elements of the Bookmark method, has been proposed to specifically address the case in which there are well-defined constructs that characterize qualitatively distinct levels of performance (see Wilmot, Schoenfeld, Wilson, Champney, & Zahner, 2011, for an applied example of this method).

Some methods of standard-setting are designed specifically for use with instruments that contain polytomous tasks. Hambleton, Jaeger, Plake, and Mills (2000) give an overview of standard-setting methods for complex performance tasks. The methods they discuss include some in which polytomous tasks are used to place respondents on either side of a single cutpoint (e.g., Hambleton & Plake, 1995), and others in which multiple cutpoints are established (e.g., Reckase, 2000).

However, these methods do not always connect the *scoring procedures* for a polytomous

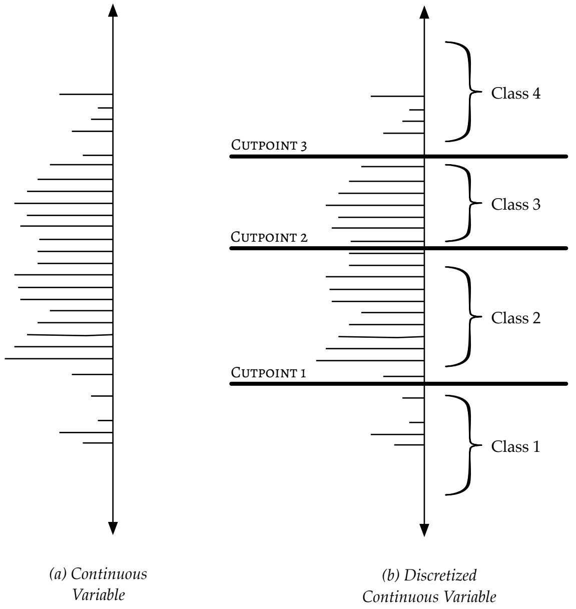
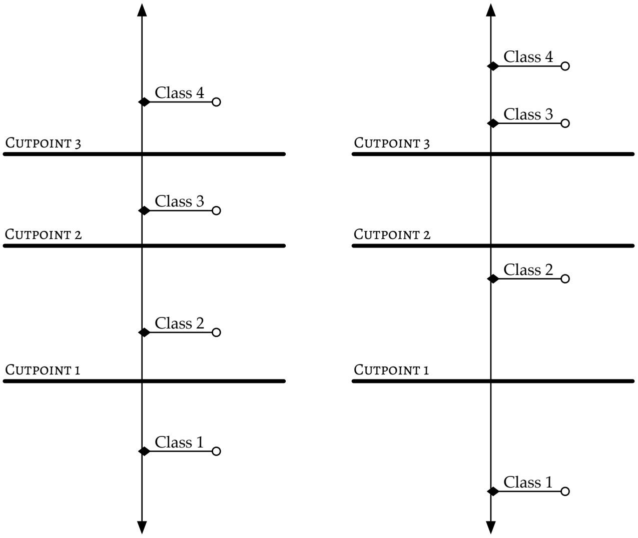


Figure 2: Creating Classes From a Single Distribution.

item to the *standard setting* procedures. In cases such as the *Striving Readers* example, in which the item scoring is motivated by a strong substantive theory describing performance at the various levels, one possibility is to forgo typical standard-setting procedures in favor of an analytic approach that assumes that a respondent at the border of levels 1 and 2 is one that is equally likely to respond at level 1 as level 2 on a typical item. This approach allows for a direct estimation of cutpoints based on a reparameterization of the Partial Credit Model, and is discussed at length in Diakow et al. (2013).

In this paper we extend the work of Diakow et al. (2013) by explicitly modeling latent performance classes through the use of Latent Class Analysis (LCA; Hagenaars & McCutcheon, 2002; Lazarsfeld & Henry, 1968). The combination of these two approaches yields a model that explicitly estimates the location of both the latent classes and the cutpoints. With this analysis, we can empirically examine whether the latent classes align with the empirically estimated levels, as shown in pane (a) of Figure 3, or if the classes are concentrated in one or more of the levels as in pane (b) of Figure 3. The simultaneous estimation of the estimated cutpoints and class locations allows us then to interpret the class locations in relation to the theory-based levels defined in the construct.



(a) Latent Classes Aligned with Levels

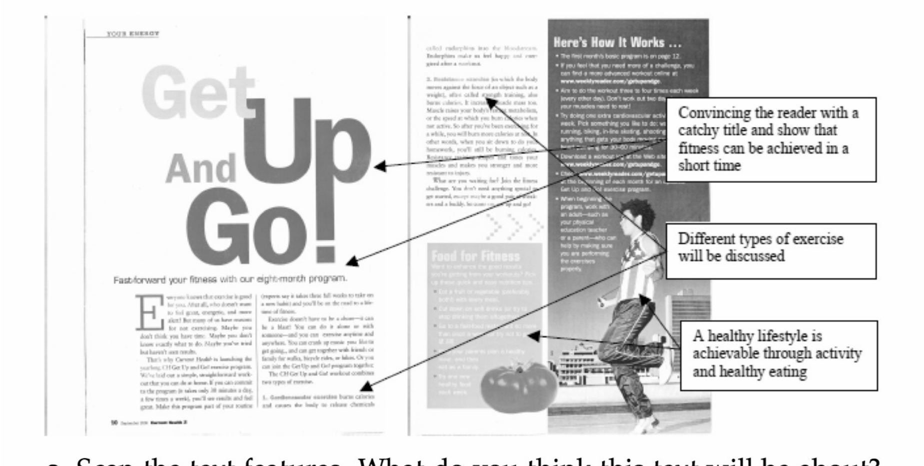
(b) Latent Classes Unaligned with Levels

Figure 3: Diagram representing the Latent Class Level PCM.

The *Striving Readers* project

The *Striving Readers* project provides an example of the assessment context for the empirical setting of ordinal performance levels and the type of data needed for the proposed method.

The literacy intervention *Strategies for Literacy Independence across the Curriculum* (SLIC) focuses on teaching students how different text forms can be used to present particular types of information, and how to use the features of a text form to gain information about the text's content (McDonald, Thornley, Staley, & Moore, 2009; Institute for Education Sciences, 2006). The intervention was funded by the Institute for Education Sciences. The goal of the partner *Striving Readers* project was to develop an assessment framework for SLIC, including the construct, items, scoring guides, and assessments. In partnership with the curriculum developers and district personnel, a team from the Berkeley Evaluation and Assessment Research (BEAR) Center used the BEAR Assessment System (BAS) (Wilson, 2005) to create and refine the set of assessments (Dray, Brown, Lee, Diakow, & Wilson, 2011). An example *Striving Readers* item is shown in Figure 4.


YOUR ENERGY
Get Up And Go!
Fast-forward your fitness with our eight-month program.
Here's How It Works ...
Food for Fitness
Convincing the reader with a catchy title and show that fitness can be achieved in a short time
Different types of exercise will be discussed
A healthy lifestyle is achievable through activity and healthy eating

2. Scan the text features. What do you think this text will be about?

Figure 4: A *Striving Readers* sample item. The white boxes and arrows describe text elements. These hints were not shown to respondents.

The construct developed by the assessment team contains five ordered levels, from *Disengaging* (the lowest level) through *Synthesizing* (the highest). All items are designed to assess the same construct and are scored polytomously from 0-4. The scoring guides link the scores given on the item directly to the constructs, such that an assigned score of 0 on an item indicates that the student's response to this item displays evidence of reading comprehension at the *Disengaging* level, and a score of 4 indicates comprehension at the *Synthesizing* level. Figure 5 shows a scoring guide for the *Striving Readers* item shown in Figure 4.

The strong connection between the levels of the *Striving Readers* construct and the *Striving Readers* items is the key feature of this assessment which will be used in this paper. The presence of this connection motivates the possibility of empirically setting performance levels in relation to the polytomous item scores and then classifying students into ability levels using their probability of achieving these levels on assessment items.

	Construct Description	Scoring Guide
4	**Synthesizing - Creating New Key Ideas** - new understanding based upon the text - new understanding based upon multiple texts - evaluating author's intent - literary and/or rhetorical criticism	Response is **complete in relation to the information contained:** *Example*: This article is convincing you to get healthy by describing a program of exercise and eating right. It also encourages to do exercise and suggests that it is not hard.
3	**Cross-checking - Coordinating Key Ideas in the Text** - claim - argument - theme - identifying author's intent	Student responds with **multiple items from tactics-based sources and cross-checks or combines items of information:** *Example*: This article is about convincing us to stay fit and healthy.
2	**Discriminating - Key Ideas in the Text** - idea structure - supporting statement - plot - characterization	Student responds with **at least two items from tactics-based sources:** *Example*: Exercise is good for you. We should all exercise.
1	**Engaging - Ideas in the Text** - topic - main idea of a paragraph	Student responds with **one item of information from tactics-based source:** *Examples:* - A fitness plan - Exercise is good for you - Everyone should exercise - Changing your diet can enhance fitness results
0	**Disengaging - Ideas Not in the Text** - not challenging existing knowledge - no new ideas	Student gives an **incorrect response:** *Example*: It's about how you should exercise

Figure 5: The *Striving Readers* construct map and the scoring guide for one item.

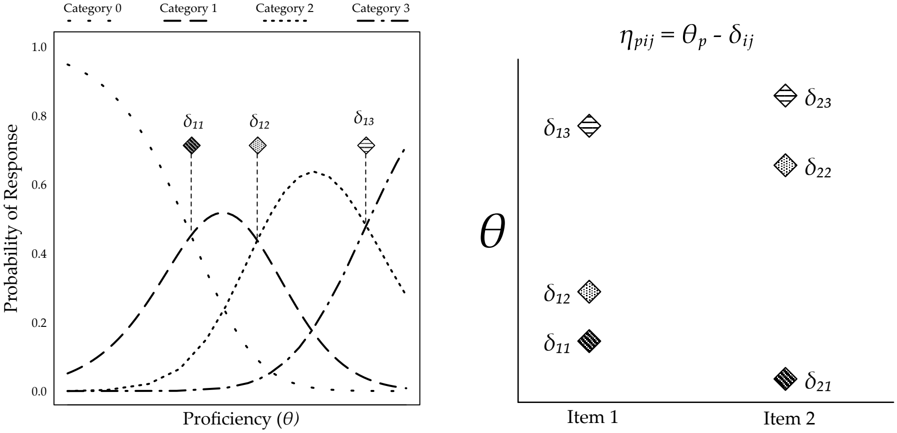
The Level Partial Credit Models

The Level PCM

Using the *Striving Readers* project as their empirical example, Diakow et al. (2013) present a reparameterization of Master's Partial Credit Model (PCM) aimed at making an explicit link between the estimated item level difficulties and the theoretical hypothesized levels. The PCM defines the logit of the probability of person p of answering item i at level j rather than level $j-1$ as:

$$\text{logit}[Pr(x_{pij} = 1 \mid \theta_p)] = \eta_{pij} = \theta_p - \delta_{ij} \tag{1}$$

where θ_p represents the ability of person p, and δ_{ij} the difficulty of category j in item i (Masters, 1982). When θ_p is equal to δ_{ij}, the respondent is equally likely to reach levels j and $j-1$. Figure 6 illustrates this model. Figure 6a shows the relationship between the δ_{ij} parameters and the probability of responding at a given level to the item. Under this representation, there are no parameters representing the main effects of items or levels, obscuring the connection between the construct levels and the model results.



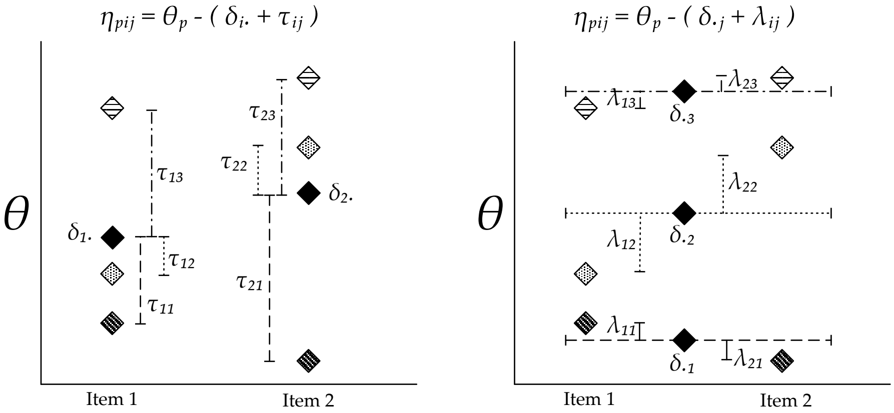
(a) The locations of δ_{ij} parameters for a single item on a category characteristic curve.

(b) The locations of δ_{ij} parameters on an item map.

Figure 6: Sample illustration showing the standard parameterization of Master's Partial Credit Model.

One common reparameterization of the PCM estimates a main effect $\delta_{i.}$ for each item i, with additional parameters τ_{ij} representing the additional deviation for each level ij from the mean difficulty for item i. This approach to the PCM is illustrated in Figure 7a.

By using a main effect for each item, this reparameterization is suited for analyses in which the difficulty of each item is the crucial variable.



(a) The item-focused reparameterization. **(b)** The level-focused reparameterization.

Figure 7: Sample illustration showing two reparameterizations of Master's Partial Credit model.

As an alternative, Diakow et al. (2013) propose a different reparameterization, the Level Partial Credit Model (L-PCM). This reparameterization, illustrated in Figure 7b, estimates a main effect parameter $\delta_{.j}$ for each level j, with additional λ_{ij} parameters representing the deviations for each item-level:

$$\text{logit}[Pr(x_{pij} = 1 \mid \theta_p)] = \eta_{pij} = \theta_p - (\delta_{.j} + \lambda_{ij}) \tag{2}$$

Under this reparameterization, a respondent p with ability level θ_p equal to $\delta_{.j}$ for some level j is, on an "average" item, equally likely to reach level j as level $j-1$. If we interpret the item levels as corresponding directly to person levels (as is the case with the *Striving Readers* assessment), then we can think of respondent p as having an ability level at the borderline between levels j and $j-1$. In other words, we can treat the $\delta_{.j}$ parameter as the cutpoint between the two levels. (A Wright Map — also known as an item-person map — illustrating this idea is shown in Figure 8.) This reparameterization method thus provides us with a direct method of estimating level cutpoints, without requiring a separate standard setting process.

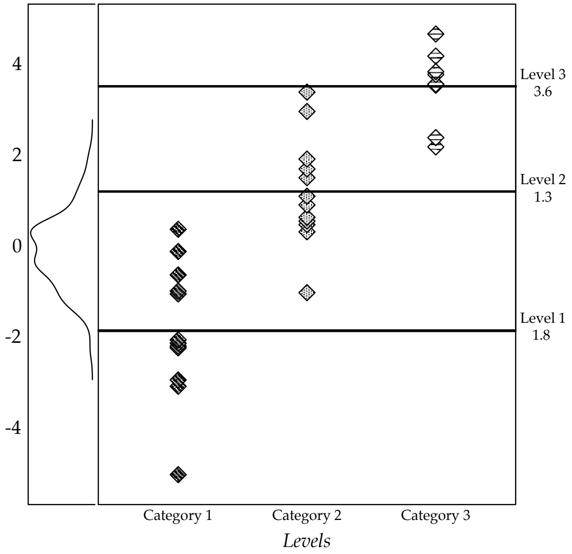


Figure 8: Wright map organizing the λ_{ij} parameters as deviations of the $\delta_{.j}$ level parameters.

The Latent Class Level-PCM

The L-PCM estimates the locations of cutpoints along a continuous latent trait, to classify estimated respondent locations among hypothesized discrete ordered levels. However, since the goal is to classify respondents into groups, another alternative is to use a latent class model to directly estimate class locations and respondent class membership. Located latent class models (Formann, 1995; Hagenaars & McCutcheon, 2002; Lindsay, Clogg, & Grego, 1991) follow latent class analysis (Hagenaars & McCutcheon, 2002; Lazarsfeld & Henry, 1968) in directly modeling proficiency groups. The Latent Class Partial Credit Model (Latent Class PCM) estimates the logit of the probability that person p in class c will answer item i at level j rather than level $j-1$ as

$$\text{logit}[Pr(x_{pij} = 1 \mid \theta_{c(p)})] = \eta_{pij} = \theta_{c(p)} - \delta_{ij} \tag{3}$$

where $\theta_{c(p)}$ represents the centroid of class c and δ_{ij} is the difficulty parameter associated with level j in item i.

The Latent Class Level Partial Credit Model (Latent Class L-PCM) combines the reparameterization approach of the L-PCM with the latent class analysis of the Latent Class

PCM. The model estimates the logit of the probability that person p in class c will answer item i at level j rather than level $j-1$ as

$$\text{logit}[Pr(x_{pij} = 1 \mid \theta_{c(p)})] = \eta_{pij} = \theta_{c(p)} - (\delta_{.j} + \lambda_{ij}) \tag{4}$$

where $\theta_{c(p)}$ is as in the Latent Class PCM, and $\delta_{.j}$ and λ_{ij} are as in the L-PCM.

The Latent Class L-PCM thus estimates the locations of both respondent classes $\theta_{c(p)}$ and level cutpoints $\delta_{.j}$. If the model fits, the respondents do group in the hypothesized classes, and the item scores do reflect ability at the hypothesized levels, then we expect the classes to be located between cutpoints. A comparison of the two sets of estimated parameters thus provides a check of model fit.

Analysis and Results

All the analyses were conducted using Latent Gold 4.5 (Vermunt & Magidson, 2005), and the plots were prepared using R (R Core Team, 2013).

Empirical data

The data used to illustrate these methods come from the *Striving Readers* project. Sixteen assessments were developed to be given to San Diego Unified School District students in four grades (7-10) four times a year (September, December, March, and June) as part of an experimental study (Dray et al., 2011). Prior to this study, a calibration study was performed in New Zealand in the summer of 2008. The New Zealand students also ranged from grades 7-10, but completed the assessments through an overlapping design that allowed for linking the assessments and vertical scaling across the grades.

The dataset used in this article consists of the responses from the New Zealand 7th graders who had complete data for one of the subtests. The test had 12 items, and there were 202 students in the sample. This sample was also used in Diakow et al. (2013). Due to the lack of responses scored in the highest category (*Synthesizing*) in the 7th grade sample, the data show only four item categories (and, for some items, only three). Accordingly, the analysis focuses on the determination of the three boundaries between the four levels and the location of the classes of respondents for those levels.

Number of latent classes

As is the case in any latent class analysis, the selection of the appropriate number of classes is an issue for the Latent Class L-PCM. In this paper, we focus on a case where the theory provides an initial hypothesis regarding the number of classes that we might expect to see across the entire population of interest. However, it is not always clear that the sample of respondents will in fact represent the entirety of that range, even if the initial hypothesis about the number of classes is correct. While the theory provides us with a starting point, it is also necessary to examine whether empirically it makes sense to conduct the analysis using a given number of classes.

In this paper we follow standard practice in the field (Nylund, Asparouhov, & Muthén, 2007), by comparing the fit of models with different numbers of latent classes in terms of both the Akaike Information Criterion (AIC; Akaike, 1987) and the Bayesian Information Criterion (BIC; Schwarz, 1978). BIC penalizes parameter usage more severely than AIC; as a result, BIC is more likely to underfit the data but will tend to select more parsimonious models while AIC is more likely to overfit the data but will tend to select models that detect more subtle features in the data (see Dziak, Coffman, Lanza, & Li, 2012, and Vrieze, 2012, for recent reviews of the comparison between AIC and BIC). In the context of latent class models, this means that when they disagree, AIC would indicate models with more latent classes and BIC models with fewer. When the decision made would differ based on which criteria is used, the choice of which information criterion to follow relies heavily on the judgment of the researchers, whose decisions are made in light of their initial theory.

Based on the structure of the *Striving Readers* construct, we have an initial hypothesis of four classes. We conducted an analysis to determine empirically the optimum number of classes to use for this sample. If, for instance, the sample did not contain many students at the hypothesized highest level, it is possible that a model with fewer levels may be more appropriate.

Figure 9 shows AIC and BIC values for different numbers of latent classes. The lowest AIC value is found when there are four latent classes, while the lowest BIC value occurs in the case where there are three latent classes. Considering that the primary purpose of the empirical example in this paper is to illustrate the Latent Class L-PCM, rather than make substantive conclusions based on the data, we decided to conduct the analysis in this paper using four classes because it is consistent with the initial hypothesis. However, when helpful, we include comparison information from the models with other number of classes. In work applying this model to draw substantive conclusions, additional evidence, rather than just a match to the initial hypothesis, should be used to decide between models when the AIC and BIC support different numbers of classes.

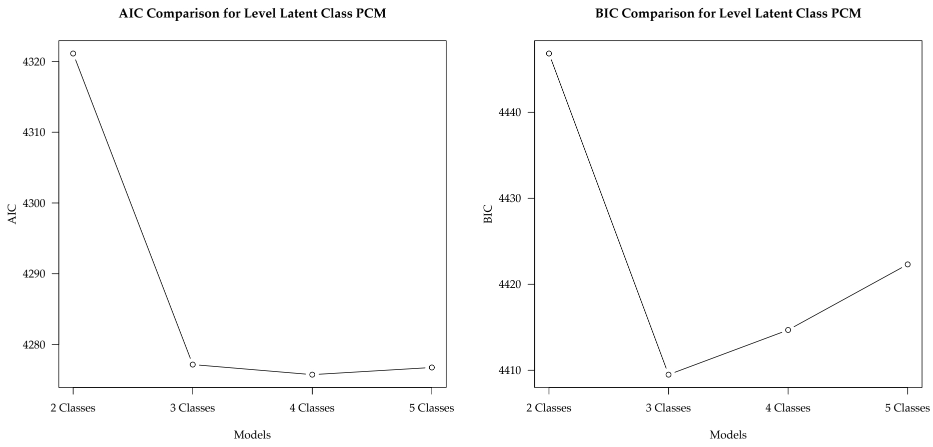


Figure 9: Comparison of AIC and BIC values for different numbers of latent classes.

Class and cutpoint comparisons

The first results that the Latent Class L-PCM provide are the set of parameters that represent (a) the category boundaries (i.e. the $\delta_{.j}$), (b) the respondent class locations ($\theta_{c(p)}$), and (c) the item specific interaction parameters (λ_{ij}). By plotting these parameters in a Wright Map, we can quickly examine the location of the classes on the latent continuum and the corresponding performance levels as determined by the $\delta_{.j}$ boundaries.

Figure 10 shows these results. The left side of the figure shows a histogram of the location of respondents in the sample, where a respondent's location is given by the average of the estimated class locations, weighted by the respondent's estimated probability of being in that class. These locations are therefore constrained to lie between those estimated for Class 1 and Class 4. The right side of the figure shows the estimated locations of the item parameters for the first, second, and third levels. (Note that there are several items with no third λ_{ij} parameter.) The estimated class locations are represented as dashed horizontal lines, and the estimated level cutpoints as solid horizontal lines.

In an ideal scenario, we would expect to recover each class as associated with a different performance level; this plot allows us to examine to what extent the recovered classes are associated to the different performance levels identified through the level boundaries. Figure 10 shows that the class locations recovered by the latent class analysis do not seem to be located in regions associated with the four levels of performance, indicating

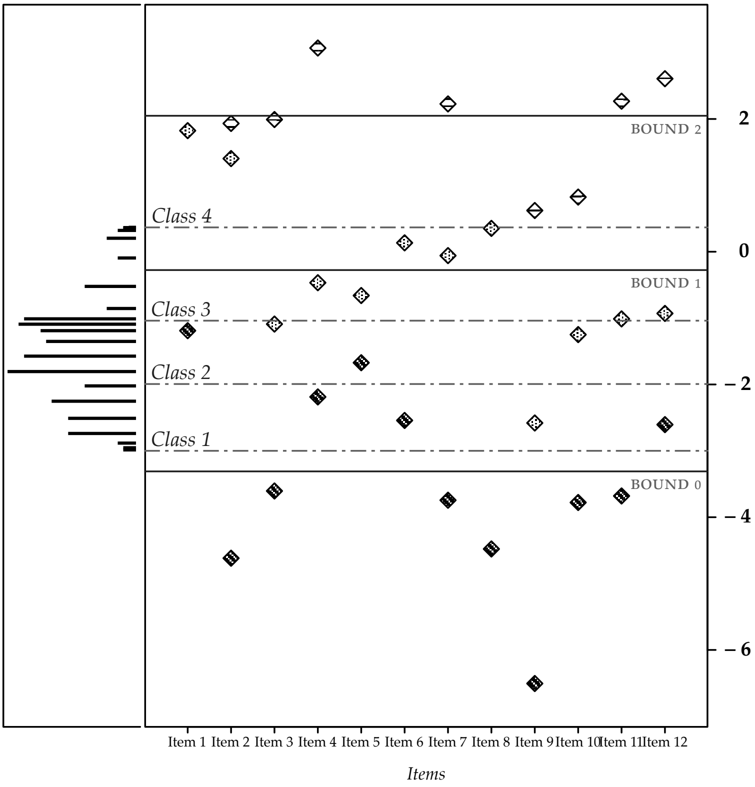


Figure 10: Results of the Latent Class L-PCM.

that only two of the performance levels seem to be represented by these classes. The lowest three classes are all estimated to lie between the cutpoints demarcating level 1, while the estimated location for the third class lies between the cutpoints for level 2.

For comparison, Figure 11 shows the results from the PCM model. On this plot, the left side shows the estimated values of θ_p for each respondent, while the right side shows the estimated item-level difficulties (i.e. λ_{ij} parameters) and level cutpoints. From this graph, it is clear that most respondents were well above the first difficulty level for most items, and well below the highest difficulty level. This lack of cover of respondents across the levels is echoed in Figure 10 by the high location of the lowest class and the low location of the highest class. The latent class analysis is unable to recover class locations for respondents not present in the sample. It distinguishes three classes within level 1, which could be considered as an emerging level 1 (close to the boundary between level 0 and level 1), a prototypical level 1, and an advanced level 1 (located close to the level 2 boundary).

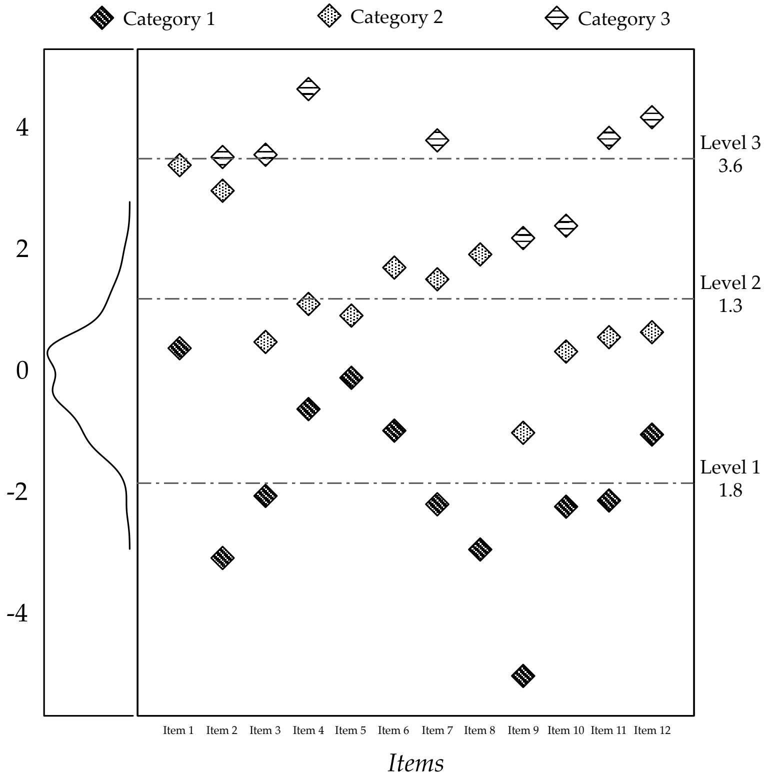


Figure 11: Results of the PCM.

Ideal cases

In addition to comparing cutpoints and class locations, Diakow et al. (2013) identify two possible methods for evaluating the appropriateness of the L-PCM that we can extend to the Latent Class L-PCM. The first, plotting ideal cases, involves examining the locations of *prototypical* and *boundary* examinees in relation to the estimated cutpoint locations. As illustrated in Figure 12, a *prototypical* level 2 respondent receives scores of 2 on all

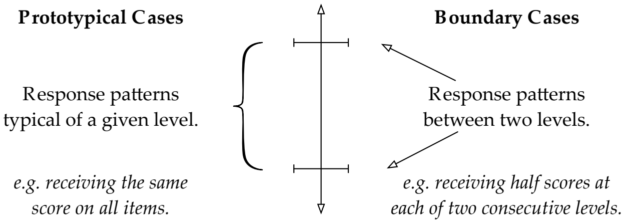

Prototypical Cases
Boundary Cases
Response patterns typical of a given level.
Response patterns between two levels.
e.g. receiving the same score on all items.
e.g. receiving half scores at each of two consecutive levels.

Figure 12: Illustration of both kinds of ideal cases: prototypes and boundary cases.

items, while a level 1/2 *boundary* respondent receives scores of 1 on half the items, and 2 on the other half.

Figure 13 plots the average item scores of respondents in the *Striving Readers* dataset against their ability location as estimated using the L-PCM, illustrating the association between each estimated θ_p value and the estimated performance levels.

In this figure, the vertical line segments through each point show the standard errors of the estimate. The dashed vertical lines show the location of prototypical respondents, while the solid vertical lines show the locations of boundary respondents. The solid horizontal lines show the level cutpoints: Between levels 0 and 1, between levels 1 and 2 and between levels 2 and 3. The shaded boxes show the expected locations of respondents within the plot. The expectation is that students with average item scores above 0.5 (i.e., above 0/1 boundary examinees) will be above the cutpoint for level 1, while students with average item scores below 1.5 (i.e., below 1/2 boundary examinees) will be below the cutpoint for level 2. As shown in Figure 13, nearly all the points from the *Striving Readers* L-PCM do fall in the expected regions.

Figure 14 shows the same plot for the Latent Class L-PCM with 4 classes. The dashed horizontal lines show the class locations. As in Figure 10, a respondent's θ estimate is equal to the average of the estimated class locations, weighted by the respondent's estimated probability of being in that class. This again means that these locations are constrained to lie between the locations of the lowest and highest classes. In the Latent Class L-PCM, as in the L-PCM, we see that nearly all the plotted points lie inside the expected shaded regions.

Figure 15 shows the estimated ability location as a function of average item score for models with 2–5 latent classes. The solid horizontal lines show the estimated locations of the level cutpoints, while the dashed horizontal lines show the estimated class locations. Interestingly, with only 2 latent classes, *both* are estimated to be located between the level 1 and level 2 cutpoints. If the goal is for the latent classes to correspond to the levels

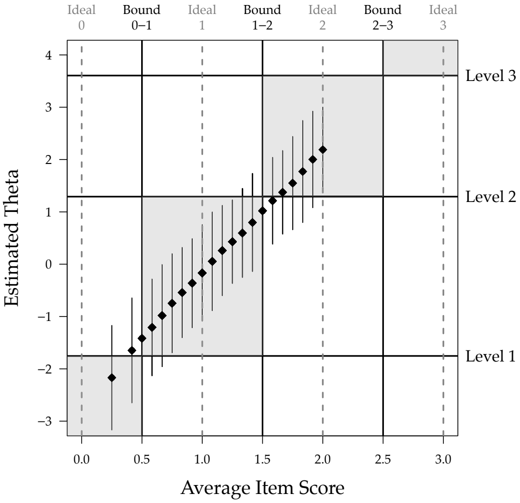


Figure 13: Plotting proficiency estimates from the L-PCM in relation to estimated cut-points and ideal cases.

demarcated by the cutpoints, then the model with 3 classes represents an improvement over the model with only 2, as it contains two classes estimated at the second level and one estimated at the third level. Adding a fourth class simply adds another at the second level, and the fifth adds for the first time a class in the first level. From these graphs, it appears the models with 3 and 5 classes correspond the most closely to the hypothesized level structure, though the model with 5 classes is likely overfitting the given sample. More importantly, these results suggest investigating whether there may be at least one additional level that can be distinguished between levels 1 and 2. The consistent empirical finding indicates that revision might be needed to the hypothesized theory of reading development.

Expected score ranges

The second method identified by Diakow et al. (2013) to evaluate the performance of the L-PCM is to examine the set of expected scores for respondents in each level. In this way, we can better understand the predicted performance of the members of each latent class in terms of the locations of their expected scores for each item and obtain additional information based on the dispersion of these expected scores.

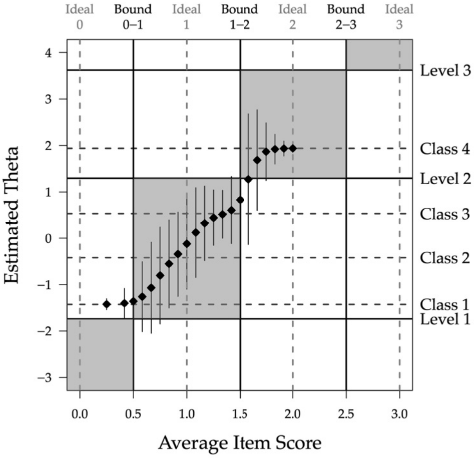


Figure 14: Plotting proficiency estimates from the Latent Class L-PCM in relation to estimated cut-points, ideal cases, and class locations.

The expected score s_i for a respondent p on item i is given as a function of respondent ability θ_p, as

$$\mathbb{E}[s_i \mid \theta_p] = \sum_j j \cdot Pr(x_{pij} = 1 \mid \theta_p) \tag{5}$$

Figure 16a shows the expected score under the L-PCM for each item as a function of θ, together with the level cutpoints. On the rightmost edge of this figure, it is possible to distinguish two sets of items: one in which the expected score approaches an upper asymptote of 3, and another, smaller set of items approaching an upper asymptote of 2. The smaller set consists of items with no student responses scored as 3, leading to no estimated λ_{ij} for the last boundary parameter, so an expected score above 2 is impossible.

For the L-PCM, there are a range of θ estimates within a given level, and thus no clear definition of what the expected scores for a “typical” respondent in that level would be. One possible approach is to take the average of the expected score function over the θ interval comprising that level. Within a middle level with a lower cutpoint a and upper cutpoint b, the average expected score is given by:

$$\frac{1}{a-b} \int_a^b \mathbb{E}[s_i \mid \theta] \; d\theta \tag{6}$$

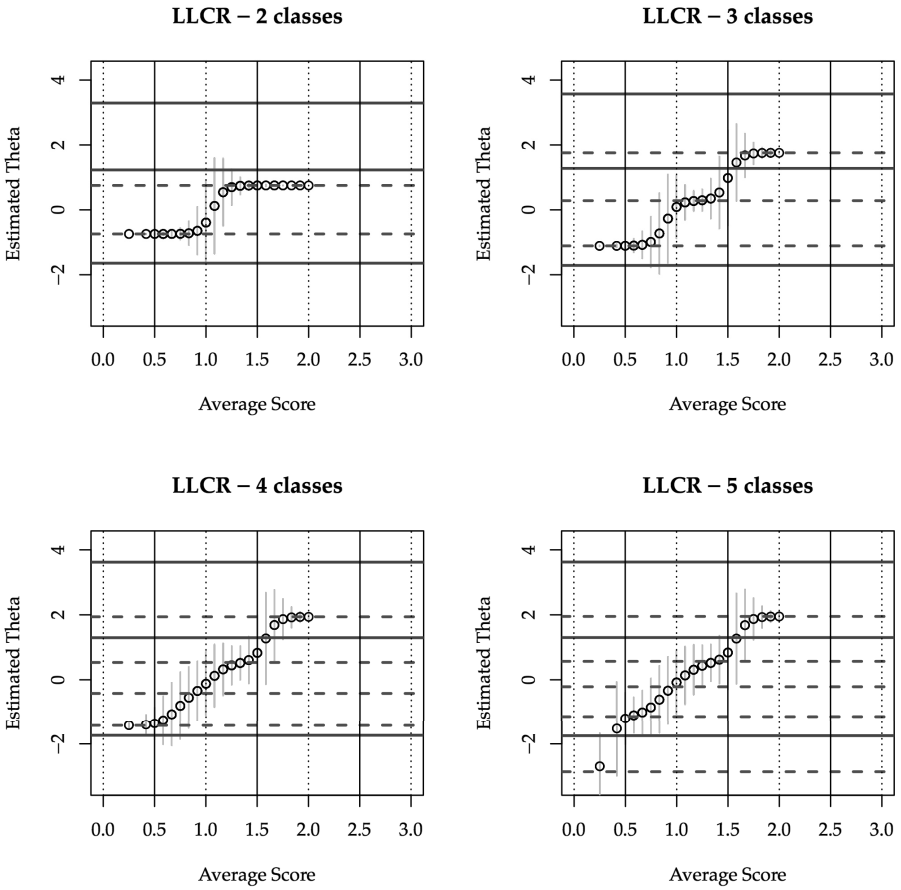
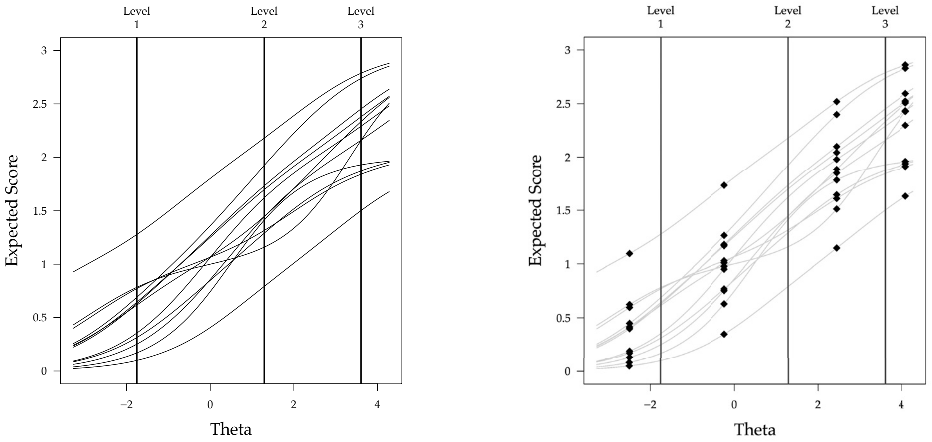


Figure 15: Proficiency estimates in relation to estimated cut-points for different numbers of ideal cases.

In addition to its potential mathematical complexity, this method also has the disadvantage that there is no clear way to define the "average" expected value for the infinitely wide lowest and highest levels. For these reasons, Diakow et al. (2013) propose an alternate method of evaluating the expected scores for a level under the L-PCM. Under this method, a single θ value within that level is selected, and the expected scores $\mathbb{E}[s_i \mid \theta]$ for each item i are then calculated. For middle levels bounded by both an upper



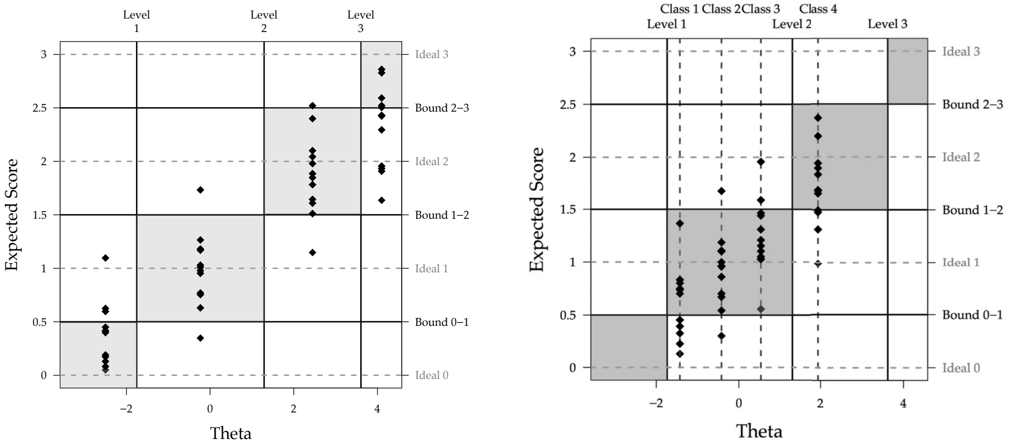
(a) Expected scores as a function of θ.

(b) Expected scores for selected respondents.

Figure 16: Expected score plots for the L-PCM.

and lower cutpoint, the midpoint of the level region is selected. A midpoint cannot be calculated for the highest and lowest levels, so a point has to be chosen arbitrarily to represent the probabilities for that level (in this case, we chose a point that was .75 logits below the first boundary and one that was .5 logits above the last boundary). However, it is worth mentioning that this issue is resolved under the Latent Class L-PCM, as the latent class location estimates provide us with a location for all the levels.

Figure 16b shows the same plot as Figure 16a, with point estimates added showing the expected scores under the L-PCM for a selected θ value within each level. Figure 17a removes the expected score curves. It also adds dashed lines indicating the points at which the expected score is exactly an integer number of points, and solid horizontal lines demarcating four regions such that points below the lowest solid line have expected scores of approximately 0, those between the lowest two lines have expected scores of approximately 1, and so on. If the items are behaving appropriately and item score levels correspond to the student ability levels as predicted, students at a certain level should have expected scores at approximately that level. The shaded areas then indicate locations in which a student classified as a certain group, based on their location relative to the cutpoints, has an expected score for an item within the intended range for that level.



(a) Expected scores for the L-PCM.

(b) Expected scores for the Latent Class L-PCM.

Figure 17: Expected scores as a function of θ for the L-PCM and the Latent Class L-PCM.

For the majority of items and levels, the expected scores are within the desired ranges. From comparing Figures 16b and 17a, it is apparent the item with an expected score below the second and third shaded boxes, as well as all the items with expected scores below 2 even for the highest θ value, are those with no third λ_{ij} parameter, so the probability of receiving the highest possible score is estimated to be 0. The majority of the points outside the shaded boxes belong to the sample respondents in the highest level, indicating that respondents in that level may have expected scores closer to 2 than to 3 on a number of items.

For the L-PCM, we selected a θ value from within each level to plot. For the Latent Class L-PCM, since respondent ability is modeled as a series of located latent classes, we can simply plot the expected scores for members of each class. This plot is shown in Figure 17b, with the locations of the latent classes given by the dashed vertical lines.

As noted above, the estimated locations for the first three classes all lie between the estimated locations of the first and second cutpoints. Using the levels demarcated by the cutpoints, the items appear to be performing acceptably, with the majority inside the shaded boxes. However, the latent classes recovered by the model do not seem to correspond as well to the expected scores. For members of the lowest class, there are a number of items with expected scores close to 0. For members of the second class, nearly all the expected scores are close to 1. But members of the third class have expected scores mainly greater than 1 on most items. Members of the highest class seem to be clearly located within the second-highest level. Thus, while the second and fourth classes have expected item scores that correspond to their placement within the level boundaries, the first and third classes have expected item scores that lie across performance levels. As above, this indicates that it may be useful to further differentiate between levels of performance within the current level 1. In addition, there is no class with items for which the expected score is closer to 3 than to 2.

Discussion

The Latent Class L-PCM presented in this paper can help practitioners connect the theory-based performance levels that motivate their work to the results of their psychometric models. In the case of the *Striving Readers* assessment, the analysis helped us learn that, while the items could be used successfully to segment the latent continuum into regions associated with each performance level, the latent subgroups present in this sample of respondents were concentrated in only two of the four levels differentiated by the items. The possibility of identifying empirically the location of the latent classes in relation to the level cutpoints and potentially rejecting an expected interpretation of the latent classes, as was the case in this analysis, is an important benefit of this model. It can save us from the risk of finding the expected number of classes and simply assuming they align to the theoretical classes.

This analysis, focused on a single assessment of the *Striving Readers* project, has some limitations worth noting. A first issue is the restricted range of proficiency among the respondents, which led to the absence of observed scores on the upper levels of the construct, and consequently made it impossible to explore the upper range of the construct. A second issue, related to this restriction in the range of proficiency, is that a few items only had responses in the first three levels, which made impossible the estimation of the last λ_{ij} parameter with a reasonable uncertainty for those items. These limitations may apply to any empirical standard setting method, and the method proposed here highlights rather than hides them.

Using the Latent Class L-PCM is relatively straightforward mathematically. However, conducting this kind of analysis demands considerable work in advance in the development of the theoretical levels, the creation of items that target those levels, and the construction of scoring rubrics that maintain that connection. We believe that this kind of upfront investment in the design and development of the constructs, assessment instruments, and scoring rubrics is good practice in general. The importance of this investment for applying the proposed model only reveals that the strength of these components underlies the validity of other standard setting methods as well.

The use of the Latent Class L-PCM could be of particular interest to practitioners who, needing a classification procedure for the respondents, would usually rely solely on a standard setting procedure. The results illustrating how this model estimates both cutpoints as well as latent class locations demonstrate how the Latent Class L-PCM can potentially be used as an additional input to judges in a more traditional standard setting context or, potentially, as an empirical alternative to the determination of cutpoints by human judges. The cutpoints and class locations established through the use of the Latent Class L-PCM could also be compared with the determinations of experts presented with the same results. The possible application of this method in a standard setting context merits further research.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika*, *52*(3), 317–332.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (p. 508-600). Washington: American Council on Education.

Bourque, M. L. (2009). *A History of NAEP Achievement Levels: Issues, Implementation, and Impact 1989-2009.* (Tech. Rep.). National Assessment Governing Board. Retrieved from http://www.nagb.org/publications/reports-papers/achievement-levels/history-naep-achievement-levels-1989-2009.html

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, N.J: Lawrence Erlbaum.

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: contemporary methods. *Educational Measurement: Issues and Practice*, *23*(4), 31–31.

Croon, M. (1990). Latent class analysis with ordered latent classe. *British Journal of Mathematical and Statistical Psychology*, *43*(2), 171–192.

Diakow, R., Torres Irribarra, D., & Wilson, M. (2013). Some comments on representing construct levels in psychometric models. In R. Millsap, L. van der Ark, D. Bolt, & C. Woods (Eds.), *New developments in quantitative psychology: Presentations from the 77th annual psychometric society meeting* (pp. 319–334). Berlin: Springer.

Dray, A., Brown, N., Lee, Y., Diakow, R., & Wilson, M. (2011). *Striving Readers BEAR Assessment Report* (Tech. Rep.). Berkeley Evaluation and Assessment Research Center.

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria* (Tech. Rep. No. 12-119). The Methodology Center and Department of Statistics, Penn State, The Pennsylvania State University.

Formann, A. K. (1995). Linear logistic latent class analysis and the Rasch model. In *Rasch models: Foundations, recent developments, and applications* (pp. 239–255). New York: Springer-Verlag.

Glaser, R., Chudowsky, N., & Pellegrino, J. W. (2001). *Knowing What Students Know: The Science and Design of Educational Assessment.* Washington, DC: National Academies Press.

Gonzales, P., Williams, T., Jocelyn, L., Roey, S., Kastberg, D., & Brenwald, S. (2008). *Highlights from TIMSS 2007: Mathematics and Science Achievement of US Fourth- and Eighth-Grade Students in an International Context* (Tech. Rep.). Washington, DC: National Center for Education Statistics.

Hagenaars, J. A., & McCutcheon, A. L. (2002). *Applied latent class analysis.* Cambridge, New York: Cambridge University Press.

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*(4), 355–366.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, *8*(1), 41–55.

Institute for Education Sciences. (2006). *Striving Readers Program.* Retrieved from `http://www2.ed.gov/programs/strivingreaders/index.html`

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Boston, MA: Houghton Mifflin.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Ed.), *IRT-based standard setting procedures utilizing behavioral anchoring. symposium conducted at the council of chief state school officers national conference on large-scale assessment.* Phoenix, AZ.

Lindsay, B., Clogg, C. C., & Grego, J. (1991). Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *Journal of the American Statistical Association*, *86*(413), 96–107.

Lubke, G. H., & Muthén, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

McDonald, T., Thornley, C., Staley, R., & Moore, D. W. (2009). The San Diego Striving Readers' project: Building academic success for adolescent readers. *Journal of Adolescent & Adult Literacy*, *52*(8), 720–722.

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*(4), 535–569.

Perie, M. (2008). A guide to understanding and developing performance-level descriptors. *Educational Measurement: Issues and Practice*, *27*(4), 15–29.

Programme for International Student Assessment. (2007). *PISA 2006: Science Competencies for Tomorrow's World. Volume 1, Analysis* (Tech. Rep.). Paris: Organisation for Economic Cooperation and Development.

R Core Team. (2013). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria.

Reckase, M. D. (2000). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts conducted by ACT.* Iowa City, IA: National Assessment Governing Board.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD choice 4.0 user's manual.* Belmont, MA: Statistical Innovations, Inc.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*(2), 228-243.

Wilmot, D. B., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning*, *13*(4), 259–291.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, N.J.: Lawrence Erlbaum.

Wilson, M., & Draney, K. (2002). A technique for setting standards and maintaining them over time. In S. Nishisato, Y. Baba, H. Bozdogan, & K. Kanefugi (Eds.), *Measurement and multivariate analysis (proceedings of the international conference on measurement and multivariate analysis, banff, canada, may 12-14, 2000)* (pp. 325–332). Tokyo: Springer-Verlag.