Psychometrika (2026), **0**, 1–19
doi:10.1017/psy.2026.10149

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Sparse Latent Class Analysis for Dichotomous Responses: Post-Estimation Refinement via Item-Level Pseudo-Likelihood

Yuxuan Xu[1], Lea Kaufmann[2], Yunxiao Chen[1], Maria Kateri[2] and Irini Moustaki[1]

[1]Department of Statistics, The London School of Economics and Political Science, UK; [2]RWTH Aachen University, Germany

Corresponding author: Yunxiao Chen; Email: y.chen186@lse.ac.uk

(Received 18 May 2026 UTC; revised 3 September 2026 UTC; accepted 9 September 2026 UTC)

Abstract

Latent class analysis (LCA) is widely used to identify unobserved subgroups in social and behavioral sciences. A long-standing challenge for LCA is the interpretability of the latent classes, due to the high complexity of the estimated item-response probability matrix. To address this, we propose a computationally efficient post-estimation refinement procedure that enhances model interpretability by a sparse model estimate. The method begins by estimating a classical, unrestricted, latent class model and determining the number of classes using the Bayesian information criterion. It is followed by a refinement step that further performs model selection on the item-specific response probabilities based on the initial estimate. This refinement penalizes the number of distinct response probability levels per item, collapsing redundant levels to yield a sparse matrix that is significantly easier to interpret than those produced by classical LCA. We provide asymptotic theory showing that the proposed procedure consistently recovers the sparse pattern of the item-response probabilities for each item and further validate its performance through extensive simulations. The practical power of the proposed method is further illustrated via an application to survey data on social role performance, where it provides a parsimonious and clear characterization of the resulting latent classes.

Keywords: latent class model; model interpretability; post-estimation refinement; pseudo-likelihood; sparsity

1. Introduction

Rooted in the seminal contributions of Lazarsfeld and Henry (1968) and Goodman (1974a, 1974b), latent class analysis (LCA) has emerged as a cornerstone for exploratory data analysis, enabling researchers to identify and characterize unobserved subgroups within heterogeneous populations. This versatile framework has found widespread applications across the social and behavioral sciences, public health, and marketing research; for a rigorous treatment of its inferential foundations and broad empirical utility, we refer the reader to Collins and Lanza (2010). Beyond these roots in exploratory data analysis, latent class models are increasingly employed to assess fine-grained psychological and educational attributes, a field commonly referred to as cognitive diagnosis or diagnostic classification (see Rupp et al., 2010 and von Davier & Lee, 2019 for comprehensive reviews). These models typically impose restrictions on the parameter space of standard latent class models and are therefore often known as restricted latent class models (Gu & Xu, 2020). Notwithstanding

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

these measurement-focused developments, the current article remains situated within the exploratory tradition of LCA.

Assigning substantive meaning to latent classes is central to exploratory LCA. This step is typically done by examining the item-response probability matrix. These parameters essentially define the "profile" of each subgroup; yet, their interpretation is often non-trivial. With many latent classes or items, identifying distinct, interpretable patterns within the matrix becomes difficult. This often yields "noisy" classes that are difficult to distinguish, creating a risk that subgroup labeling becomes inaccurate or overly subjective. To address interpretability challenges, recent advances have focused on inducing sparsity in the item-response matrix. Chen et al. (2017) proposed a regularized estimation approach for the LCA of binary response data. This approach induces sparsity in the item-response probability matrix, thereby improving the interpretability of the resulting latent classes. This approach has been further extended to polytomous response data in Robitzsch (2020). However, this regularized estimation approach requires maximizing a regularized log-likelihood function that includes a nonconvex, nonsmooth penalty term. The optimization is typically carried out via an expectation–maximization (EM) algorithm, for which each M-step requires solving multiple nonconvex, nonsmooth optimization problems. Such a heavy computational burden may render the method impractical for many real-world applications. Furthermore, within exploratory data analysis settings, sparsity assumptions and regularized estimation procedures have been applied to diagnostic classification models to enhance the interpretability of latent classes (see, e.g., Chen et al., 2015; Chen, Culpepper, et al., 2020; He et al., 2025; Wang & Lu, 2021). Recent work has also established identification results for this type of sparse structure. Bowers and Culpepper (2024) proposed the equivalence set restricted latent class model (ESRLCM), which induces the sparse structure in the item-response probability matrix through equivalence-set restrictions, and established sufficient conditions for generic identifiability based on whether the item-response structure can adequately distinguish the latent classes.

In this article, we propose a post-estimation refinement procedure that achieves the interpretative benefits of sparsity without the prohibitive computational costs of regularized estimation. Our approach operates in two stages: we first estimate a classical LCA model and determine the number of classes via the Bayesian information criterion (BIC; Schwarz, 1978), and then apply a refinement step that yields a sparse item-response probability matrix with clearer patterns, thereby making the latent classes significantly easier to interpret. This refinement procedure has two building blocks: (1) a stepwise search algorithm for exploring sparse models and (2) an extended BIC (EBIC) for comparing models with different sparsity levels, both of which rely on a novel item-specific pseudo-likelihood function constructed from the initial model estimate. We provide asymptotic theory confirming that our procedure consistently selects the number of distinct item-response probability levels. Furthermore, we demonstrate the power of this method through extensive simulations and an application to survey data on social role performance, where it clarifies respondents' complex profiles.

The item-specific pseudo-likelihood function mentioned above plays a central role in the proposed refinement procedure. This function is constructed from the posterior probabilities of observations' latent class membership, computed under the initial model estimate in the first stage. Since the initial model estimate is already root-N consistent (Collins & Lanza, 2010), these posterior probabilities are sufficiently close to those that we would obtain under the true model. Consequently, the resulting EBIC enjoys similar asymptotic properties as the classical BIC and thus yields statistical consistency in model selection. Similar pseudo-likelihood functions have been used in stepwise estimation methods for latent class models, such as the ones developed to model a distal outcome predicted by a latent categorical variable (Zhu et al., 2017). However, existing work focuses on parameter estimation, whereas, to our best knowledge, the proposed method is the first to use such a pseudo-likelihood for model selection. Further discussions about stepwise methods for latent class models can be found in Zhu et al. (2017), Bakk and Kuha (2018), and Bakk and Kuha (2021).

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

The rest of the article is organized as follows. Section 2 introduces the proposed refinement method and establishes its theoretical guarantees. Section 3 presents a comprehensive simulation study to evaluate the method's performance. Section 4 illustrates the practical utility of the approach by applying it to data from a health survey. Finally, Section 5 presents a discussion of the results and concludes.

2. The proposed method

2.1. Latent class analysis

We consider N respondents answering J items. Let Y_{ij} denote the response from respondent i to item j. To simplify the discussion, we assume that the responses are binary, that is, $Y_{ij} \in \{0, 1\}$. As discussed in Section 5, the proposed method can be extended to data with polytomous items.

Suppose the data follow a latent class model with K classes. Let $\xi_i \in \{1, 2, \ldots, K\}$ indicate the latent class membership of respondent i. The model assumes that ξ_i, $i = 1, \ldots, N$, are independent and identically distributed (i.i.d.) random variables, following a categorical distribution, with $P(\xi_i = k) = \nu_k$, where $\nu_1, \ldots, \nu_K$ are unknown parameters satisfying $\nu_k \geq 0$ and $\nu_1 + \cdots + \nu_K = 1$. Moreover, the model assumes that an individual's response probabilities depend only on their latent class membership. That is,

$$P(Y_{ij} = 1 | \xi_i = k) = \beta_{jk},$$

where β_{jk}, $j = 1, \ldots, J$ and $k = 1, \ldots, K$, are unknown parameters. Finally, the local independence assumption is adopted, which says that for each respondent i, $Y_{i1}, \ldots, Y_{iJ}$ are conditionally independent given ξ_i.

For the parsimony of notation, we use $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)^\top$, $\mathbf{B} = (\beta_{jk})_{J\times K}$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jK})^\top$ to denote the parameter vectors or matrix, and $\boldsymbol{\nu}^* = (\nu_1^*, \ldots, \nu_K^*)^\top$, $\mathbf{B}^* = (\beta_{jk}^*)_{J\times K}$, and $\boldsymbol{\beta}_j^* = (\beta_{j1}^*, \ldots, \beta_{jK}^*)^\top$ the corresponding true values. Under the assumptions of the latent class model, the marginal log-likelihood function takes the form

$$l(\boldsymbol{\nu}, \mathbf{B}) = \sum_{i=1}^{N} \log\Big(\sum_{k=1}^{K} \nu_k \prod_{j=1}^{J} \beta_{jk}^{Y_{ij}} (1 - \beta_{jk})^{1-Y_{ij}} \Big).$$

Classical LCA estimates the unknown model parameters by the maximum likelihood estimator (MLE), that is,

$$(\hat{\boldsymbol{\nu}}, \hat{\mathbf{B}}) = \arg\max_{(\boldsymbol{\nu}, \mathbf{B})} l(\boldsymbol{\nu}, \mathbf{B}). \tag{1}$$

This optimization is typically solved using an EM algorithm. Under correct model specification and mild regularity conditions, the MLE is root-N consistent, in the sense that

$$\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\| = O_p(1/\sqrt{N}) \text{ and } \|\hat{\mathbf{B}} - \mathbf{B}^*\|_F = O_p(1/\sqrt{N}),$$

where $\|\cdot\|$ denotes the Euclidean norm and $\|\cdot\|_F$ denotes the matrix Frobenius norm.[1]

When the number of latent classes K is unknown, the optimal number of latent classes is chosen by fitting models with increasing numbers of classes and selecting the one that minimizes the BIC value. We refer readers to Chapter 4 of Collins and Lanza (2010) for a review of the inferential foundations of the classical LCA. For ease of exposition, we treat the number of classes K as known when motivating and describing the proposed refinement method in Sections 2.2 and 2.3 below.

[1]There is a label indeterminacy issue, also known as the label switching problem, that is related to this consistency result. Here, we assume that the true and estimated labels have been consistently aligned, which can be achieved, for example, by assuming $\nu_1^* > \cdots > \nu_K^*$ and imposing the same ordering on $\hat{\nu}_1, \ldots, \hat{\nu}_K$.

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

Table 1. An example of the item-response probabilities for a sparse latent class model

	Latent class			
Item	1	2	3	4
1	0.10	0.10	0.10	0.90
2	0.20	0.20	0.80	0.80
3	0.15	0.75	0.15	0.75
4	0.30	0.30	0.30	0.85
5	0.25	0.25	0.90	0.90
6	0.20	0.55	0.55	0.95

2.2. Sparse LCA

Interpreting the latent classes is an essential step for LCA. This step is based on $\hat{\mathbf{B}}$, which characterizes the item-response behaviors of all the learned latent classes. However, when J or K is moderately large, interpreting the latent classes can be a challenge, due to the large number of parameters in $\hat{B}$. To tackle this challenge, Chen et al. (2017) proposed a sparse latent class model, which assumes that the number of distinct item-response probability levels for each item can be smaller than the total number of classes.

More specifically, let $m_j^* = \text{card}(\boldsymbol{\beta}_j^*)$ denote the true number of distinct item-response probability levels for item j. The sparse latent class model assumes that $m_j^* < K$ for many items. However, even when the true model is sparse, due to the randomness in data, each $\hat{\boldsymbol{\beta}}_j$ has K distinct values with probability tending to 1 as the sample size goes to infinity, and thus, cannot recover the sparse pattern in the true model. To tackle this issue, Chen et al. (2017) proposed a regularized estimation method that simultaneously learns the number of distinct item-response probability levels, denoted by $\tilde{m}_j$, and estimates the item-response probabilities, denoted by $\bar{\mathbf{B}} = (\bar{\beta}_{jk})_{J\times K}$, satisfying that the number of distinct levels in $(\bar{\beta}_{j1}, \ldots, \bar{\beta}_{jK})^\top$ is $\tilde{m}_j$, where $\tilde{m}_j$ can be strictly smaller than K. Consequently, $\bar{\mathbf{B}}$ tends to have substantially fewer parameters than those in $\hat{\mathbf{B}}$, the MLE for classical LCA, and thus, tends to be much easier to interpret. Consistency for both the model selection and parameter estimation has been established for this regularized estimator.

Despite the interpretational advantage and theoretical guarantees, this regularized estimator requires maximizing a regularized log-likelihood function, which involves complex EM cycles in which each M-step involves solving multiple non-smooth optimization problems. Such a heavy computational burden renders the method impractical for many real-world applications.

Example 1. Table 1 gives a toy example of the true item-response probability matrix $\mathbf{B}$ for a sparse latent class model, with $J = 6$ items and $K = 4$ latent classes. In this example, the elements of $\boldsymbol{\beta}_j$ have two distinct levels for items 1–5 and three distinct levels for item 6. If such a sparse pattern can be recovered from the data, one can infer that the first latent class has the lowest item-response probabilities across all items, while the fourth latent class has the highest item-response probabilities. In addition, each of the second and third classes has a combination of the lowest and highest item-response probabilities for items 1–5 and a median level of item-response probability for item 6. In a classical latent class model, however, each item would typically be assigned four distinct response-probability levels, even when the true underlying structure is sparse, making the resulting classes harder to interpret.

2.3. Proposed refinement procedure

We propose a refinement procedure to estimate a sparse latent class model, which remains statistically consistent while avoiding the high computational burden of the regularized estimator. Like the regularization approach in Chen et al. (2017), our goal is to find an estimate of the latent class model, denoted

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

by $(\bar{\boldsymbol{\nu}}, \bar{\mathbf{B}})$, such that $\bar{\mathbf{B}}$ is sparse, in the sense that each row of $\bar{\mathbf{B}}$ has a small number of distinct values, that is, each item has a small number of unique item-response probability levels.

The proposed estimator has two building blocks, including (1) a stepwise search algorithm for exploring sparse models and (2) an EBIC for comparing models with different sparsity levels, both of which rely on an item-specific pseudo-likelihood function constructed based on the initial model estimate from the first stage. In what follows, we first define the pseudo-likelihood, then describe the stepwise search algorithm, and finally the EBIC.

2.3.1. Item-specific pseudo-likelihood

Let

$$\hat{\gamma}_{ik} = \frac{\hat{\nu}_k \prod_{j=1}^{J} \hat{\beta}_{jk}^{Y_{ij}} (1-\hat{\beta}_{jk})^{1-Y_{ij}}}{\sum_{h=1}^{K} \hat{\nu}_h \prod_{j=1}^{J} \hat{\beta}_{jh}^{Y_{ij}} (1-\hat{\beta}_{jh})^{1-Y_{ij}}}$$

be the posterior probability of respondent i belonging to the kth latent class evaluated under the MLE for the classical LCA, and we denote $\hat{\boldsymbol{\Gamma}} = \{\hat{\gamma}_{ik}\}_{N\times K}$. The pseudo-likelihood function for item j is defined as

$$Q_j(\boldsymbol{\beta}_j; \hat{\boldsymbol{\Gamma}}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \hat{\gamma}_{ik} \left(Y_{ij} \log(\beta_{jk}) + (1 - Y_{ij}) \log(1-\beta_{jk}) \right). \tag{2}$$

This is a weighted Bernoulli log-likelihood with the latent class memberships replaced by the corresponding posterior probabilities $\hat{\gamma}_{ik}$; for brevity, we refer to it as the pseudo-likelihood throughout. We should also note that this function is the objective function for $\boldsymbol{\beta}_j$ in the M-step of the EM algorithm for solving (1).

This item-specific pseudo-likelihood enables independent model selection item by item, substantially reducing the complexity of the model selection problem.

2.3.2. Stepwise search algorithm

For each item j and each possible number of distinct item-response probability levels $m_j \in \{1, \ldots, K\}$, ideally, we hope to solve the following maximization problem:

$$\tilde{\boldsymbol{\beta}}_j^{(m_j)} = \arg\max_{\boldsymbol{\beta}_j} Q_j(\boldsymbol{\beta}_j; \hat{\boldsymbol{\Gamma}}), \quad \text{s.t. } \operatorname{card}(\boldsymbol{\beta}_j) = m_j. \tag{3}$$

Recall that $\operatorname{card}(\boldsymbol{\beta}_j)$ denotes the number of distinct values within the j-th item-response probability vector $\boldsymbol{\beta}_j$. In other words, $\tilde{\boldsymbol{\beta}}_j^{(m_j)}$ re-estimates the response probabilities for item j, under the constraint that the item has m_j distinct item-response probability levels. The solutions $\tilde{\boldsymbol{\beta}}_j^{(1)}, \ldots, \tilde{\boldsymbol{\beta}}_j^{(K)}$ are then compared using the EBIC introduced below. Unfortunately, (3) becomes computationally infeasible when K is sufficiently large, due to the combinatorial nature of the discrete constraint $\operatorname{card}(\boldsymbol{\beta}_j) = m_j$. For example, when $K = 10$, solving (3) for $m_j = 1, \ldots, 10$ requires comparing a total of 115,975 models, which is computationally demanding.

To tackle this computational challenge, we instead use a stepwise search algorithm that leverages the ordering of the initial estimate $\hat{\boldsymbol{\beta}}_j$. The algorithm iteratively identifies partitions of the entries of $\boldsymbol{\beta}_j$ such that the parameters in the same subset of the partition share the same value. More specifically, let $\mathcal{C} = \{C_1, \ldots, C_{m_j}\}$ be a partition of $\{1, \ldots, K\}$, satisfying that (1) each subset C_l is nonempty, (2) C_ls are mutually exclusive, and (3) the union of C_l is $\{1, \ldots, K\}$. A partition defines a constrained parameter space for $\boldsymbol{\beta}_j$:

$$\Theta_j(\mathcal{C}) = \{\boldsymbol{\beta}_j : \beta_{jk} = \beta_{jk'}, \forall\, k, k' \in C_l,\ l = 1, \ldots, m_j, \text{ and } \beta_{jk} < \beta_{jk'}, \text{ if } k \in C_l, k' \in C_{l'}, l < l'\},$$

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

satisfying $\text{card}(\boldsymbol{\beta}_j) = m_j$ for all $\boldsymbol{\beta}_j \in \Theta_j(\mathcal{C})$. As $\hat{\boldsymbol{\beta}}_j$ is a statistically consistent estimate of $\boldsymbol{\beta}_j^*$, the ordering of the estimated parameters in $\hat{\boldsymbol{\beta}}_j$ is useful when we search for the partition based on the true parameters $\boldsymbol{\beta}_j^*$.

For ease of exposition, we assume that $\hat{\boldsymbol{\beta}}_j$ has K distinct entries, which is typically the case in practice and asymptotically guaranteed. Let π_j be a permutation of $\{1, \ldots, K\}$ that defines the order statistic for $\hat{\beta}_{j1}, \ldots, \hat{\beta}_{jK}$, satisfying that $\hat{\beta}_{j,\pi_j(1)} < \cdots < \hat{\beta}_{j,\pi_j(K)}$. The algorithm starts with $\tilde{\boldsymbol{\beta}}_j^{(K)} = \hat{\boldsymbol{\beta}}_j$. This solution defines a partition $\mathcal{C}_j^{(K)} = \big\{\{\pi_j(1)\}, \ldots, \{\pi_j(K)\}\big\}$, where each subset contains exactly one latent class and the subsets are ordered based on the initial estimate $\hat{\boldsymbol{\beta}}_j$. It is easy to see that

$$\tilde{\boldsymbol{\beta}}_j^{(K)} = \arg\max_{\boldsymbol{\beta}_j} Q_j(\boldsymbol{\beta}_j; \hat{\boldsymbol{\Gamma}}) \quad \text{s.t. } \boldsymbol{\beta}_j \in \Theta_j(\mathcal{C}_j^{(K)}).$$

The algorithm then proceeds to find a partition $\mathcal{C}_j^{(K-1)}$ by merging two subsets of $\mathcal{C}_j^{(K)}$ and then learns

$$\tilde{\boldsymbol{\beta}}_j^{(K-1)} = \arg\max_{\boldsymbol{\beta}_j} Q_j(\boldsymbol{\beta}_j; \hat{\boldsymbol{\Gamma}}) \quad \text{s.t. } \boldsymbol{\beta}_j \in \Theta_j(\mathcal{C}_j^{(K-1)}).$$

More specifically, we consider merging two adjacent subsets, with the intuition that β_{jk}^*s that share the same value should cluster together in $\hat{\boldsymbol{\beta}}_j$. We consider $K-1$ candidate partitions

$$\begin{aligned}
\mathcal{C}_{j1}^{(K-1)} &= \big\{\{\pi_j(1), \pi_j(2)\}, \{\pi_j(3)\}, \ldots, \{\pi_j(K)\}\big\}, \\
\mathcal{C}_{j2}^{(K-1)} &= \big\{\{\pi_j(1)\}, \{\pi_j(2), \pi_j(3)\}, \ldots, \{\pi_j(K)\}\big\}, \\
&\cdots \\
\mathcal{C}_{j,K-1}^{(K-1)} &= \big\{\{\pi_j(1)\}, \{\pi_j(2)\}, \ldots, \{\pi_j(K-1), \pi_j(K)\}\big\},
\end{aligned}$$

each of which contains $K-1$ subsets and is obtained by merging two adjacent subsets in $\mathcal{C}_j^{(K)}$. For each candidate partition, $\mathcal{C}_{jb}^{(K-1)}$, $b = 1, \ldots, K-1$, we compute a constrained solution

$$\tilde{\boldsymbol{\beta}}_{jb}^{(K-1)} = \arg\max_{\boldsymbol{\beta}_j} Q_j(\boldsymbol{\beta}_j; \hat{\boldsymbol{\Gamma}}) \quad \text{s.t. } \boldsymbol{\beta}_j \in \Theta_j(\mathcal{C}_{jb}^{(K-1)}).$$

By construction, $\text{card}(\tilde{\boldsymbol{\beta}}_{jb}^{(K-1)}) = K-1$ for all b. The best $(K-1)$-level solution is given by the one that yields the largest pseudo-likelihood. That is, $\tilde{\boldsymbol{\beta}}_j^{(K-1)} = \tilde{\boldsymbol{\beta}}_{jb_j^{(K-1)}}^{(K-1)}$, with associated partition $\mathcal{C}_j^{(K-1)} = \mathcal{C}_{j,b_j^{(K-1)}}^{(K-1)}$, where

$$b_j^{(K-1)} = \arg\max_{1 \le b \le K-1} Q_j\big(\tilde{\boldsymbol{\beta}}_{jb}^{(K-1)}; \hat{\boldsymbol{\Gamma}}\big).$$

Iteratively, suppose that the best s-level solution $\tilde{\boldsymbol{\beta}}_j^{(s)}$ has been obtained, with associated partition $\mathcal{C}_j^{(s)} = \{C_{j1}^{(s)}, \ldots, C_{js}^{(s)}\}$, for $s = 2, 3, \ldots, K$. The derivation of the best $(s-1)$-level solution $\tilde{\boldsymbol{\beta}}_j^{(s-1)}$ is similar to that of $\tilde{\boldsymbol{\beta}}_j^{(K-1)}$. That is, the candidate partitions for the $(s-1)$-level solution are given by

$$\begin{aligned}
\mathcal{C}_{j1}^{(s-1)} &= \big\{C_{j1}^{(s)} \cup C_{j2}^{(s)}, C_{j3}^{(s)}, \ldots, C_{js}^{(s)}\big\}, \\
\mathcal{C}_{j2}^{(s-1)} &= \big\{C_{j1}^{(s)}, C_{j2}^{(s)} \cup C_{j3}^{(s)}, \ldots, C_{js}^{(s)}\big\}, \\
&\cdots \\
\mathcal{C}_{j,s-1}^{(s-1)} &= \big\{C_{j1}^{(s)}, C_{j2}^{(s)}, \ldots, C_{j,s-1}^{(s)} \cup C_{js}^{(s)}\big\}.
\end{aligned}$$

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

For each candidate partition, we compute a constrained solution

$$\tilde{\boldsymbol{\beta}}_{jb}^{(s-1)} = \arg\max_{\boldsymbol{\beta}_j} Q_j(\boldsymbol{\beta}_j; \hat{\boldsymbol{\Gamma}}) \quad \text{s.t. } \boldsymbol{\beta}_j \in \Theta_j(\mathcal{C}_{jb}^{(s-1)}),\ b = 1, \ldots, s-1.$$

Then, $\tilde{\boldsymbol{\beta}}_j^{(s-1)} = \tilde{\boldsymbol{\beta}}_{jb_j^{(s-1)}}^{(s-1)}$, with associated partition $\mathcal{C}_j^{(s-1)} = \mathcal{C}_{jb_j^{(s-1)}}^{(s-1)}$, where

$$b_j^{(s-1)} = \arg\max_{1 \le b \le s-1} Q_j\left(\tilde{\boldsymbol{\beta}}_{jb}^{(s-1)}; \hat{\boldsymbol{\Gamma}}\right).$$

By running the above calculation for $s = K, K-1, \ldots, 2$, we obtain $\tilde{\boldsymbol{\beta}}_j^{(1)}, \ldots, \tilde{\boldsymbol{\beta}}_j^{(K)}$ as the candidate 1- to K-level models. This process is summarized in Algorithm 1. Compared with an exhaustive search over all possible sparse models, the stepwise algorithm above is significantly more computationally efficient. That is, Algorithm 1 compares only $K(K-1)/2$ models, whereas the exhaustive search compares a number of models that is superexponential in K. When $K = 10$, the stepwise method reduces the number of models to compare from 115,975 in the exhaustive search to 45.

Although the algorithm does not exhaustively search over all possible sparse models, the theoretical results in Section 2.4 show that, as the sample size N goes to infinity, the true sparse structure for item j is captured by $\tilde{\boldsymbol{\beta}}_j^{(m_j^*)}$ with probability tending to 1. Consequently, together with the EBIC introduced below, consistent model selection can be achieved.

Algorithm 1 Stepwise search for item-level sparsity.

Input: Item j; number of latent classes K; unrestricted item-response estimate $\hat{\boldsymbol{\beta}}_j$; pseudo-likelihood function $Q_j(\,\cdot\,; \hat{\boldsymbol{\Gamma}})$.
Let π_j be a permutation of $\{1, \ldots, K\}$ such that $\hat{\beta}_{j,\pi_j(1)} < \cdots < \hat{\beta}_{j,\pi_j(K)}$.
Set $\tilde{\boldsymbol{\beta}}_j^{(K)} \leftarrow \hat{\boldsymbol{\beta}}_j$.
Initialize the ordered partition, $\mathcal{C}_j^{(K)} \leftarrow \{\{\pi_j(1)\}, \ldots, \{\pi_j(K)\}\}$.
for $s = K, K-1, \ldots, 2$ **do**
    Write the current ordered partition as $\mathcal{C}_j^{(s)} = \{C_{j1}^{(s)}, \ldots, C_{js}^{(s)}\}$.
    for $b = 1, \ldots, s-1$ **do**
        Form $\mathcal{C}_{jb}^{(s-1)}$ by merging the adjacent blocks $C_{jb}^{(s)}$ and $C_{j,b+1}^{(s)}$ in $\mathcal{C}_j^{(s)}$.
        Compute

$$\tilde{\boldsymbol{\beta}}_{jb}^{(s-1)} = \arg\max_{\boldsymbol{\beta}_j} Q_j(\boldsymbol{\beta}_j; \hat{\boldsymbol{\Gamma}}) \quad \text{s.t. } \boldsymbol{\beta}_j \in \Theta_j(\mathcal{C}_{jb}^{(s-1)}).$$

    end for
    Select $b_j^{(s-1)} = \arg\max_{1 \le b \le s-1} Q_j\left(\tilde{\boldsymbol{\beta}}_{jb}^{(s-1)}; \hat{\boldsymbol{\Gamma}}\right)$.
    Set $\tilde{\boldsymbol{\beta}}_j^{(s-1)} \leftarrow \tilde{\boldsymbol{\beta}}_{jb_j^{(s-1)}}^{(s-1)}$.
    Set $\mathcal{C}_j^{(s-1)} \leftarrow \mathcal{C}_{jb_j^{(s-1)}}^{(s-1)}$ and keep the inherited block order.
end for
Output: Candidate solutions $\{\tilde{\boldsymbol{\beta}}_j^{(m)} : m = 1, \ldots, K\}$.

Example 1 (continued). To further illustrate Algorithm 1, we use a simulated dataset generated from the toy example in Table 1, with sample size $N = 1,000$. We consider Item 1, for which the true item-response probability vector $\boldsymbol{\beta}_1^*$ is $(0.10, 0.10, 0.10, 0.90)$. The initial estimate gives $\tilde{\boldsymbol{\beta}}_1^{(4)} = \hat{\boldsymbol{\beta}}_1 = (0.141, 0.103, 0.083, 0.883)$, where the estimates of the first three parameters are close to each other but

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

not exactly the same. Sorting the entries of $\hat{\boldsymbol{\beta}}_1$ in increasing order gives $\pi_1(1) = 3$, $\pi_1(2) = 2$, $\pi_1(3) = 1$, and $\pi_1(4) = 4$, as well as the ordered partition $\mathcal{C}_1^{(4)} = \big\{\{3\}, \{2\}, \{1\}, \{4\}\big\}$.

To find $\tilde{\boldsymbol{\beta}}_1^{(3)}$, there are three candidate partitions obtained by merging adjacent blocks:

$$\mathcal{C}_{11}^{(3)} = \big\{\{2,3\}, \{1\}, \{4\}\big\}, \quad \mathcal{C}_{12}^{(3)} = \big\{\{3\}, \{1,2\}, \{4\}\big\}, \quad \mathcal{C}_{13}^{(3)} = \big\{\{3\}, \{2\}, \{1,4\}\big\}.$$

These partitions yield the constrained estimates $\tilde{\boldsymbol{\beta}}_{11}^{(3)} = (0.141, 0.094, 0.094, 0.883)$, $\tilde{\boldsymbol{\beta}}_{12}^{(3)} = (0.124, 0.124, 0.083, 0.883)$, and $\tilde{\boldsymbol{\beta}}_{13}^{(3)} = (0.490, 0.103, 0.083, 0.490)$, with pseudo-likelihood values -353.713, -354.332, and -532.043, respectively. Since $\tilde{\boldsymbol{\beta}}_{11}^{(3)}$ gives the largest pseudo-likelihood, we obtain $\tilde{\boldsymbol{\beta}}_1^{(3)} = \tilde{\boldsymbol{\beta}}_{11}^{(3)} = (0.141, 0.094, 0.094, 0.883)$, with associated partition $\mathcal{C}_1^{(3)} = \mathcal{C}_{11}^{(3)} = \big\{\{2,3\}, \{1\}, \{4\}\big\}$.

By further executing Algorithm 1, we obtain $\tilde{\boldsymbol{\beta}}_1^{(2)} = (0.114, 0.114, 0.114, 0.883)$ and $\tilde{\boldsymbol{\beta}}_1^{(1)} = (0.323, 0.323, 0.323, 0.323)$. We note that when the number of levels is correctly set at 2, the estimate $\tilde{\boldsymbol{\beta}}_1^{(2)}$ recovers the same sparsity pattern as the true parameter in Table 1, which is $\mathcal{C}_1^{(2)} = \{\{1,2,3\}, \{4\}\}$.

2.3.3. Extended BIC

To determine the optimal sparsity level for each item, we construct an EBIC as

$$\mathrm{EBIC}_j(m_j) = -2Q_j(\tilde{\boldsymbol{\beta}}_j^{(m_j)}; \hat{\boldsymbol{\Gamma}}) + m_j \log(N) + 2m_j \log(\rho). \tag{4}$$

This EBIC is constructed following the idea proposed in the EBIC for high-dimensional regression models (Chen & Chen, 2008, 2012), where ρ is a constant that characterizes the user's prior belief on model sparsity. More specifically, the first two terms in (4) mimic the standard BIC (Schwarz, 1978), but replace a real likelihood function with the pseudo-likelihood. These terms arise from the Laplace approximation to the marginal likelihood in a Bayesian formulation in which a prior distribution is imposed on all possible models and their model parameters. The last term arises from the prior distribution over models with different sparsity levels, with $\rho \geq 1$ being the ratio of the prior probability of a randomly chosen m-level model to that of a randomly chosen $(m+1)$-level model. We set $\rho \geq 1$ to ensure that sparse models are preferred. When ρ is set to be 1, the third term becomes zero, so that the EBIC becomes a pseudo-BIC. The role of ρ will be clarified in Section 3.3.

Based on the EBIC, we select the number of distinct levels for item j as

$$\tilde{m}_j = \arg\min_{m_j} \mathrm{EBIC}_j(m_j).$$

Given $\tilde{m}_j$, we infer the sparse item-response probability pattern based on the refined estimate $\tilde{\boldsymbol{\beta}}_j^{(\tilde{m}_j)}$.

Next, we comment on the EBIC. As we set ρ to be a constant, the third term in (4) is dominated by the second term $m_j \log(N)$ as the sample size N goes to infinity. Consequently, it will not affect the asymptotic property of the EBIC. However, this term is useful when the number of classes is relatively large compared with the sample size. In that case, the model space is large, and the effective sample size is relatively small. As a result, the standard BIC tends to over-select, as in high-dimensional regression settings (Chen & Chen, 2008, 2012). Our simulation results in Section 3 show that this additional penalty term substantially improves model selection accuracy. In particular, $\rho = 20$ provides a practical finite-sample compromise: it reduces over-selection relative to weaker penalties while avoiding the pronounced under-selection induced by larger penalties. Moreover, as shown in Section 2.4, the EBIC consistently learns the true number of levels m_j^* for each item j, that is, as the sample size N goes to infinity, with probability tending to 1, $\tilde{m}_j = m_j^*$. Combined with the model selection consistency of $\tilde{\boldsymbol{\beta}}_j^{(m_j^*)}$ from the stepwise search algorithm, the refined estimator $\tilde{\boldsymbol{\beta}}_j^{(\tilde{m}_j)}$ consistently recovers the sparse pattern of $\boldsymbol{\beta}_j^*$.

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

Example 1 (continued). We illustrate the use of EBIC by continuing the analysis of Example 1. From the stepwise search, we obtained the four candidate models with pseudo-likelihood values

$$Q_1(\tilde{\boldsymbol{\beta}}_1^{(4)};\hat{\boldsymbol{\Gamma}}) = -353.47,\ Q_1(\tilde{\boldsymbol{\beta}}_1^{(3)};\hat{\boldsymbol{\Gamma}}) = -353.71,\ Q_1(\tilde{\boldsymbol{\beta}}_1^{(2)};\hat{\boldsymbol{\Gamma}}) = -355.60,\ Q_1(\tilde{\boldsymbol{\beta}}_1^{(1)};\hat{\boldsymbol{\Gamma}}) = -629.11.$$

With $N = 1,000$ and $\rho = 20$, the EBIC values for the four candidate models are

$$\mathrm{EBIC}_1(4) = 758.54,\quad \mathrm{EBIC}_1(3) = 746.12,\quad \mathrm{EBIC}_1(2) = 736.99,\quad \mathrm{EBIC}_1(1) = 1,271.12.$$

Hence,

$$\tilde{m}_1 = \arg\min_{m\in\{1,2,3,4\}} \mathrm{EBIC}_1(m) = 2,$$

which correctly recovers the true number of distinct levels for Item 1. And the corresponding estimate $\tilde{\boldsymbol{\beta}}_1^{(2)} = (0.114, 0.114, 0.114, 0.883)$ recovers the same sparsity pattern as the true parameter vector in Table 1.

2.3.4. Final estimator

We obtain the final estimate of the latent class model via constrained marginal maximum likelihood estimation, using the item-level clustering pattern determined by EBIC. That is,

$$(\bar{\boldsymbol{\nu}}, \bar{\mathbf{B}}) = \arg\max_{(\boldsymbol{\nu},\mathbf{B})} l(\boldsymbol{\nu},\mathbf{B}), s.t.\ \beta_{jg} = \beta_{jh} \text{ if } \tilde{\beta}_{jg}^{(\tilde{m}_j)} = \tilde{\beta}_{jh}^{(\tilde{m}_j)}, j = 1,\ldots,J, \text{ and } h,g = 1,\ldots,K. \tag{5}$$

The standard error for each $\bar{\beta}_{jk}$ and $\bar{\nu}_k$ is also obtained based on the Fisher information matrix of this marginal likelihood. We summarize the full proposed procedure in Algorithm 2.

Algorithm 2 Proposed refinement procedure.

Input: Binary response matrix $\mathbf{Y}$; number of latent classes K; initial LCA estimate $(\hat{\boldsymbol{\nu}}, \hat{\mathbf{B}})$; posterior probability matrix $\hat{\Gamma}$; EBIC tuning parameter ρ.
for $j = 1,\ldots,J$ **do**
  Construct the item-specific pseudo-likelihood $Q_j(\boldsymbol{\beta}_j;\hat{\Gamma})$ as in (2.3.1).
  Apply Algorithm 1 to obtain candidate estimates $\{\tilde{\boldsymbol{\beta}}_j^{(m)} : m = 1,\ldots,K\}$.
  for $m = 1,\ldots,K$ **do**
    Compute

$$\mathrm{EBIC}_j(m) = -2Q_j(\tilde{\boldsymbol{\beta}}_j^{(m)};\hat{\Gamma}) + m\log N + 2m\log\rho.$$

  end for
  Select

$$\tilde{m}_j = \arg\min_{1\le m\le K} \mathrm{EBIC}_j(m),$$

  and let $\tilde{\mathcal{C}}_j$ be the partition induced by $\tilde{\boldsymbol{\beta}}_j^{(\tilde{m}_j)}$.
end for
Estimate $(\bar{\boldsymbol{\nu}}, \bar{\mathbf{B}})$ by maximising the marginal likelihood subject to $\boldsymbol{\beta}_j \in \Theta_j(\tilde{\mathcal{C}}_j)$ for all $j = 1,\ldots,J$.
Output: Selected item-level sparsity pattern $\{\tilde{m}_j, \tilde{\mathcal{C}}_j : j = 1,\ldots,J\}$; final sparse LCA estimate $(\bar{\boldsymbol{\nu}}, \bar{\mathbf{B}})$.

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

2.4. Theoretical results

In what follows, we show that the sparse item-response probability pattern implied by $\tilde{\boldsymbol{\beta}}_j^{(\tilde{m}_j)}$ for each item j is statistically consistent, for all $j = 1, \ldots, J$. This result further implies that the statistical inference based on the final estimate $(\bar{\mathbf{B}}, \bar{\boldsymbol{\nu}})$ is asymptotically valid. We make the following assumptions.

Assumption 1. The MLE in (2.1) satisfies

$$\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}^*\| = O_p(1/\sqrt{N}) \text{ and } \|\hat{\mathbf{B}} - \mathbf{B}^*\|_F = O_p(1/\sqrt{N}),$$

where $\boldsymbol{\nu}^*$ and $\mathbf{B}^*$ are the corresponding true model parameters.

Assumption 2. The true parameters lie in the interior of the parameter space; that is, $0 < \beta_{jk}^* < 1$ and $\nu_k^* > 0$ for all $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

Since J and K are fixed, Assumption 2 implies that there exists a constant $\varepsilon > 0$ such that $\varepsilon \le \beta_{jk}^* \le 1 - \varepsilon$ and $\nu_k^* \ge \varepsilon$ for all $j = 1, \ldots, J$ and $k = 1, \ldots, K$.

Assumption 3. The parameter ρ in the EBIC is a constant that does not depend on the sample size N.

Theorem 1. *Suppose that Assumptions 1–3 hold. Then, for each item $j = 1, \ldots, J$, the probability $P(\tilde{m}_j = m_j^*)$ converges to 1, as the sample size N goes to infinity. Moreover, for each item $j = 1, \ldots, J$, the refined estimator $\tilde{\boldsymbol{\beta}}_j^{(\tilde{m}_j)}$ satisfies:*

1. *$P(\tilde{\beta}_{jg} = \tilde{\beta}_{jh})$ converges to 1, for all g and $h \in \{1, \ldots, K\}$ such that $\beta_{jg}^* = \beta_{jh}^*$, and*
2. *$P(\tilde{\beta}_{jg} \neq \tilde{\beta}_{jh})$ converges to 1, for all g and $h \in \{1, \ldots, K\}$ such that $\beta_{jg}^* \neq \beta_{jh}^*$,*

as the sample size N goes to infinity.

The proof of the theorem is given in Section A of the Supplementary Material. Theorem 1 gives two results about the statistical consistency of the proposed method. The first result, $P(\tilde{m}_j = m_j^*)$ converges to 1, states that the EBIC selection step consistently identifies the true number of distinct response-probability levels for each item. The second result, which concerns $P(\tilde{\beta}_{jg} = \tilde{\beta}_{jh})$ and $P(\tilde{\beta}_{jg} \neq \tilde{\beta}_{jh})$, further establishes the consistency in recovering the sparse pattern for each item: latent classes with the same true item-response probability are asymptotically grouped together, whereas latent classes with different true item-response probabilities are asymptotically kept in separate groups. Thus, the refinement procedure consistently recovers both the item-specific model complexity and the equality pattern among the K latent classes.

3. Simulation study

3.1. Simulation settings

We consider two simulation settings. Setting I has $K = 4$ latent classes and $J = 32$ binary items, and Setting II has $K = 8$ latent classes and $J = 64$ binary items. For each setting, we generate one true class proportion vector $\boldsymbol{\nu}$ and one true item-response probability matrix $\mathbf{B}$ and then keep these true parameters fixed across all sample sizes and replications. The true class proportions are $\boldsymbol{\nu} = (0.241, 0.259, 0.202, 0.298)$ in Setting I and $\boldsymbol{\nu} = (0.106, 0.137, 0.130, 0.101, 0.124, 0.117, 0.150, 0.136)$ in Setting II. For each setting, we consider sample sizes $N \in \{500, 750, 1,000, 1,500, 2,000\}$ and run 100 independent replications for each sample size. The two settings also differ in the level configurations of the true item-response probabilities. In Setting I, all 32 items have two distinct response-probability levels, where a level denotes a distinct value among the K class-specific response probabilities for a given item rather than an observed response category. In Setting II, the 64 items consist of 48 two-level items

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

and 16 three-level items, yielding a richer and more complex item-level structure. The full true class proportions and item-response probabilities are reported in Section B of the Supplementary Material.

In the refinement stage, we consider multiple candidate values of the EBIC tuning parameter ρ, namely, 1, 5, 10, 20, 40, 80, 160, and 320. Recall that the additional EBIC penalty is $2m_j \log\rho$, so this term is zero when $\rho = 1$ and increases with ρ, placing a stronger penalty on denser models.

In the main simulation analysis, the unrestricted first-stage model for $(\hat{\mathbf{B}}, \hat{\boldsymbol{\nu}})$ is fitted using the true value of K, so that the reported results focus on the proposed item-level refinement.[2] Furthermore, we also conducted additional simulations to compare the proposed procedure with the regularized method of Chen et al. (2017); the results are reported in Section E of the Supplementary Material.

3.2. Evaluation criteria

Multiple criteria are used to evaluate the performance of the proposed method. First, we evaluate the selection of the item-specific numbers of probability levels. For each simulated dataset, we record the numbers of items that are under-selected, correctly selected, and over-selected relative to the true values $\{m_j^*\}_{j=1}^J$. We also record the total number of incorrectly selected items per replication, where an item is counted as incorrect whenever $\tilde{m}_j \neq m_j^*$.

Second, we assess recovery of the sparse item-response probability pattern itself. This is done by comparing the partition of the latent classes induced by $\boldsymbol{\beta}_j^*$, denoted $\mathcal{C}_j^*$, with that induced by $\tilde{\boldsymbol{\beta}}_j$, denoted $\tilde{\mathcal{C}}_j$, using the adjusted Rand index (ARI; Hubert & Arabie, 1985). For each item j, we compute

$$\mathrm{ARI}(\mathcal{C}_j^*, \tilde{\mathcal{C}}_j) = \frac{\sum_{g,h}\binom{n_{gh}}{2} - \frac{\sum_g \binom{n_{g\cdot}}{2}\sum_h \binom{n_{\cdot h}}{2}}{\binom{K}{2}}}{\frac{1}{2}\left[\sum_g \binom{n_{g\cdot}}{2} + \sum_h \binom{n_{\cdot h}}{2}\right] - \frac{\sum_g \binom{n_{g\cdot}}{2}\sum_h \binom{n_{\cdot h}}{2}}{\binom{K}{2}}}.$$

Here, n_{gh} denotes the number of latent classes that fall into group g under $\mathcal{C}_j^\star$ and into group h under $\tilde{\mathcal{C}}_j$, and $n_{g\cdot} = \sum_h n_{gh}$ and $n_{\cdot h} = \sum_g n_{gh}$ are the corresponding marginal counts. For each simulated dataset and each value of ρ, we compute the average item-level ARI, $(1/J)\sum_{j=1}^J \mathrm{ARI}(\mathcal{C}_j^\star, \tilde{\mathcal{C}}_j)$.

Finally, we assess parameter estimation accuracy for both the item-response probabilities and the class proportions. Specifically, for each simulated dataset and each value of ρ, we compute the mean squared error of the final refined estimator:

$$\mathrm{MSE}(\tilde{\mathbf{B}}) = \frac{1}{JK}\min_{\pi}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\tilde{\beta}_{j,\pi(k)} - \beta_{jk}^*\right)^2,$$

where the minimum is taken over all permutations π on $\{1, \ldots, K\}$ to account for label indeterminacy. Let $\hat{\pi}$ denote the permutation that attains this minimum. We then compute the corresponding error for the class proportions using the same alignment:

$$\mathrm{MSE}(\tilde{\boldsymbol{\nu}}) = \frac{1}{K}\sum_{k=1}^{K}\left(\tilde{\nu}_{\hat{\pi}(k)} - \nu_k^*\right)^2.$$

To evaluate the effect of the refinement step, we compute the same quantities for the unrestricted first-stage estimator $(\hat{\mathbf{B}}, \hat{\boldsymbol{\nu}})$ and compare them with those of the final refined estimator $(\tilde{\mathbf{B}}, \tilde{\boldsymbol{\nu}})$.

[2] Additional simulations have been conducted to assess the selection of the number of classes K by BIC. The results, which are given in Section C of the Supplementary Material, show that K can be accurately selected under the current simulation settings.

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

Table 2. Item-level selection errors under different EBIC tuning values ρ

Setting I										
ρ	Under-selected items					Over-selected items				
	500	750	1, 000	1, 500	2, 000	500	750	1, 000	1, 500	2, 000
1	0.00	0.00	0.00	0.00	0.00	1.45	1.32	1.06	0.76	0.69
5	0.00	0.00	0.00	0.00	0.00	0.25	0.28	0.24	0.14	0.14
10	0.00	0.00	0.00	0.00	0.00	0.11	0.19	0.13	0.08	0.10
20	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.07**	**0.11**	**0.08**	**0.06**	**0.06**
40	0.00	0.00	0.00	0.00	0.00	0.05	0.03	0.01	0.04	0.03
80	0.00	0.00	0.00	0.00	0.00	0.04	0.01	0.00	0.03	0.02
160	0.00	0.00	0.00	0.00	0.00	0.03	0.01	0.00	0.01	0.00
320	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.01	0.00
Setting II										
ρ	Under-selected items					Over-selected items				
	500	750	1, 000	1, 500	2, 000	500	750	1, 000	1, 500	2, 000
1	0.27	0.02	0.00	0.00	0.00	8.94	7.13	6.64	5.47	4.25
5	1.01	0.12	0.01	0.00	0.00	2.00	1.65	1.47	1.17	0.86
10	1.45	0.17	0.01	0.00	0.00	1.09	0.71	0.68	0.69	0.45
20	**2.33**	**0.32**	**0.01**	**0.00**	**0.00**	**0.40**	**0.41**	**0.36**	**0.38**	**0.21**
40	3.24	0.46	0.02	0.00	0.00	0.16	0.21	0.17	0.17	0.11
80	4.04	0.64	0.06	0.00	0.00	0.09	0.07	0.08	0.11	0.08
160	5.18	0.86	0.12	0.00	0.00	0.07	0.05	0.02	0.05	0.04
320	6.24	1.28	0.14	0.00	0.00	0.03	0.03	0.01	0.04	0.01

Note: For each setting and sample size, the table reports the average numbers of under-selected and over-selected items per replication. The row for $\rho = 20$ is highlighted in bold because this value is used in the remaining simulation summaries.

3.3. Simulation results

We first discuss the choice of ρ, which controls a clear trade-off between over-selection and under-selection in the refinement step. As ρ increases, denser models are penalized more heavily, so over-selection tends to decrease. At the same time, a larger ρ makes the procedure more prone to under-selection. Table 2 shows the average numbers of under-selected and over-selected items under different values of ρ. In Setting I, this trade-off is mild. Because this setting is relatively simple, the overall performance is strong across all sample sizes. Under-selection is essentially absent across all sample sizes and all candidate values of ρ, while over-selection decreases steadily as ρ increases and also becomes smaller as N increases. In contrast, the trade-off is much more pronounced in Setting II. For each fixed ρ, both types of selection error generally decrease as N increases. For each fixed sample size, increasing ρ sharply reduces over-selection, but this improvement is accompanied by more under-selection, especially when N is small. For example, at $N = 500$, over-selection drops from 8.94 items at $\rho = 1$ to 0.40 at $\rho = 20$, while under-selection increases from 0.27 to 2.33.

For the remaining analyses, we use $\rho = 20$. This choice is motivated by the fact that the real-data application in Section 4 has sample and item sizes similar to those of the $N = 750$ case under Setting II. For this case, $\rho = 20$ provides a reasonably balanced compromise between over-selection and under-selection. In particular, at $N = 750$ in Setting II, increasing ρ from 1 to 20 reduces the average number of over-selected items from 7.13 to 0.41, while the average number of under-selected items

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

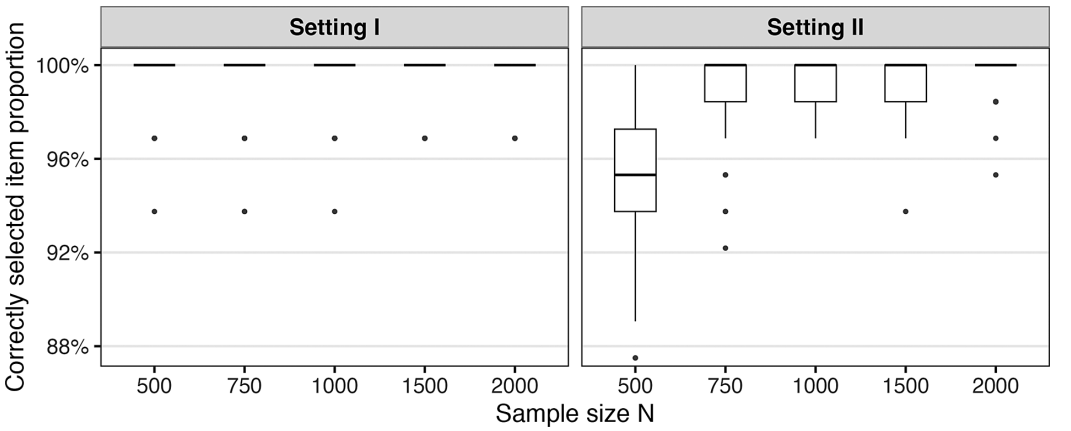


Figure 1. Proportion of items with correctly selected m_j in each replication at $\rho = 20$. Columns correspond to Settings I and II.

is still moderate at 0.32; larger values of ρ reduce over-selection only slightly further but induce more under-selection. This asymmetry is important because under-selection can introduce substantial bias in the refined estimates of $\bar{\mathbf{B}}$, whereas moderate over-selection is less harmful for estimation accuracy. We therefore use $\rho = 20$ in the remaining simulation summaries and in the real-data analysis. Results for the other candidate values of ρ are reported in Section D of the Supplementary Material.

Figure 1 summarizes, for each replication, the proportion of items for which the number of levels, m_j, is correctly selected. In Setting I, this proportion is concentrated very close to one across all sample sizes, indicating that EBIC almost always recovers the true item-level complexity. Even in the smallest samples, the worst replications incorrectly select at most two items out of 32, and the distribution becomes increasingly concentrated at perfect selection as N increases. Setting II is more challenging, but the boxplots still shift steadily upward with N. At $N = 500$, the median proportion of correctly selected items is already above 0.95, meaning that the typical replication incorrectly selects no more than three items out of 64. Once N reaches 750 or above, the median is essentially at perfect selection, and by $N = 2,000$, most replications achieve perfect selection. Overall, these results show that the EBIC step accurately identifies the true number of item-level probability levels, and that its performance improves with sample size, in line with theoretical results.

Figure 2 provides a complementary summary based on the mean item-level ARI, which measures how well the estimated class partition matches the true sparse partition once m_j has been selected. In Setting I, the ARI remains high, and the median is essentially at perfect partition recovery across all sample sizes; as N increases, the remaining outliers become less frequent. In Setting II, recovery is weaker at $N = 500$, with a mean item-level ARI of around 0.95, but it improves quickly with sample size and exceeds 0.98 once N reaches 750. By $N = 2,000$, most replications recover the true sparse pattern almost perfectly, apart from a small number of outliers. Taken together, these results show that the selected item-level sparsity pattern converges rapidly to the truth, with most remaining recovery errors concentrated in the smallest-sample Setting II scenario.

Figure 3 compares the MSE of the unrestricted first-stage estimate with that of the final refined estimate, separately for $\mathbf{B}$ and $\boldsymbol{\nu}$. For both parameters, the MSE decreases with increasing sample size, consistent with theoretical results. For $\mathbf{B}$, the refined estimator uniformly improves on the unrestricted estimator across both simulation settings and all sample sizes. In Setting I, the mean MSE decreases from 1.73×10^{-3} to 8.32×10^{-4} at $N = 500$, and from 4.27×10^{-4} to 2.04×10^{-4} at $N = 2,000$. In Setting II, the corresponding decrease is from 3.27×10^{-3} to 1.69×10^{-3} at $N = 500$, and from 7.90×10^{-4} to 2.20×10^{-4} at $N = 2,000$. These gains indicate that, once the item-level sparse structure is recovered with sufficient accuracy, imposing the selected equality constraints substantially improves estimation of the item-response probability matrix.

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

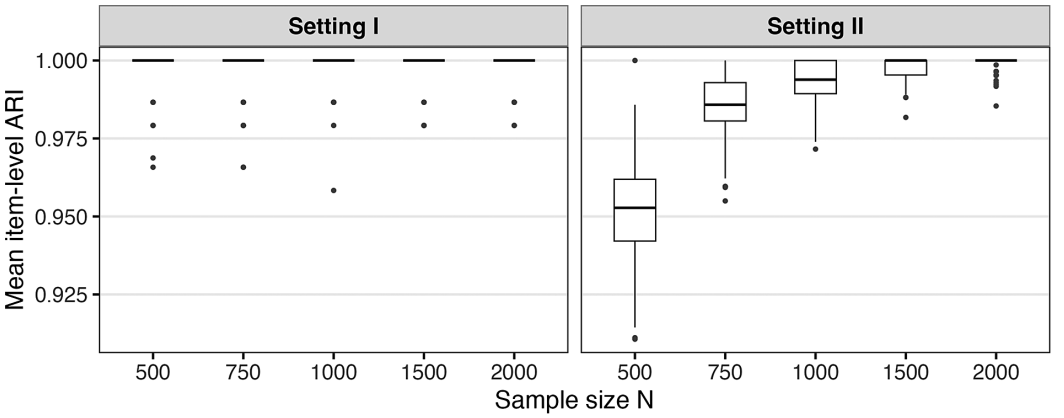


Figure 2. Mean item-level ARI in each replication at $\rho = 20$. Columns correspond to Settings I and II.

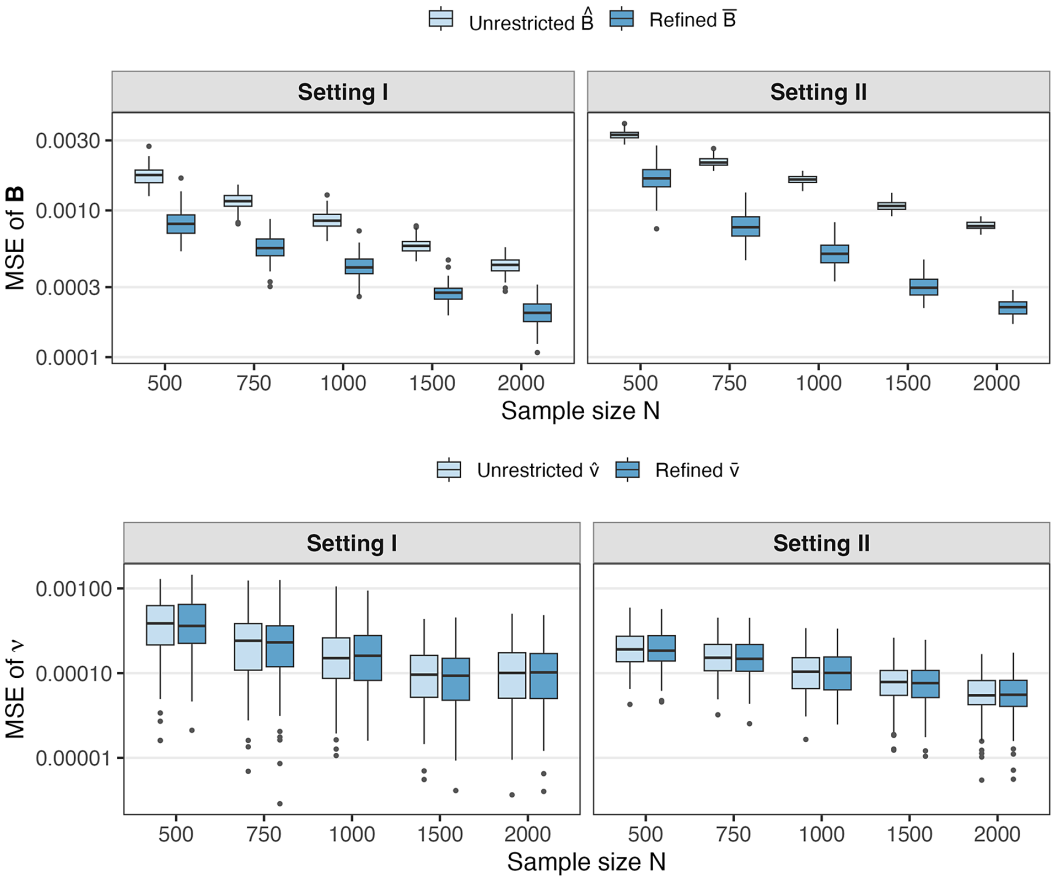


Figure 3. Mean squared error of the item-response probability matrix **B** and class proportions ν across replications at $\rho = 20$. The top panel reports the MSE for **B**, and the bottom panel reports the MSE for ν. Within each setting and sample size, the two boxplots compare the unrestricted first-stage estimate with the final refined estimate.

For the class proportions $\boldsymbol{\nu}$, the unrestricted and refined estimates are much closer, as expected, because the refinement step is designed primarily to improve estimation of **B** rather than the class proportions. The MSE of $\boldsymbol{\nu}$ nevertheless decreases with sample size in both settings. In Setting I, the mean refined MSE decreases from 4.53×10^{-4} at $N = 500$ to 1.18×10^{-4} at $N = 2,000$; in Setting II, it

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

decreases from 2.14×10^{-4} to 6.23×10^{-5} over the same range. Thus, the main estimation gain from refinement occurs for $\mathbf{B}$, while estimation of $\boldsymbol{\nu}$ is largely preserved.

4. An application to social role performance data

4.1. Data

We applied the proposed refinement procedure to the Social Role Performance (SRPPER) item bank from the PROMIS Wave 1 calibration study (Cella et al., 2010). The analysis included respondents with complete responses to all SRPPER items, yielding a response matrix with $N = 737$ respondents and $J = 56$ items, listed in Table G1 in the Supplementary Material.

The original SRPPER items in the PROMIS dataset were measured on a five-point ordinal scale with response categories Never, Rarely, Sometimes, Often, and Always. Because the items included both positively and negatively worded statements, the numeric scores from 1 to 5 were assigned to align all items in the same direction, with higher values indicating more severe limitations in social role performance. In the analysis, we dichotomized each item as $Y_{ij} = 1$ for responses coded 4 or 5 and $Y_{ij} = 0$ otherwise. Therefore, the item-response probabilities reported below can be interpreted as the probabilities of endorsing a more severe response under this binary coding.

4.2. Analysis workflow

The real-data analysis followed the strategy developed in Section 2.3. We first fitted unrestricted latent class models with candidate numbers of classes ranging from 1 to 8 and selected $K = 8$ latent classes based on BIC. We then applied the proposed item-level refinement procedure to this selected unrestricted solution. The refinement step used $\rho = 20$, as suggested by the simulation results (see Section 3.3 and Setting II in Table 2), to select item-specific equality groups among the eight latent classes. Finally, the sparse latent class model was re-estimated under these selected equality constraints, producing the refined item-response probability matrix $\bar{\mathbf{B}}$ and the estimated class proportions.

4.3. Results

The refinement substantially reduced the complexity of the item-response probability matrix. Among the 56 SRPPER items, the selected number of distinct probability levels was $\tilde{m}_j = 3$ for 7 items, $\tilde{m}_j = 4$ for 38 items, and $\tilde{m}_j = 5$ for 11 items. Thus, instead of estimating eight unrelated class-specific probabilities for every item, the refined model represents each item using only three to five ordered probability levels. This reduction is directly tied to the goal of the proposed method: it preserves heterogeneity across items while making the class differences more interpretable. The unrestricted first-stage estimates, together with the corresponding comparison heatmap using the same item grouping, are reported in Section F.1 of the Supplementary Material. Relative to the refined matrix, the unrestricted estimates are visibly noisier: many item-specific probabilities vary gradually across classes rather than collapsing into a small number of repeated levels, so the overall severity ordering remains visible but the domain-specific contrasts are much harder to identify directly from the raw estimates.

For presentation, the eight latent classes were labeled in ascending order of average item-response probability. The estimated class proportions after this relabeling are 0.142, 0.086, 0.075, 0.089, 0.116, 0.071, 0.117, and 0.302. The corresponding average item-response probabilities are 0.054, 0.234, 0.304, 0.509, 0.586, 0.751, 0.851, and 0.976, showing a clear ordering from the least severe to the most severe latent profiles.

Figure 4 displays the final refined item-response probability matrix $\bar{\mathbf{B}}$. Compared with the unrestricted heatmap reported in Section F.1 of the Supplementary Material, where the probabilities often vary gradually across classes and the relevant level patterns are difficult to see, the refined matrix makes the level structure visually transparent. Classes assigned to the same item-specific level have

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

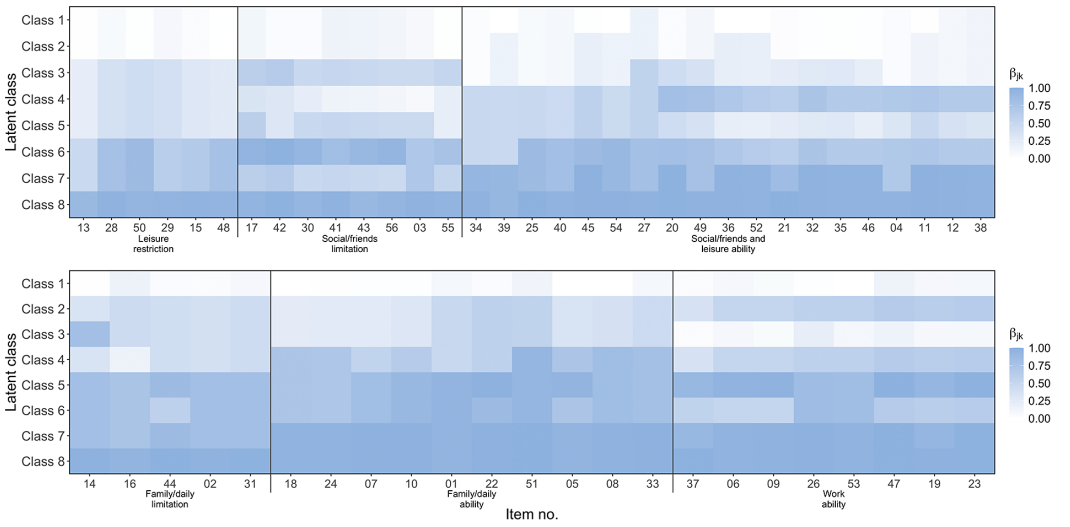


Figure 4. Heatmap of the final refined item-response probability matrix for the PROMIS Social Role Performance data. For readability, the 56 items are displayed across two stacked panels. Within each block, items are ordered by hierarchical clustering of their refined probability profiles; the labels beneath the *x*-axis summarize the content of each block. Rows correspond to relabeled latent classes ordered by increasing average probability of endorsing the more severe binary response.

identical refined probabilities, and classes assigned to adjacent levels can be interpreted as having low, intermediate, or high probabilities of endorsing the severe response. Consequently, when the items are ordered by hierarchical clustering and the latent classes are ordered from the lowest to the highest class-average severe-response probability, the heatmap reveals clearer block structures across items.

The hierarchical clustering of the items reveals six contiguous item blocks, highlighted in Figure 4 and defined in Table G1 in the Supplementary Material. They are (i) *leisure restriction items* (6 items), (ii) *social/friends limitation items* (8 items), (iii) *social/friends and leisure ability items* (19 items), (iv) *family and daily-role limitation items* (5 items), (v) *family and daily-role ability items* (10 items), and (vi) *work ability items* (8 items). This ordering is not driven by topical content alone. Instead, the clustering also separates item polarity: family and daily-role items split into limitation versus ability wording, while the broader social/leisure domain separates into a shared ability block and two negatively worded blocks corresponding primarily to leisure and to social/friends functioning.

Along the vertical axis, the relabeled Classes 1–8 form an overall severity ordering. Class 1 is characterized by the lowest or near-lowest item-specific probability level across essentially all blocks, indicating little endorsement of severe responses throughout leisure, social/friends, family/daily-role, and work domains. At the opposite end, Class 8 is uniformly high across all six blocks, with almost all items assigned to the highest or near-highest level, representing a globally severe profile.

The intermediate classes differ not only in overall severity but also in which domains become elevated first. In Class 2, the first three blocks, corresponding to leisure restriction, social/friends limitation, and social/friends and leisure ability, remain at low levels, whereas the family/daily-role and work blocks begin to rise. Class 3 shows a different pattern: the limitation blocks for social/friends and leisure become elevated, especially the social/friends limitation block, while the shared social/friends and leisure ability block remains relatively intact. In the family/daily-role domain, Class 3 is similar to Class 2, but its work block stays much closer to the low level. Class 4 is marked by a clearer increase in the family/daily-role ability block and in the social/friends and leisure ability block, suggesting that positive-functioning items in these domains are now differentiating the class more strongly than in the preceding profiles.

From Class 5 onward, the profiles move into a broadly moderate-to-high range. Class 5 is the first class showing a more general mid-to-high pattern across the heatmap, with the

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

family/daily-role and work blocks rising earliest and most clearly. Class 6 is also moderate-to-high overall, but is distinguished by especially severe social/friends difficulties: the social/friends limitation block stands out more strongly, whereas the work block is not especially high relative to the other domains. Class 7 is higher still, with the three ability blocks—social/friends and leisure ability, family/daily-role ability, and work ability—already reaching the highest level for many items, while the two limitation-oriented domains remain mostly at intermediate-to-high levels. Thus, neighboring classes are not separated by severity alone; they also differ in the domain-specific pattern of impairment and in whether the strongest differentiation appears first in limitation-oriented versus ability-oriented item blocks.

Overall, the refinement turns the dense 56×8 probability matrix into a structured profile in which the latent classes differ not only in the severity and emphasis of social-role dysfunction, but also in their response tendency across positively versus negatively worded items. In particular, the heatmap suggests that some respondents may be relatively less likely to endorse the more severe wording on negatively framed limitation/restriction items while showing a different tendency on positively framed ability items. This means that the class structure reflects both substantive differences in social-role functioning and systematic differences in how respondents interact with item wording polarity across domains.

5. Discussion

In this article, we introduce a post-estimation refinement procedure that bridges the gap between the exploratory power of LCA and its interpretability. By decoupling the initial estimation and the sparsity-inducing model selection, our approach sidesteps the computational bottlenecks with regularized estimation methods while maintaining the asymptotic guarantees necessary for reliable inference.

The proposed refinement procedure extends naturally to the LCA of polytomous data, where each item j can have $C_j \geq 2$ response categories. Under this setting, the item-response probability matrix expands into a three-way tensor, where each entry β_{jkc} represents the probability of class k providing response c to item j. In this case, a refinement procedure would not only aim to collapse the K multinomial distributions $\boldsymbol{\beta}_{jk} = (\beta_{jk1}, \ldots, \beta_{jkC_j})^\top$, $k = 1, \ldots, K$, into a set of $m_j < K$ unique distributions, but also collapse the multinomial probabilities $\beta_{jk1}, \ldots, \beta_{jkC_j}$ into a smaller number of unique values for each class k. An item-specific pseudo-likelihood can be easily defined for this setting by replacing the weighted Bernoulli log-likelihood in (2.3.1) with a weighted multinomial log-likelihood, based on which an EBIC can be defined accordingly. A new stepwise search algorithm is needed to accommodate the dual collapsing tasks—one over the latent classes and the other over the response categories. One way to achieve it is to perform the search in two stages. In the first stage, we use a stepwise search algorithm, as in Algorithm 1, to collapse the multinomial distributions over the K latent classes. In the second stage, for each collapsed class from the first stage, we further collapse the item-response probabilities over the C_j response categories. Similar consistency results to those in Theorem 1 can be established. We leave this extension for future investigation.

Another future direction is to extend the sparse LCA framework to multi-group LCA (Clogg & Goodman, 1985; McCutcheon, 2002), an extension of classical LCA that has received wide real-world applications. The multi-group LCA concerns data with observed groups (e.g., gender, ethnicity, and country) and aims to assess differences between these groups using LCA. The item-response probabilities in multi-group LCA can be complex, as they depend on both latent classes and observed groups. A sparse estimate may lead to clearer item-response probabilities and thus improve the interpretability of the resulting multi-group LCA model.

The proposed procedure is based on a pseudo-likelihood approach. An alternative strategy would be to adopt a full likelihood framework incorporating an L_0-type penalty on the differences $\beta_{jk} - \beta_{jk'}$ for $1 \leq k < k' \leq K$, thereby encouraging fusion of item-response probabilities across classes. This idea is in the spirit of the fusion approach for levels of categorical predictors in high-dimensional logistic regression proposed by Kaufmann and Kateri (2024). Future research could develop this likelihood-based formulation and compare its performance with the method proposed here.

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

Finally, we note that the current methodology and theoretical framework were developed for the low-dimensional regime, in which the sample size N is much larger than the number of items J. However, modern applications of LCA often encounter high-dimensional settings in which J is comparable to, or even exceeds, N. Because the proposed method and theory may not scale well to such settings, new frameworks are required. To address this high-dimensional LCA problem, future research could adapt the methodologies and theoretical insights developed for high-dimensional latent factor models, such as those in Chen et al. (2019), Chen, Li, et al. (2020) and Chen and Li (2022).

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10149.

Data availability statement. The code for implementing the proposed method is publicly available at https://github.com/florence07/Sparse-LCA-Refinement.

Author contributions. Y.C., M.K., and I.M. conceived the research idea. Y.X., L.K., and Y.C. developed the method and theory and conducted simulation studies and real-data analysis. All authors contributed to the final manuscript.

Funding statement. Y.X. was supported by a Ph.D. scholarship from the China Scholarship Council (CSC). L.K. received financial support from the London School of Economics Research Investment Fund for her 2025 research visit to the Department of Statistics at the London School of Economics and Political Science.

Competing interests. Y.C. is an Associate Editor for Psychometrika but was excluded completely from the review process. The other authors declare no competing interests.

AI declaration. During the preparation of this manuscript, the authors used ChatGPT (OpenAI) and Gemini (Google) to assist with improving the clarity, grammar, and readability of the text, and Codex to enhance code readability and documentation. The authors reviewed, edited, and verified all content generated with these tools and take full responsibility for the accuracy and integrity of the work.

References

Bakk, Z., & Kuha, J. (2018). Two-step estimation of models between latent classes and external variables. *Psychometrika*, *83*(4), 871–892.

Bakk, Z., & Kuha, J. (2021). Relating latent class membership to external variables: An overview. *British Journal of Mathematical and Statistical Psychology*, *74*(2), 340–362.

Bowers, J., & Culpepper, S. A. (2024). Equivalence set restricted latent class models (ESRLCM). Preprint, arXiv:2406.03653.

Cella, D., Riley, W., Stone, A., Rothrock, N., Reeve, B., Yount, S., Amtmann, D., Bode, R., Buysse, D., Choi, S., Cook, K., DeVellis, R., DeWalt, D., Fries, J. F., Gershon, R., Hahn, E. A., Lai, J.-S., Pilkonis, P., Revicki, D., … Hays, R. D. (2010). The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005–2008. *Journal of Clinical Epidemiology*, *63*(11), 1179–1194.

Chen, J., & Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, *95*(3), 759–771.

Chen, J., & Chen, Z. (2012). Extended BIC for small-n-large-p sparse GLM. *Statistica Sinica*, *22*(2), 555–574.

Chen, Y., Culpepper, S., & Liang, F. (2020). A sparse latent class model for cognitive diagnosis. *Psychometrika*, *85*(1), 121–153.

Chen, Y., & Li, X. (2022). Determining the number of factors in high-dimensional generalized latent factor models. *Biometrika*, *109*(3), 769–782.

Chen, Y., Li, X., Liu, J., & Ying, Z. (2017). Regularized latent class analysis with application in cognitive diagnosis. *Psychometrika*, *82*(3), 660–692.

Chen, Y., Li, X., & Zhang, S. (2019). Joint maximum likelihood estimation for high-dimensional exploratory item factor analysis. *Psychometrika*, *84*(1), 124–146.

Chen, Y., Li, X., & Zhang, S. (2020). Structured latent factor analysis for large-scale data: Identifiability, estimability, and their implications. *Journal of the American Statistical Association*, *115*(532), 1756–1770.

Chen, Y., Liu, J., Xu, G., & Ying, Z. (2015). Statistical analysis of Q-matrix based diagnostic classification models. *Journal of the American Statistical Association*, *110*(510), 850–866.

Clogg, C. C., & Goodman, L. A. (1985). Simultaneous latent structure analysis in several groups. *Sociological Methodology*, *15*, 81–110.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, Behavioral, and health sciences*. John Wiley & Sons.

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press

Goodman, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable: Part I—A modified latent structure approach. *American Journal of Sociology*, *79*(5), 1179–1259.
Goodman, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, *61*(2), 215–231.
Gu, Y., & Xu, G. (2020). Partial identifiability of restricted latent class models. *The Annals of Statistics*, *48*(4), 2082–2107.
He, S., Culpepper, S. A., & Douglas, J. A. (2025). A sparse latent class model incorporating response times. *British Journal of Mathematical and Statistical Psychology*, *78*(3), 757–784.
Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.
Kaufmann, L., & Kateri, M. (2024). Simultaneous factors selection and fusion of their levels in penalized logistic regression. *Electronic Journal of Statistics*, *18*(2), 4235–4291.
Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis*. Houghton Mifflin Company.
McCutcheon, A. L. (2002). Basic concepts and procedures in single- and multiple-group latent class analysis. In J. A. Hagenaars, & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 56–86). Cambridge University Press.
Robitzsch, A. (2020). Regularized latent class analysis for polytomous item responses: An application to SPM-LS data. *Journal of Intelligence*, *8*(3), 30.
Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
von Davier, M., & Lee, Y.-S. (Eds.). (2019). *Handbook of diagnostic classification models*. Springer.
Wang, C., & Lu, J. (2021). Learning attribute hierarchies from data: Two exploratory approaches. *Journal of Educational and Behavioral Statistics*, *46*(1), 58–84.
Zhu, Y., Steele, F., & Moustaki, I. (2017). A general 3-step maximum likelihood approach to estimate the effects of multiple latent categorical variables on a distal outcome. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(5), 643–656.

Cite this article: Xu, Y., Kaufmann, L., Chen, Y., Kateri, M. and Moustaki, I. (2026). Sparse Latent Class Analysis for Dichotomous Responses: Post-Estimation Refinement via Item-Level Pseudo-Likelihood. *Psychometrika*, 1–19. https://doi.org/10.1017/psy.2026.10149

https://doi.org/10.1017/psy.2026.10149 Published online by Cambridge University Press