iejee

International Electronic Journal of Elementary Education
March 2025, Volume 17, Issue 2

Improving Context Scale Interpretation Using Latent Class Analysis for Cut Scores

Liqun Yin[a], Ummugul Bezirhan[b], Matthias von Davier[c]

Received : 6 January 2025
Revised : 23 January 2025
Accepted : 2 March 2025
DOI : 10.26822/iejee.2025.378

[a] **Corresponding Author:** Liqun Yin, TIMSS & PIRLS International Study Center, Boston College USA.
E-mail: yinld@bc.edu
ORCID: https://orcid.org/0009-0005-1919-3548

[b] Ummugul Bezirhan, TIMSS & PIRLS International Study Center, Boston College, USA.
E-mail: bezirhan@bc.edu
ORCID: https://orcid.org/0000-0002-8771-4780

[c] Matthias von Davier, TIMSS & PIRLS International Study Center, Boston College, USA.
E-mail: vondavim@bc.edu
ORCID: https://orcid.org/0000-0003-1298-9701

Abstract

This paper introduces an approach that uses latent class analysis to identify cut scores (LCA-CS) and categorize respondents based on context scales derived from large-scale assessments like PIRLS, TIMSS, and NAEP. Context scales use Likert scale items to measure latent constructs of interest and classify respondents into meaningful ordered categories based on their response data. Unlike conventional methods reliant on human judgments to define cut points based on item content, model-based approaches such as LCA find statistically optimal groups, a categorical latent variable, that explains item score differences based on score distribution differences between latent classes. Cut scores for these classes are determined by conditional probability calculations that relate class membership to observed scores, finding the intersection point of adjacent smoothed probability distributions and connecting it to the construct. Demonstrated through application to PIRLS 2021 data, this is useful to validate existing categorizations of the context scale by human experts, and can also help to enhance classification accuracy, particularly for scales exhibiting highly skewed distributions across diverse countries. Recommendations for researchers to adopt this LCA-CS approach are provided, demonstrating its efficiency and objectivity compared to judgment-based methods.

Keywords:

Context Scales, Latent Class Analysis, Cut Scores, Large-scale Assessments

Introduction

In educational assessments of achievement, standard-setting has been used for meaningful interpretation of test scores and for making decisions that impact students' educational trajectories, such as screening students for instruction, grade promotion, selection, or admission (e.g., Cizek & Bunch, 2007; Cizek, 2012; Jiao et al., 2011). Performance standards, which are set through carefully determined cut scores, serve to classify examinees into defined proficiency levels, in doing so, guiding stakeholders' understanding of individuals' competencies relative to a given domain (Cizek, 2012). Therefore, standard-setting is central to establishing that assessments function not only as measurement tools but also as benchmarks for educational quality and progress.

Traditionally, standard-setting methods implemented for achievement instruments have relied on subject matter experts (SMEs) to interpret the content of assessment items and determine cut scores that align with descriptions of performance levels (Cizek, 1993). These methods are generally categorized as test-centered, where SMEs focus on individual test items, or examinee-centered, where judgments are based on examinee performance rather than specific item content (Jaeger, 1989). Methods such as the Angoff procedure (Angoff, 1971), bookmark method (Mitzel et al., 2013), and contrasting groups method (Livingston & Zieky, 1989) are widely used in standard-setting. In these approaches, SMEs discuss the difficulty of test items and the expected performance of a "borderline" examinee to set a threshold for each proficiency level (Cizek, 2005; Peabody et al., 2023). The Angoff and bookmark methods are test-centered, as they focus on the properties of individual test items, with SMEs evaluating item difficulty to estimate the performance of a minimally competent examinee. In contrast, the contrasting groups method is examinee-centered, as it relies on SMEs classifying examinees directly based on their overall performance relative to the standard.

In addition to test-centered and examinee-centered distinctions, standard-setting methods can be classified as holistic or analytical, norm-referenced, or criterion-referenced. Holistic methods involve evaluating overall performance levels, while analytical methods break down performance into specific competencies or skills. Norm-referenced methods set performance standards by comparing the examinee's performance to a reference group, whereas criterion-referenced methods define standards based on specific performance criteria or competencies (Cizek, 2012). Similar to the test-centered versus examinee-centered distinction, these categorizations, while conceptually useful, tend to overlap in practice, as most standard-setting approaches combine elements of various methodologies to comprehensively evaluate examinee proficiency levels.

Although well-established, these methods require intensive cognitive effort from experts to consider both the test content's characteristics and the abilities of the target population. They are susceptible to inconsistencies due to variations in judgment, especially across diverse contexts (Brown, 2007; Cizek, 2012).

To address the limitations of traditional judgment-based approaches, recent research has explored data-driven methods for setting cut scores (Binici & Cuhadar, 2022; Brown, 2007; Peabody et al., 2023; Templin & Jiao, 2012). Latent class analysis (LCA; Dayton & MacReady, 1976, 2006; Lazarsfeld & Henry, 1968) has emerged as an appealing alternative for establishing cut scores in a statistically objective manner. LCA, a categorical latent variable modeling technique, identifies groups within a population based on response patterns rather than judgment, thus reducing the subjectivity typically associated with standard setting. This approach segments examinees into homogeneous latent classes according to a statistical optimization criterion, effectively distinguishing groups based on the item response distributions within each class. Unlike conventional methods that presuppose a continuous latent trait, LCA models assume that different, discrete latent classes account for variation in observed scores. This enables LCA to categorize individuals into performance levels based on empirical relationships among responses rather than a-priori content-based judgments.

Brown (2007) evaluated the effectiveness of LCA alongside the Angoff procedure and profile rating method for a middle school statistics assessment. This study utilized LCA to categorize students based on response patterns, providing an empirical, data-driven alternative to judgment-based approaches. The results showed that the traditional methods showed strong agreement, with students categorized similarly 85.7% of the time. The LCA showed an even higher alignment with the Angoff method (92.2%) but slightly lower agreement with the Profile method (77.1%), indicating that LCA could reliably classify students into proficiency levels while reducing reliance on expert judgment. Similarly, Binici and Cuhadar (2022) applied LCA to an operational large-scale science assessment administered in one of the southern states in the United States to validate performance standards derived from traditional methods. Their work examined whether LCA could provide additional validity evidence into the classification accuracy of existing cut scores. By analyzing the latent structure within student response patterns, Binici and Cuhadar (2022) demonstrated that LCA could complement conventional judgment-based methods by offering a statistically derived basis for performance standards. These studies showcase the advantages of LCA in creating objective and data-driven cut scores, primarily focusing on setting performance standards for achievement data. While applying LCA to standard settings is not entirely new, its application to background scales remains relatively underexplored.

In large-scale international assessments such as PIRLS (Progress in International Reading Literacy Study) and TIMSS (Trends in International Mathematics and Science Study), context questionnaires are widely used to gather data on students' background through student, school, and home questionnaires. Many of these context items are designed to measure common and dominant underlying latent constructs, such as student motivation, family support, and school resources, which aid in understanding the various factors that relate to student performance. The item

response theory (IRT) based scaling approach is then utilized to derive context scale scores for the items measuring the same latent construct.

In operational settings, context scales are often divided into regions aligned with raw score points and transformed reporting scale cut points. The interpretation of these regions is content-referenced, meaning that each boundary aligns with a combination of response categories. These cut points are often defined through SME judgments. Hence, experts determine what constitutes high or low levels on each scale, sometimes solely based on reviewing the items and response categories, without referencing how respondents use the scale. However, these content-referenced cut-score definitions can result in score regions that contain few or no students, especially when evaluating skewed scale distributions across countries with diverse educational backgrounds.

Current study introduces an LCA-based cut score (LCA-CS) determination approach that addresses the limitations of traditional, judgment-based cut score definitions on context scales. This approach uses LCA with a predefined number of classes determined as the number of ordered categories experts wish to distinguish. LCA identifies groups of examinees based on their observed responses, providing posterior probabilities of class membership for each individual. Examines are then assigned to the most likely class based on the maximum posterior class probability, therefore classifications are statistically grounded rather than subjective expert judgment. After LCA identifies latent classes, which are homogeneous groups within the data, the latent classes are sorted based on the expected mean score for each class. This step reflects the principles of located and ordered latent class models (Clogg 1979; Croon, 1990; Formann 1992; Lazersfeld & Henry, 1968) that the classes are represented by scores on a latent continuum. In our case, the construct's scale score provides this continuum, ensuring that class order is directly related to the underlying latent trait. This can be interpreted as the probability of selecting increasingly positive categories on a rating scale, in the case of context scales, or for cognitive skills, selecting the correct response, which increases as one progresses through a set of latent classes from the lowest to the highest (Croon, 2002), making it particularly useful in contexts where subgroups within a latent trait are to be identified rather than measuring differences between individuals. However, for ordered latent class approach to hold, it is also necessary to verify that the expected scores follow the same order across all items. Additionally, the differences between the expected scores for adjacent classes should be sufficiently large to demonstrate meaningful separation.

Furthermore, we modeled the conditional score distributions for each class independently to identify cut scores that separate adjacent classes. For this, we assume that each latent class represents a homogeneous group, and the conditional distribution of scores within each class follows a normal distribution. The use of conditional normal approximations for score distributions reflects widely applied practices in latent variable modeling, where parametric assumptions are employed to smooth score distributions (e.g., Heinen, 1993, 1996; Embretson & Riese, 2013; Mislevy, 1983; Rost & von Davier, 1995; Smit et al., 2003; Templin & Jiao, 2012). While Formann (1992) emphasizes the relationship between categorical latent variables and response probabilities in linear logistic latent class models, our model ties class membership to a latent continuum. Smoothing these distributions helps cut-score boundaries not to be overly sensitive to random fluctuations in the data. This is particularly important in large-scale assessments where sample sizes and response patterns vary widely across contexts.

When applying LCA, the intersection points of smoothed posterior probabilities between adjacent classes define the cut scores. Then, these cut points are mapped back to the underlying construct. The model integrates categorical class definitions with continuous construct measurement by anchoring these cut scores to the IRT scale. Templin and Jiao (2012) argue for combining latent class models with continuous scaling to enhance the psychometric validity of classifications, while Rost (1990) emphasizes the compatibility of latent class and trait models for defining ordered categories along a latent continuum. Similarly, Croon (1990) and Formann (1992) offer theoretical frameworks for modeling ordered latent classes that align with continuous latent constructs, providing a basis for statistically grounded and construct-aligned classifications. Leveraging these principles, our approach bridges the strengths of LCA and IRT to develop a replicable, robust, and easy-to-implement method for cut-score determination, making the classification more apt for secondary analysis and interpretation of the results.

To demonstrate the applicability of our model, we utilize PIRLS 2021 data to validate the classifications on context scales and enhance classification accuracy, particularly for scales with skewed distributions across countries with diverse educational backgrounds. This data-driven approach strengthens examinee categorization by extending the application of LCA-based approaches to standard setting and proficiency scaling into new domains, supporting reliable, data-driven standard setting across different educational contexts. Overall, this study highlights the advantages of LCA-CS as a viable alternative or complementary method to traditional judgment-based approaches for determining cut scores on context scales.

Methods

The latent class model (e.g., Lazarsfeld & Henry, 1968; von Davier & Lee, 2019) is a statistical technique for identifying latent subgroups within a population based on categorical observed variables. Suppose we observe J polytomous items *(j=1,2,...,J)* where each item has K_j *(k = 1,...,K_j)* response categories, and we observe responses for examinees *i* = 1,2,...,N. The observed responses to these variables are denoted as X_{ijk}, where X_{ijk} = 1 if examinee i selects the k-th response category to the j-th item, and 0 otherwise. The latent class model assumes that the observed joint distribution of the manifest variables can be expressed as a weighted sum of conditional distributions in C latent classes. Each class represents a cross-classification table of response probabilities, parameterized by π_{jck}, the probability of selecting the *k*-th response to the *j*-th item in class *c*. For each variable *j*, $\sum_{k=1}^{K_j} \pi_{jck}$ = 1. The weights p_c, referred to as the mixing proportions, represent the prior probabilities of class membership satisfying $\sum_{c=1}^{C} p_c$ =1.

A key assumption in LCA is conditional independence, meaning that the observed variables are independent of one another, given membership in a latent class. This assumption, analogous to the local independence property in IRT, allows the model to decompose the observed joint distribution of responses into class-conditional probabilities (Yamamato, 1987). The model is fully identified by the matrix of conditional probabilities, π_{jck}, and the class distribution, p_c which together parameterize the probability of observed responses.

Under conditional independence, the probability of observing a specific set of responses for an individual *i* in a class *c* is given by:

$$f(X_i;\pi_c) = \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jck})^{X_{ijk}}$$

The probability of the observed responses across all classes is then

$$P(X_i|\pi,p) = \sum_{c=1}^{C} p_c \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jck})^{X_{ijk}}$$

The parameters of the model p_c and π_{jck} are estimated by maximizing the log-likelihood function:

$$lnL = \sum_{i=1}^{N} ln\left(\sum_{c=1}^{C} p_c \prod_{j=1}^{J}\prod_{k=1}^{K_j}(\pi_{jck})^{X_{ijk}}\right)$$

Posterior probabilities for class membership are computed using Bayes' rule:

$$P(c|X_i) = \frac{p_c f(X_i;\pi_c)}{\sum_{q=1}^{C} p_q f(X_i;\pi_c)}$$

where *c* = 1,2,...,C.

Latent classes are ordered if there is a permutation $\eta(c)$ of the class membership variable C so that the expected responses of all items *j* are ordered across classes. That is,

$$\sum_{k=1}^{K_j} k\pi_{j[\eta(c)]k} < \sum_{k=1}^{K_j} k\pi_{j[\eta(c)+1]k} \quad \text{for all } j = 1,..,J$$

This ensures that an ordered or continuous latent trait that leads to equivalent conditional probabilities can be identified. To test this, the classes are ordered by their expected sum score, i.e., the expected score is increasing with (reordered) class index. Then, the same property, the monotonicity of the expected scores, is checked for each item on the scale (Rost, 1990).

LCA for Identifying Cut Points

The proposed approach uses the latent class model to identify cut points on a scale from the response data. It first uses LCA to define a categorical latent variable that explains differences in item scores based on score distribution differences between homogeneous groups (latent classes). Next, a series of calculations are needed to identify cut points on the context scale. The details of these steps are described below. The following descriptions are based on three classes for simplicity and clarity, though the procedure generalizes to any number of classes.

1. Run latent class analysis (LCA) with a pre-specified number of classes. This number is usually identified based on literature or by context experts. In large-scale assessments such as TIMSS & PIRLS, the goal is to define cutpoints for three groups with high, medium, and low expected scores on the context scales.

2. Assign test takers to classes based on the posterior probability $P(C = c|X_1...X_J)$ of being a member of class c given responses $X_1...X_j$ to a set of items. Each test taker is assigned to the class based on the maximum posterior probability among the specified classes.

3. Re-order classes so that the expected score increases with the class index. That is, *E(score|C=c) > E(score|C=c+1)* if class *c = 1* represents the class with higher scores, where *E(score|C=c)* is the expected score given class c. Meanwhile, check whether the expected scores of each item are in the same order as the ordered classes.

4. Calculate the probability of a score given a class, *P(score|C)*. This probability is approximated assuming that each class is a

homogeneous group with a conditional normal ability distribution, $N(\mu_c,\sigma_c)$, where μ_c and σ_c are the mean and standard deviation of scores within the class. The result is an approximate conditional probability distribution, the probability of a score given a class, *P(score|C)*.

5. Calculate the conditional probability approximation of a "class" given a score using Bayes' theorem. Standard results yield,

$$P(C|score) = \frac{P(score|C)P(C)}{P(score)}$$

where *P(score|C)* is obtained from step 4. *P(C)* is the class size, and *P(score)* is the marginal probability for each score point.

6. Identify the cut score points and connect them to the construct, either the raw points or the scale score. The cut points are identified by locating the intersection point of adjacent smoothed posterior probability distributions, obtained from step 5, so that *P(C=c|cut point) > P(C=c+1|cut point)* and *P(C=c+1|cut point-1>) P(C=c|cut point-1)*, if class *c* = 1 represents the class with higher scores.

7. Classify the respondents into one of the three regions based on the identified cut points. Once the cut points are determined using this method, the subsequent procedures of assigning respondents to categories mirror those of the judgment-based cut point specification method or other methods.

For reporting or interpretation of the regions divided by these cut points, the minimum responses needed to meet or exceed the cut scores could be determined by calculating the expected responses for each item based on the IRT model and estimated item parameters. This involves selecting the most likely response for each item given the associated scale cut score, starting with the response category with the highest probability across all items, then moving to the next highest probability on another item until the total raw scores of expected responses are achieved to have the same values as the identified raw cut scores. Note that any response pattern that matches the raw score associated with the scale cut score is compatible with this approach if the scale score is derived using Rasch IRT model, just as in the judgement-based approach.

Application of the LCA-CS Method for Creating Scale Regions

PIRLS and Context Scales Reporting

This section describes applying the approach to define scale regions using data from PIRLS 2021. PIRLS is designed to measure reading achievement at the fourth-grade level and school and teacher practices related to reading instruction. Students complete a reading assessment and a questionnaire asking about their attitudes toward reading and reading habits. In addition, parents, teachers, and school principals are given questionnaires to gather information about students' home and school experiences in developing reading literacy. Since 2001, PIRLS has provided high-quality data for monitoring progress in students' reading achievement in their fourth year of schooling and measuring trends in achievement over time, covering 20 years of trends.

In PIRLS 2021, the fifth assessment cycle, 57 countries and 8 benchmarking entities participated. All students were administered the same questionnaires after the achievement booklet administration. PIRLS 2021 collected data from approximately 400,000 students, their parents, teachers, and school principals (Mullis et al. 2023). The PIRLS context questionnaire included several item sets intended to measure a latent construct. These constructs included the availability of home resources for learning, participation in literacy and numeracy activities in the home, the school's emphasis on academic success, students' attitudes about learning, and many others. In total, 22 context scales were derived from the PIRLS 2021 data collected from students, their parents, teachers, or principals using the Rasch partial credit model (PCM; Masters, 1982; Masters & Wright, 1997). The estimated Rasch scale scores were converted into a (10, 2) reporting metric for each scale, based on the countries included in the calibration (Yin & Reynolds, 2023). The reporting metric of the scale is set during the PIRLS cycle when the scale is first used or if a scale was revised by adding or changing items or revising response options.

Respondents were classified into three regions corresponding to high, middle, and low values on the construct to facilitate interpretation of the context scale results. The cut scores on the scale delimiting the regions were described in terms of combinations of response categories, the score combinations needed to reach medium or high score regions were defined based on review by content experts. Details on this procedure can be found in Yin & Reynolds (2023).

Once the raw cut points were identified, the corresponding scale cut scores were located utilizing the fact that the raw score is a sufficient statistic in the Rasch model (Andersen, 1977). This conversion was done assuming all questions in the set were answered. This judgment-based method works well under certain conditions, and the scale is well-centered and has sufficient variance along the range of possible scores. However, when the item responses are highly skewed across countries, the content-referenced cut-score definitions might produce score regions that do not contain students for some reporting groups, or even in some countries. The classification is not very useful, if, for example, only 'medium' and 'high' groups are populated, but no students are assigned to the 'low' group. For analytic purposes, such a case would reduce the reporting to only two groups.

In these cases, the proposed LCA-based approach can improve the situation. In the example of the PIRLS 2021 data, the proposed LCA-CS method validates existing categorizations on the context scales and enhances classification accuracy, particularly for scales exhibiting highly skewed distributions across diverse countries.

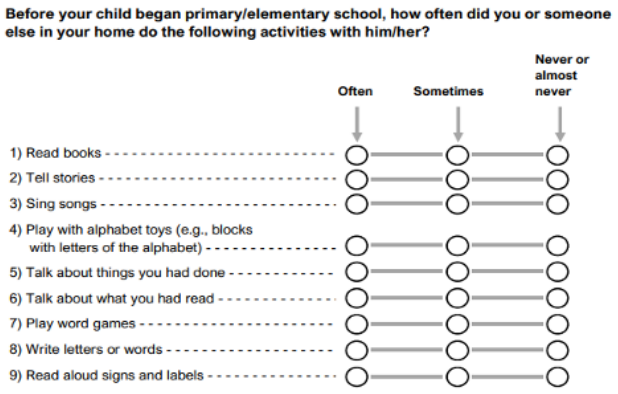
Description of Example Scale

This study uses the "Home Early Literacy Activities Before Primary School" scale as an example to demonstrate the LCA-CS method for specifying cut points.

The Home Early Literacy Activities scale was initially developed in PIRLS 2011 and has been continued for subsequent cycles. It includes nine component items from parents' questionnaires, focusing on how often parents engage their children in early literacy activities, as listed in Table 1 (Mullis et al., 2023). All 9 questions have three response options, "Often', "Sometimes", and "Never or almost never", with assigned numeric values of 2, 1, and 0 to the corresponding response categories. Therefore, the maximum available total raw points of this scale were 18.

Table 1:

Questions Included in PIRLS 2021 Home Early Literacy Activities Before Primary School Scale



The distribution of this scale is highly skewed, with almost no respondents falling into the low category for most countries when using cut scores provided by content experts. The categorization was based on scale cut scores of 10.7 and 6.2, derived from raw cut points of 14 and 4 based on minimal response profiles provided by content experts described earlier.

Applying the LCA-CS Method

To apply the proposed LCA-CS method for identifying the raw cut points, the SAS procedure PROC LCA (Lanza et al., 2015), one specialized function designed for latent class analysis in SAS program, was used for estimating the latent class model. The LCA was based on the combined data from all 40 calibration countries (Yin & Reynolds, 2023), countries that administered the assessment as scheduled at the end of the 4th school year, with complete responses to the 9 items. A total of 171,796 respondents were included in the LCA model, estimated assuming three classes to align with the reporting goals for PIRLS 2021 international results. The NSTARTS value in PROC LCA was set as 20 to find the best estimates and avoid local maxima of the likelihood function when conducting the analysis.

The posterior probability of the three classes for each respondent is part of the derived statistics that can be obtained through the SAS LCA procedure. Next, the rest of the steps from the previous section were applied. Table 2 shows the results after step 5, the re-calculated conditional probability approximations of the three classes given a score, *P(C|score)*. In the table, class 1 represents the class with the highest expected score, while class 3 represents the class with the lowest expected score. The left two columns are raw possible total points of complete responses of nine items and the associated unique transformed Rasch scale scores, which were retrieved from Appendix 15B in the PIRLS 2021 context scaling chapter (Yin & Reynolds, 2023). The last three columns are the conditional probability approximations, or smoothed posterior probabilities, for the three classes.

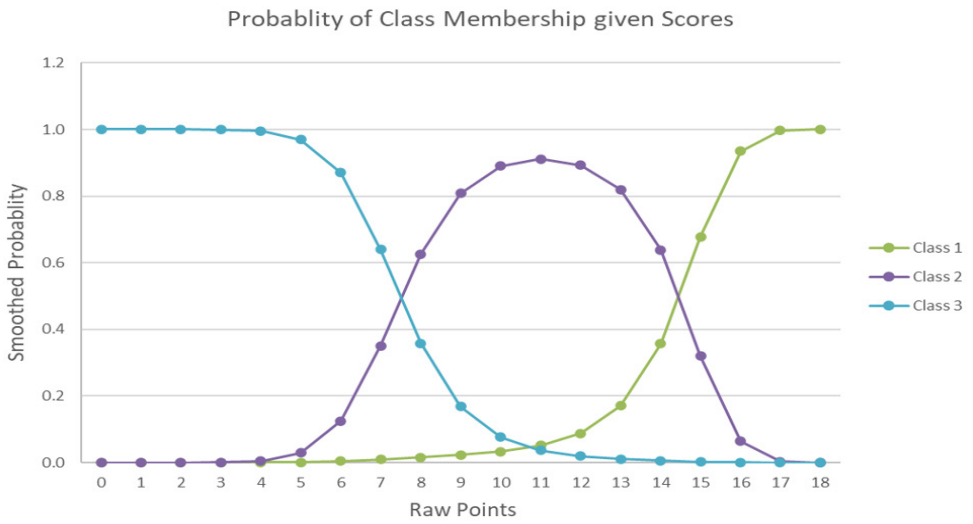
Table 2:

Conditional Probability Approximations of Classes given a Raw Score Point

Raw Points	Scale Score	Number of respondents	Smoothed Conditional Probability		
			Class 1	Class 2	Class 3
0	2.0717	511	0.00	0.00	1.00
1	3.9169	402	0.00	0.00	1.00
2	4.8778	698	0.00	0.00	1.00
3	5.5848	987	0.00	0.00	1.00
4	6.1700	1566	0.00	0.00	1.00
5	6.6863	2470	0.00	0.03	0.97
6	7.1652	3537	0.00	0.13	0.87
7	7.6184	5215	0.01	0.35	0.64
8	8.0567	7310	0.02	0.63	0.36
9	8.4885	12000	0.02	0.81	0.17
10	8.9179	12752	0.03	0.89	0.08
11	9.3525	15367	0.05	0.91	0.04
12	9.7989	18101	0.09	0.89	0.02
13	10.2674	19713	0.17	0.82	0.01
14	10.7707	19484	0.36	0.64	0.01
15	11.3376	17865	0.68	0.32	0.00
16	12.0220	14082	0.94	0.06	0.00
17	12.9578	9721	1.00	0.00	0.00
18	14.7746	10015	1.00	0.00	0.00

Figure 1 displays the smoothed posterior probability distribution for each class. Cut points were identified by locating the intersections of adjacent probability distributions and connecting them to the construct. From Figure 1, the intersections occur between 7 and 8 for classes 2 and 3, and between 14 and 15 for classes 1 and 2. To align with the judgment-based raw cut points approach using whole numbers, 8 and 15 were chosen as the raw cut points.

Figure 1:
Plot of the Conditional Probability Approximations of Classes given a Raw Score Point



Once the raw cut points were determined, the subsequent procedures of assigning respondents to categories mirror those of the judgment-based cut point specification method described in creating the PIRLS 2021 context scales chapter (Yin & Reynolds, 2023). According to the equivalence table of the raw scores and transformed scale scores presented in Table 2, the corresponding scale scores are 8.0567 and 11.3376 for raw points 8 and 15, respectively. Following the same rounding rules as the judgment-based cut point specification methods, the rounded scale scores, 8.1 (rounded up) and 11.3 (rounded down), were the final scale cut scores. These two cut scores were then used to classify all the respondents into one of three regions, including those from the countries with delayed administrations due to pandemic-related delays.

Categorization Results Using the LCA-CS Method

The following section presents the categorization results applied to the Home Early Literacy Activities scale using the LCA-CS method to identify the cut scores.

Table 3 shows the percentage of students whose parents were classified into each of the three regions using two different categorization methods. The standard errors (SEs) associated with the percentages, except for the percentage of 2 or smaller, are listed in parentheses. This table reports the results based on all PIRLS 2021 countries with comparable data, including those not included in the LCA model and item calibrations. The rightmost column shows each country's average scale score and associated SE. The results in the left part of Table 3 are the PIRLS 2021 published results (Mullis et al., 2023), showing percentages derived from conventional methods reliant on human judgments to define raw cut points based on item content. In contrast, the percentages for the three regions in the right part of the table were obtained using the LCA-CS procedures.

In Table 3, within the low region of the scale, there are many very small percentages, 2%, 1%, and even 0s, when using the judgment-based categorization. In practice, reporting the achievement levels for such a small percentage of students in a region is associated with a large error, and PIRLS does not report groups smaller than 2% in size. Therefore, the results from this categorization provided limited value for interpreting the relationship between achievement and home early literacy activities.

In contrast, using the LCA-CS procedures, the distribution of percentages across the three regions is less skewed across countries, enhancing the interpretation of the achievement and the related context. Based on the categorical latent variable modeling technique, the low category is no longer empty for all countries, which identifies groups based on a statistically optimal criterion. Additionally, the percentages in the middle region closely align with those from the judgment-based approach at the country level and internationally. This supports the existing categorizations on the context scales for the middle region, indicating that most respondents are likely in the "Medium" region of the scale. Overall, the categorization based on this method provides more value for interpreting home early literacy activities with students' reading achievement.

Discussion

With growing interest in understanding how learning contexts relate to student achievement, many items in large-scale assessment questionnaires are designed to measure a common underlying context construct linked to achievement. For interpretation, respondents

Table 3:

Percent of Students in Each Region of Home Early Literacy Activities Scale Using Two Categorization Methods

Country	Percent of Students (Judgment-based Method)			Percent of Students (LCA-CS Method)			Average Scale Score
	High	Medium	Low	High	Medium	Low	
Kazakhstan	66 (0.9)	34 (0.9)	0 ~	52 (0.9)	44 (0.8)	3 (0.5)	11.3 (0.04)
Russian Federation	64 (1.3)	35 (1.2)	1 ~	51 (1.3)	44 (1.0)	6 (0.7)	11.3 (0.07)
Northern Ireland s	64 (0.9)	35 (0.9)	1 ~	54 (1.0)	42 (1.0)	4 (0.4)	11.5 (0.04)
Georgia	59 (1.1)	40 (1.1)	1 ~	45 (1.0)	49 (1.0)	5 (0.6)	11.0 (0.05)
Croatia	58 (1.1)	42 (1.1)	0 ~	43 (1.1)	52 (1.1)	5 (0.5)	11.0 (0.05)
Malta r	57 (1.2)	42 (1.2)	0 ~	44 (1.1)	50 (1.1)	6 (0.6)	11.1 (0.05)
Albania	57 (1.5)	41 (1.4)	2 ~	44 (1.6)	47 (1.7)	9 (1.3)	10.9 (0.08)
Uzbekistan	57 (1.7)	43 (1.7)	0 ~	40 (1.6)	55 (1.4)	4 (0.5)	10.8 (0.06)
Ireland	56 (1.1)	43 (1.0)	1 ~	43 (1.1)	50 (0.9)	7 (0.6)	11.0 (0.05)
Kosovo	55 (1.3)	44 (1.3)	1 ~	40 (1.3)	55 (1.2)	5 (0.6)	10.8 (0.04)
Montenegro	55 (0.9)	45 (0.9)	0 ~	41 (0.8)	54 (0.7)	5 (0.4)	10.9 (0.03)
North Macedonia	55 (1.2)	43 (1.2)	2 ~	43 (1.2)	51 (1.2)	6 (1.1)	10.9 (0.09)
Serbia	54 (1.2)	46 (1.2)	0 ~	40 (1.3)	54 (1.0)	5 (1.0)	10.8 (0.05)
Poland	53 (0.9)	47 (1.0)	0 ~	40 (0.9)	54 (1.0)	6 (0.5)	10.8 (0.04)
Spain	52 (0.8)	47 (0.8)	1 ~	39 (0.8)	53 (0.8)	8 (0.4)	10.7 (0.03)
Italy	52 (0.9)	47 (0.9)	1 ~	39 (0.8)	54 (0.8)	7 (0.4)	10.7 (0.03)
Cyprus	51 (0.6)	48 (0.7)	1 ~	39 (0.7)	53 (0.7)	9 (0.5)	10.7 (0.03)
Slovak Republic	49 (1.1)	49 (1.2)	2 ~	36 (1.0)	54 (1.2)	10 (1.6)	10.5 (0.07)
Slovenia	49 (1.0)	51 (1.0)	1 ~	37 (0.9)	55 (0.8)	8 (0.6)	10.6 (0.04)
Latvia	48 (1.1)	51 (1.1)	1 ~	35 (0.9)	57 (1.0)	8 (0.5)	10.5 (0.04)
Israel s	47 (1.0)	52 (1.0)	1 ~	36 (1.0)	54 (0.9)	10 (0.7)	10.6 (0.04)
Hungary r	47 (1.0)	52 (1.0)	1 ~	31 (0.9)	61 (1.0)	8 (0.6)	10.5 (0.03)
Czech Republic	46 (0.8)	54 (0.8)	0 ~	33 (0.8)	60 (0.8)	7 (0.4)	10.5 (0.03)
United Arab Emirates s	42 (0.7)	56 (0.7)	2 ~	31 (0.6)	56 (0.5)	13 (0.4)	10.3 (0.03)
Bulgaria	41 (1.1)	50 (1.1)	9 (1.2)	30 (1.0)	50 (1.3)	20 (1.5)	9.9 (0.09)
France	41 (0.9)	57 (0.9)	2 ~	30 (0.9)	58 (0.9)	12 (0.6)	10.2 (0.04)
Denmark	41 (0.9)	58 (0.9)	1 ~	28 (0.9)	60 (0.9)	12 (0.6)	10.3 (0.04)
Germany s	40 (1.1)	59 (1.1)	1 ~	27 (1.0)	64 (1.1)	9 (0.6)	10.3 (0.04)
Norway (5)	39 (0.7)	59 (0.7)	1 ~	28 (0.7)	61 (0.7)	11 (0.5)	10.2 (0.03)
Saudi Arabia r	39 (1.0)	58 (1.1)	3 (0.4)	29 (0.9)	58 (1.1)	13 (0.7)	10.2 (0.05)
South Africa r	38 (0.9)	58 (0.8)	4 (0.5)	28 (0.8)	56 (0.9)	16 (0.8)	10.1 (0.05)
Bahrain	38 (0.7)	60 (0.7)	2 ~	26 (0.9)	60 (0.9)	14 (0.6)	10.1 (0.03)
Sweden s	38 (1.1)	61 (1.1)	1 ~	27 (0.9)	59 (1.0)	13 (0.9)	10.2 (0.04)
Austria	37 (0.9)	61 (0.9)	1 ~	25 (0.8)	62 (1.1)	13 (0.8)	10.1 (0.04)
Portugal	37 (0.9)	62 (0.9)	1 ~	25 (0.8)	62 (0.7)	12 (0.5)	10.1 (0.03)
Azerbaijan	36 (1.0)	62 (1.0)	2 ~	23 (0.9)	63 (1.0)	14 (0.8)	10.1 (0.05)
Singapore	35 (0.8)	62 (0.8)	4 (0.3)	26 (0.7)	54 (0.7)	19 (0.6)	10.0 (0.04)
Oman	34 (1.0)	65 (1.0)	2 ~	21 (0.9)	67 (1.0)	11 (0.7)	10.0 (0.04)
Qatar r	33 (1.0)	65 (1.0)	2 ~	23 (1.0)	63 (0.9)	14 (0.8)	9.9 (0.04)
Finland	33 (0.7)	66 (0.7)	1 ~	23 (0.7)	65 (0.9)	12 (0.5)	10.0 (0.02)
Turkiye	31 (1.1)	57 (1.2)	13 (1.6)	22 (1.0)	52 (1.4)	27 (1.8)	9.3 (0.12)
Belgium (French) r	30 (1.0)	67 (1.0)	2 ~	20 (0.9)	63 (1.1)	17 (0.8)	9.8 (0.04)
Brazil	30 (1.0)	63 (1.2)	7 (0.9)	21 (0.9)	55 (1.3)	24 (1.1)	9.6 (0.06)
Jordan	29 (1.0)	66 (0.9)	5 (0.6)	19 (0.8)	61 (1.1)	20 (1.2)	9.6 (0.06)
Belgium (Flemish)	27 (0.8)	71 (0.9)	2 ~	17 (0.7)	62 (0.8)	21 (0.9)	9.6 (0.04)
Egypt	27 (1.3)	67 (1.3)	7 (0.7)	17 (1.1)	60 (1.2)	23 (1.3)	9.4 (0.07)
Iran, Islamic Rep. of	24 (1.1)	71 (1.2)	5 (0.9)	15 (0.9)	61 (1.2)	24 (1.2)	9.4 (0.07)
Chinese Taipei	18 (0.5)	76 (0.6)	6 (0.4)	12 (0.4)	60 (0.7)	28 (0.7)	9.1 (0.03)
Hong Kong SAR	16 (0.8)	81 (0.8)	3 (0.3)	10 (0.7)	66 (0.8)	24 (0.8)	9.2 (0.04)
Morocco	13 (0.7)	67 (1.4)	19 (1.6)	8 (0.5)	49 (1.5)	44 (1.7)	8.2 (0.10)
Macao SAR	10 (0.4)	85 (0.4)	5 (0.3)	6 (0.3)	58 (0.8)	36 (0.7)	8.7 (0.02)
International Average	**42 (0.1)**	**55 (0.1)**	**3 (0.1)**	**31 (0.1)**	**56 (0.1)**	**13 (0.1)**	
New Zealand x	59 (1.1)	40 (1.1)	1 ~	49 (1.1)	45 (1.0)	7 (0.5)	11.2 (0.05)
Netherlands x	39 (1.3)	60 (1.4)	1 ~	27 (1.2)	62 (1.2)	11 (0.8)	10.2 (0.05)

An "r" indicates data are available for at least 70% but less than 85% of the students.
An "s" indicates data are available for at least 50% but less than 70% of the students.
An "x" indicates data are available for at least 40% but less than 50% of the students—interpret with caution.
A tilde (~) indicates insufficient data to report result. A dash (-) indicates comparable data not available.

are classified into high, middle, and low regions utilizing specified cut-points on the context scale. The achievement in each group is then reported. This enables the relationship between achievement and the context to be observed across diverse groups. Conventional methods rely on expert judgments to define cut points based on item content, which works well with balanced response distributions. However, when the item responses are highly skewed across diverse groups or populations, these content-referenced cut-score definitions likely produce regions with few or no respondents, limiting the interpretation of the achievement and context relationship, as illustrated in Table 3.

The proposed LCA-CS method addresses these challenges by leveraging LCA to calculate the posterior probability of class membership for a pre-specified number of classes for each respondent with complete responses. With the assumption that each class is a homogeneous group with a conditional normal ability distribution, the conditional probability approximations of class membership are obtained by a series of calculations, as illustrated in the previous sections. These conditional probabilities of a class membership given a score provide the basis for finding the cut scores on the constructed context scale to apply to all respondents with a valid scale score. As demonstrated by applying the method to the PIRLS 2021 Home Learning Activity data, the proposed LCA-CS method statistically optimized the distribution of students across categories and enhanced the adequacy of categorization. This implies that this data-driven LCA-CS method could serve as an improved approach for identifying cut scores for educational researchers or practitioners, especially when the responses are highly skewed across diverse groups.

Our study aligns with the growing body of literature emphasizing the importance of incorporating statistical modeling techniques into educational assessment to enhance the validity of classification decisions (e.g., Brown, 2007; Templin & Jiao, 2012; Binici & Cuhadar, 2022). While both Brown (2007) and Binici and Cuhadar (2022) focused on the application of LCA-based method to achievement data demonstrating its utility as an empirical, data-driven alternative to judgement-based methods for classifying examinees, our research extends the application of LCA-based method to contextual data. In this domain, where response distributions are often skewed across diverse groups, LCA-based classifications can improve the adequacy of categorization. Furthermore, our findings resonate with those of Binici and Cuhadar (2022), who demonstrated that LCA-based methods can validate performance standards derived from traditional judgment-based approaches. Similarly, in the context of our study, the LCA-based method proved effective for validating existing judgement-based categorizations on the context scales.

In conclusion, the LCA-CS method offers a promising, statistically sound alternative for defining cut scores on context scales in large-scale assessments. By addressing the limitations of traditional methods and optimizing the distribution of respondents across categories, this approach provides meaningful insights into the relationship between learning contexts and achievement. The LCA-CS method, as introduced in this study, utilized scales derived from a Rasch model with a pre-specified number of classes provided by analytic goals. When this required number of classes is unavailable, the LCA method can be used to determine the optimal number of classes based on model fit statistics and practical needs. This study introduced the LCA-CS method and demonstrated its implementation with real data from a large-scale assessment. Future studies should focus on developing diagnostics to evaluate the effectiveness of the LCA-CS method compared to judgment-based cut points. In addition, future research could extend this approach to scales based on more general IRT models, such as the Generalized Partial Credit Model, using similar procedures.

References

Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika* 42, 69–81. https://link.springer.com/content/pdf/10.1007/BF02293746.pdf

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Binici, S., & Cuhadar, I. (2022). Validating performance standards via latent class analysis. *Journal of Educational Measurement*, 59(4), 502-516.

Brown, R. S. (2007). Using latent class analysis to set academic performance standards. *Educational Assessment*, 12(3-4), 283-301.

Cizek, G. J. (1993). Reconsidering standards and criteria. *Journal of Educational Measurement*, 30(2), 93–106. https://doi.org/10.1111/j.1745-3984.1993.tb01068.x

Cizek, G. J. (2005). EMW. *Defending Standardized Testing*, 23.

Cizek, G. J. (Ed.). (2012). *Setting performance standards: Foundations, methods, and innovations*. Routledge.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. SAGE Publications Ltd.

Clogg, C. C. (1979). Some latent structure models for the analysis of Likert-type data. *Social Science Research*, 8(4), 287-301.

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

Croon, M. (2002). Ordering the classes. *Applied latent class analysis*, 137-162.

Dayton, C. M., & Macready, G. B. (1976). A probabilistic model for validation of behavioral hierarchies. *Psychometrika*, 41(2), 189–204. https://doi.org/10.1007/BF02291838

Dayton, C. M., & Macready, G. B. (2006). 13 Latent Class Analysis in Psychometrics. *Handbook of statistics*, 26, 421-446.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Formann, A. K. (1992). Linear logistic latent class analysis for polytomous data. *Journal of the American Statistical Association*, 87(418), 476-486.

Heinen, T. (1993). *Discrete Latent Variable Models*. Tilburg: University Press.

Heinen, T. (1996). *Latent class and discrete latent trait models: Similarities and differences. Sage Publications*, Inc.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (pp. 485–514). New York: Macmillan.

Jiao, H., Lissitz, R. W., Macready, G., Wang, S. & Liang, S. (2011). Exploring levels of performance using the mixture Rasch model for standard setting. *Psychological Test and Assessment Modeling*, 53(4), 499-522.

Lanza, S. T., Dziak, J. J., Huang, L., Wagner, A. T., & Collins, L. M. (2015). *Proc LCA & Proc LTA users' guide* (Version 1.3.2). University Park: The Methodology Center, Penn State. Available from methodology.psu.edu.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. New York, NY: Houghton Mifflin

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, 2(2), 121–141.

Masters, G. N. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, 47(2), 149-174.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory*. Berlin: Springer.

Mislevy, R. J. (1983). Item response models for grouped data. *Journal of Educational Statistics*, 8(4), 271-288.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2013). The bookmark procedure: Psychological perspectives. In *Setting Performance Standards* (pp. 263-296). Routledge.

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 International Results in Reading*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

Peabody, M. R., Muckle, T. J., & Meng, Y. (2023). Applying a Mixture Rasch Model-Based Approach to Standard Setting. *Educational Measurement: Issues and Practice*, *42*(3), 5-12.

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement*, *14*(3), 271-282.

Rost, J., & von Davier, M. (1995). Polytomous mixed Rasch models. In G. H. Fischer & I. W. Molenaar (Eds.), Rasch models – Foundations, Recent Developments, and Applications (pp. 371–379). Springer. https://doi.org/10.1007/978-1-4612-4230-7_20

Smit, J. A., Kelderman, H., & van der Flier, H. (2003). Latent trait latent class analysis of an Eysenck Personality Questionnaire. *Methods of Psychological Research Online*, 8(3), 23-50.

Templin, J., & Jiao, H. (2012). Applying model-based approaches to identify performance categories. In *Setting performance standards* (pp. 379-397). Routledge.

von Davier, M., & Lee, Y.-S. (2019). Introduction: From latent classes to cognitive diagnostic models. In Handbook of Diagnostic Classification Models (pp. 1-17). Springer. https://doi.org/10.1007/978-3-030-05584-4_1

Yamamoto, K. (1987). *A model that combines IRT and latent class models*. University of Illinois at Urbana-Champaign.

Yin, L., & Reynolds, K. A. (2023). Creating and interpreting the PIRLS 2021 context questionnaire scales. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and Procedures: PIRLS 2021 Technical Report* (pp. 15.1-15.161). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb6994