Journal of Computerized Adaptive Testing

Volume 12 Number 3
July 2025

Estimating the Joint Item-Score Density Using an Unrestricted Latent Class Model: Advancing Flexibility in Computerized Adaptive Testing

Anastasios Psychogyiopoulos, Niels Smits, and L. Andries van der Ark
University of Amsterdam

The *Journal of Computerized Adaptive Testing* is published by the International Association for Computerized Adaptive Testing
www.iacat.org/jcat

ISSN: 2165-6592

International Association for Computerized Adaptive Testing
IACAT
Advancing the Science and Practice of Human Assessment

Journal of Computerized Adaptive Testing

Volume 12, Number 3, July 2025

DOI 10.7333/2507-1203136

ISSN 2165-6592

Estimating the Joint Item-Score Density Using an Unrestricted Latent Class Model: Advancing Flexibility in Computerized Adaptive Testing

Anastasios Psychogyiopoulos, Niels Smits, and L. Andries van der Ark
University of Amsterdam

Computerized adaptive testing (CAT) reduces cognitive fatigue and response burden while maintaining measurement precision by administering items tailored to the respondent. However, the assumptions of item response theory models—commonly used in CAT—might be too stringent for some tests. This study investigated the bias and accuracy of a flexible CAT procedure, called LSCAT (for latent-class sum-score CAT). In the calibration phase, an unrestricted latent class model estimates the joint item-score density ($\boldsymbol{\pi}$) and the total-score density ($\boldsymbol{\pi}_+$); in the operational phase, the respondents' expected total scores are estimated. The paper's first study indicated that using the Bayesian information criterion (BIC) to determine the number of latent classes produced the most accurate estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$. The second study showed that the unrestricted latent class model more accurately estimated $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ than the two-parameter logistic model, especially under a complex data-generating mechanism. As a proof of concept, the third study compared the precision of LSCAT and a traditional CAT procedure using the two-parameter logistic model with a single empirical dataset. The two CAT procedures were approximately equally precise. Although the two procedures had the same fixed efficiency, LSCAT was more efficient for the high- and low-scoring respondents, while traditional CAT was more efficient for respondents in the middle.

Keywords: CAT simulation, computerized adaptive testing, density-estimation, item-calibration, latent class analysis, model selection.

In computerized adaptive testing (CAT; e.g., Van der Linden & Glas, 2010; Wainer & Dorans, 2000) the items are selected based on the examinee's responses to previous items: Less able examinees are presented with easier questions, whereas more able examinees are presented with more difficult items (Eggen & Verschoor, 2006; Wainer & Dorans, 2000). CAT is used to measure an examinee's ability in various domains and is often linked to distal outcomes. For example, in healthcare CAT is used to measure patient-reported outcomes (e.g., Choi et al., 2010; Flens et al., 2017) and predict physical, mental, and social health (Gibbons et al., 2012). In education, CAT is used for university admission, for example, and predicts future educational

achievement (e.g., Conrad, 1977; Kuncel et al., 2001). By using CAT, more precise results are acquired in less time resulting in more efficient assessments (e.g., Weiss, 2004).

Over the past few decades, significant progress has been made in CAT techniques, from the earliest methods (Clark, 1976; Lord, 1969) to more recent developments (Magis et al., 2017; von Davier, Mislevy, et al., 2021). Traditionally, examiners often aim to measure a specific ability trait (Wainer & Dorans, 2000), which is why the construction of a CAT often relies on unidimensional item response theory (IRT) models (Van der Linden, 2018; Van der Linden & Glas, 2010). When the construct being measured comprises multiple non-orthogonal traits, multidimensional CATs can be constructed (Chalmers, 2012, 2016; Segall, 1996) using multidimensional IRT models (Reckase, 2009).

Despite the widespread use of IRT methods, other non-IRT alternatives have also been developed to accommodate tests that do not satisfy the statistical assumptions of IRT models. Van Buuren and Eggen (2017) introduced CAT employing an unrestricted latent class model (ULCM) to assign respondents to proficiency categories of a nominal latent variable. By means of cognitive diagnosis models (see von Davier & Lee, 2019, for an overview), cognitive diagnostic CAT (CD-CAT; e.g., Cheng, 2009; Sorrel et al., 2021) is used to assign respondents to attribute profiles, which are also categories of a nominal latent variable. Other non-IRT alternatives include the use of decision trees to overcome the absence of unidimensionality (Yan et al., 2004) or local independence assumptions (Ueno & Songmuang, 2010). Such more advanced CAT applications are needed not just to go beyond restricted statistical models but also to integrate techniques from advanced computing methods and incorporate a wide variety of input data in order to stay current with technological developments (Veldkamp, 2022; von Davier, Di Cerbo, et al., 2021).

Van der Ark and Smits (2023) proposed a framework for CAT, which they coined FlexCAT, consisting of two main components: the *engine* and the *score*. The *engine* refers to the model used to estimate the joint item-score density (i.e., a vector containing the probabilities for observing each item-response pattern), denoted as $\boldsymbol{\pi}$. For example, three dichotomous items result in the following eight response patterns (i.e., 000, 001, 010, 011, 100, 101, 110 and 111). If in the population, all respondents would respond randomly with probability 0.5, then all response patterns would be equally likely, and $\boldsymbol{\pi} = (.125, .125, .125, .125, .125, .125, .125, .125)^{\mathrm{T}}$. The *score* refers to the variable used to communicate the measurement value of the respondent. For instance, within the FlexCAT framework, for IRT-based CAT the *engine* is an IRT model (for instance, a two-parameter logistic model; 2PLM) from which estimates of $\boldsymbol{\pi}$ can be derived, and the *score* is the estimated latent trait value (θ). In ULCM-based CAT proposed by Van Buuren and Eggen (2017), the *engine* is a ULCM and the *score* is the respondent's estimated latent class membership. In most CAT applications, the score constitutes a part of the engine, which is computationally efficient and might be intuitively appealing. For example, in IRT-CAT, direct estimation of the joint item-score density, $\boldsymbol{\pi}$, can be circumvented because the IRT model also produces estimates of scores. Note that an estimate of $\boldsymbol{\pi}$ can be derived from the item-parameter estimates, in combination with the assumed distribution of the latent trait, θ. Within the FlexCAT framework, the engine and score may be independent, providing more flexibility in choosing the score of interest. For example, in decision-tree-based CAT (Ueno & Songmuang, 2010; Yan et al., 1998, 2004), the engine is the decision tree and the score is the expected total score of the respondent.

In this paper, the focus is on FlexCAT with the ULCM as an engine, along with both the item-response pattern for scores and the (unweighted) total score, hence denoted as LRCAT and LSCAT respectively. For some CAT applications, the ULCM can be an attractive engine because of its flexibility compared to other latent variable models. The ULCM assumes only local independence of the item scores given latent class membership (e.g., Vermunt &

Magidson, 2004, pp. 2-3), putting relatively few constraints on the data, resulting in a relatively good model fit. Thus, the ULCM might be an attractive engine if the data are not consistent with an IRT model. Van der Ark and Smits (2023) discussed several other reasons why the ULCM is an attractive engine.

First, using the ULCM also allows CAT for prediction purposes, both with an internal criterion (e.g., "Does the respondent belong to the 3% highest scoring respondents?") and an external criterion (e.g., "Will the respondent fall back into criminal behavior?") because a ULCM can model items and internal and external criteria simultaneously.

Second, the ULCM is able to handle tests and questionnaires that contain items with different numbers of response categories. Areas where these questionnaires are commonly used include mental care (e.g., Jelínek et al., 2021), youth care (e.g., Ebesutani et al., 2011), and quality of life research (e.g., Kim, 2014).

In traditional applications of the ULCM (e.g., Hagenaars & McCutcheon, 2002), the selected number of latent classes is limited, say 2 to 6, to facilitate the interpretation of the latent classes. Van Buuren and Eggen (2017) also constructed their ULCM-based CAT using a limited number of latent classes, a plausible approach in their case since these latent classes were also used as scores, necessitating a clear interpretation. The ULCM is used here differently, as a density estimator (e.g., Linzer, 2011; Vermunt et al., 2008): Its sole purpose is to estimate $\boldsymbol{\pi}$ as accurately as possible. Contrasting the ULCM for density estimation with traditional latent class analysis, Vermunt et al. (2008) identified four aspects of using the ULCM as a density estimator. First, the latent classes need not be interpreted, suggesting that for density estimation the number of latent classes is determined by the fit of $\boldsymbol{\pi}$, not by substantive concerns. Second, in density estimation, overfitting is less problematic than underfitting. This suggests that when in doubt about the number of latent classes needed for a good model fit, the more flexible model might be preferred. Third, identifiability issues are of no concern for density estimation. Finally, local optima—which are difficult to avoid with a larger number of latent classes—typically yield ULCM parameter estimates that result in estimates of $\boldsymbol{\pi}$ nearly as good as those based on the global maximum. Van der Palm et al. (2016b) showed that a hierarchical version of the ULCM could effectively detect higher-order effects in multivariate discrete densities. Other applications where a ULCM has been used as a density estimator include smoothing large contingency tables (Linzer & Lewis, 2011), imputing missing data (Van der Palm et al., 2016a), and estimating test-score reliability (Van der Ark et al., 2011).

Both the total score and the item-response pattern are considered to be reasonable choices for a score. The total score is of interest because it is still used in many psychological and educational tests. Similarly, the item-response pattern is a relevant choice since it is the foundation of most scores, including the latent trait θ used in IRT. If the engine is the ULCM, then the θ—the standard score in CAT—is not an obvious choice for a score. The ULCM can provide an estimate of $\boldsymbol{\pi}$, but then an IRT model would be required to estimate θ from $\widehat{\boldsymbol{\pi}}$ It can be expected that estimating θ directly from the observed data or estimating θ from the smoothed data in $\widehat{\boldsymbol{\pi}}$ would provide very similar estimates.

The primary objective of this article was to evaluate the accuracy of the ULCM as a density estimator of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ when discrete item scores are used; inaccurate estimates could distort individual measurement and prediction. In the next sections, after discussing the ULCM and ULCM model fit, three simulation studies are presented addressing three specific research questions. Study 1 addressed the question of how the number of latent classes in the ULCM should be selected to obtain the most accurate estimate of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$. Study 2 addressed the question of whether the ULCM can provide accurate estimates of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ when the data-generating process is complex due to higher-order interactions among the item scores. Whereas such complexity might not be typical for practical CAT applications, the ability of ULCM to

provide an accurate estimate of any $\boldsymbol{\pi}$ can be considered an asset of the method. Finally, in Study 3, LSCAT was employed in practice using CAT simulations with empirical data. Additionally, an IRT-CAT simulation was conducted with the 2PLM as an engine and θ as the score using the same data. The outcome of Study 3 demonstrated how both CAT methods performed in comparison to a full-item test.

The Unrestricted Latent Class Model for Density Estimation

The ULCM

Before discussing the utility of the ULCM as a density estimation tool and the challenges of ULCM fit, some relevant notation is needed to aid understanding of the approach used in this paper. Assume that a test contains J items, each with $m+1$ ordered response categories $(0, 1, \ldots, m)$. If $m = 1$, the items are dichotomously scored, whereas if $m > 1$, the items are polytomously scored. Let Y_j $(j = 1, \ldots, J)$ be a random variable denoting the score on item j, and let y_j be the realization of Y_j. Let $\mathbf{Y} = \left(Y_1, \ldots, Y_J\right)^{\mathsf{T}}$ be a multivariate random variable containing all item scores, which is referred to as the *item-response pattern*, with realization $\mathbf{y} = \left(y_1, \ldots, y_J\right)^{\mathsf{T}}$. As each of the J items has $m+1$ categories, the total number of values $\mathbf{y}$ is able to take equals $L = (m+1)^J$, ranging from $\mathbf{y} = (0,0, \ldots,0)^{\mathsf{T}}$ to $\mathbf{y} = (m, m, \ldots, m)^{\mathsf{T}}$. Let $Y_+ = \sum_j X_j$ be a random variable denoting the total score—i.e., the sum of the item scores—with realization y_+; y_+ can take $L_+ = mJ + 1$ values ranging from $y_+ = 0$ to $y_+ = mJ$. For simplicity of notation, the probabilities of observing a certain item-response pattern, $P(\mathbf{Y} = \mathbf{y})$, are collected in the $L \times 1$ vector $\boldsymbol{\pi}$, whereas the probabilities of observing a certain total score, $P(Y_+ = y_+)$, are collected in the $L_+ \times 1$ vector $\boldsymbol{\pi}_+$. Let $\mathbf{Q}$ be an appropriate $L \times L_+$ design matrix in which $q_{rs} = 1$ if item-response pattern $\mathbf{y}_r$ sums to the total score s, and $q_{rs} = 0$, otherwise. Then $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ are related by

$$\boldsymbol{\pi}_+ = \mathbf{Q}^{\mathsf{T}}\boldsymbol{\pi}. \tag{1}$$

Under the ULCM, it is assumed that a latent variable Ξ with $\mathcal{K}$ discrete categories $(1, \ldots, \mathcal{K})$—usually referred to as latent classes—explains all the associations in the data. Let $P(\Xi = \xi)$ denote the probability that a randomly selected respondent belongs to latent class $(\xi = 1, \ldots, \mathcal{K})$, also known as the *class weight*. Furthermore, let $P(\mathbf{Y} = \mathbf{y}|\Xi = \xi)$ denote the conditional probability of obtaining item-response pattern $\mathbf{y}$ given that the respondent belongs to latent class ξ, also referred to as the response probability. The conditional independence assumption in the ULCM states that item scores depend only on class membership, which means that

$$P(\mathbf{Y} = \mathbf{y}|\Xi = \xi) = \prod_j \left(Y_j = y_j|\Xi = \xi\right). \tag{2}$$

Hence, under the ULCM, the probability that a randomly selected respondent obtains item-response pattern $\mathbf{y}$ equals

$$P(\mathbf{Y} = \mathbf{y}) = \sum_\xi P\left(\Xi = \xi\right) \times \prod_j P\left(Y_j = y_j|\Xi = \xi\right). \tag{3}$$

Vector $\boldsymbol{\pi}$ can be derived from Equation 3, and $\boldsymbol{\pi}_+$ can be computed from $\boldsymbol{\pi}$ using Equation 1. The matter at issue is whether the ULCM can accurately estimate $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$.

ULCM Fit and Selection

An important issue is the selection of the number of latent classes. Let ULCM(K) denote the estimated ULCM with K latent classes. Global goodness-of-fit tests for contingency tables, such as the likelihood ratio test or Pearson's chi-square test, are seldom used to select the number of latent classes. Comparing the fit of ULCM(K) to the data (i.e., the saturated model) requires a sufficiently large sample size for each probability in $\boldsymbol{\pi}$ (Koehler & Larntz, 1980). Because the size of $\boldsymbol{\pi}$ increases exponentially as the number of items increases, these tests are only useful for very small item sets. For variants of these tests that compare the fit of ULCM(K) to ULCM(K+1), the regularity conditions required to ensure that the test statistic follows a chi-square distribution, do not apply (Chen et al., 2020; Holt & Macready, 1989; Vermunt & Magidson, 2004). This renders these tests useless resulting in biased p-values. Resampling techniques might be implemented to estimate the distribution of the test statistic or combinations of local fit statistics might be inspected, such as the bivariate residual (Vermunt & Magidson, 2004). These methods can be time consuming, particularly when dealing with extensive models intended for CAT.

More popular is the use of information criteria, statistics that balance model fit and model complexity, and sometimes sample size. For instance, the Akaike Information Criterion (AIC; Akaike, 1998) can be written as the likelihood ratio test statistic minus the number of model parameters. For $K = 1, 2, \ldots$ the information criterion of ULCM(K) is computed. The number of latent classes that produces the lowest information-criterion value, denoted K^*, is selected as the best estimate of the *true* number of latent classes, $\mathcal{K}$. As the fit of the ULCM solely depends on the number of latent classes, information criteria might be useful tools for selecting the best-fitting model. In addition to AIC, popular information criteria include the consistent AIC (CAIC; Bozdogan, 1987), AIC3, (Bozdogan, 1993) as well as information criteria that are derived from Bayes factors, such as the Bayesian information criterion (BIC; Schwarz, 1978) and the adjusted BIC (aBIC; Sclove, 1987). Nylund et al. 2007 found that aBIC generally outperforms AIC and BIC, and Lukočienė and Vermunt (2010) observed a similar trend with AIC3 outperforming AIC and BIC. However, the performance depends on factors such as sample size and number of items (Whittaker & Miller, 2021). For instance, BIC and aBIC tend to perform well with large samples, whereas AIC, CAIC, and AIC3 tend to perform better in smaller sample sizes (Morgan, 2015; Nylund et al., 2007).

The cited studies comparing the performance of information criteria did not account for the specific conditions that apply to CAT: Large samples and many items, which require a large number of latent classes. Sample sizes in the cited studies varied from 200 to 1,200, and the number of items typically ranged between eight and ten, never exceeding 15. The most common choice for the true number of latent classes, denoted by $\mathcal{K}$, was three and never exceeded eight.

Study 1

The aim of Study 1 was to determine the conditions under which the joint densities of the item-scores ($\boldsymbol{\pi}$) and the total score ($\boldsymbol{\pi}_+$) can be accurately estimated. The $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ estimates, denoted by $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_+$, depend on the chosen information criterion which in turn depends on specific factors, namely the sample size, the number of items, and the true number of classes. Inaccurate estimates could distort individual measurement and prediction when used in

LSCAT. Thus, a simulation study was necessary to determine the conditions under which $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ can be used with confidence.

Method

Data generation. Data were generated from population models. A population model was derived by estimating a ULCM(K) with the dichotomous scores of 1,407 students to $\boldsymbol{J}$ items of the Czech Medical School Admission Test (MSAT-B; see Drabinová and Martinková, 2017**,** for an overview). The ULCM-parameter estimates were used as population values, from which $\boldsymbol{\pi}$ was computed using Equation 3, and $\boldsymbol{\pi}_+$ using Equation 1. Data were generated by a sample of size N from the model.

Independent variables. Three between-subject independent variables were included.

1. *Sample size*, N, had three levels: 250; 1,000; and 2,000. $N = 2,000$ mimics realistic conditions for CAT calibration, $N = 250$ serves as a minimum requirement for precise pre-test IRT calibration (Şahin & Weiss, 2015), and $N = 1,000$ as a plausible value in between.
2. The *number of items*, J, had two levels: 9 and 18. $J = 18$ approximates the computational limitation for estimating all 2^J elements of $\boldsymbol{\pi}$, and $J = 9$ is similar to values used in the existing literature on ULCM fit.
3. *Number of latent classes in the population*, $\mathcal{K}$, had three levels: 4, 8, and 12. $\mathcal{K} = 4$ is close to values typically used in the literature on investigating ULCM model fit, $\mathcal{K} = 8$ is the highest value used in the literature, and $\mathcal{K} = 12$ represents a very complex model.

Information criterion, the only within-subject independent variable, had four levels: AIC, AIC3, BIC, and aBIC, representing four of the more popular information criteria.

Dependent variables. There were two dependent variables, corresponding to the accuracy with which $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$, respectively, were estimated. The Kullback-Leibler divergence (KL; Kullback & Leibler, 1951) was used to assess the difference between the estimated densities $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_+$, and the population probability densities $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$. $\mathrm{KL}(\boldsymbol{\pi}||\hat{\boldsymbol{\pi}})$ is the logarithmic difference between the probabilities of the true density $\boldsymbol{\pi}$ and the estimated density $\hat{\boldsymbol{\pi}}$, and is defined as

$$\mathrm{KL}(\boldsymbol{\pi}||\hat{\boldsymbol{\pi}}) = \sum \boldsymbol{\pi} \times \log_2\left(\frac{\boldsymbol{\pi}}{\hat{\boldsymbol{\pi}}}\right) = H(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}}) - H(\boldsymbol{\pi}) \tag{4}$$

where $H(\boldsymbol{\pi}, \hat{\boldsymbol{\pi}})$ is the cross-entropy between $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$, and $H(\boldsymbol{\pi})$ is the entropy of $\boldsymbol{\pi}$. KL ranges from 0 to infinity, with higher values indicating greater information loss when approximating $\boldsymbol{\pi}$ with $\hat{\boldsymbol{\pi}}$. Thus, larger values of KL are less favorable.

Dependent variables $\mathrm{KL}(\boldsymbol{\pi}||\hat{\boldsymbol{\pi}})$ and $\mathrm{KL}(\boldsymbol{\pi}_+||\hat{\boldsymbol{\pi}}_+)$ were computed as follows. For $K = 1, 2, \ldots, K^{stop}$, the information criterion of ULCM(K) was computed, where K^{stop} denotes the number of latent classes for which, according to the information criterion, the model fit had not improved for three consecutive values of K. Hence, the best fitting model is ULCM(K^*) = ULCM($K^{stop} - 3$). Then, $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_+$ were computed from ULCM(K^*) using Equations 3 and 1. Note that $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ had already been calculated from the population model ULCM($\mathcal{K}$). Finally, $\mathrm{KL}(\boldsymbol{\pi}||\hat{\boldsymbol{\pi}})$ and $\mathrm{KL}(\boldsymbol{\pi}_+||\hat{\boldsymbol{\pi}}_+)$ were computed using Equation 4.

Running the simulation. A full factorial design was conducted containing $\boldsymbol{3}$ $(\boldsymbol{N})$ $\times$ $\boldsymbol{2}$ $(\boldsymbol{J})$ $\times$ $\boldsymbol{3}$ $(\boldsymbol{K})$ $\times$ $\boldsymbol{4}$ (information criterion) $=$ $\boldsymbol{72}$ conditions, each with 100 replications. The package

`poLCA` (Linzer & Lewis, 2011) in R (R Core Team, 2023) was used for model estimation. All other computations were done in personal R code available from OSF.

Results

The results from Study 1 are presented separately for $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$. A notable shared observation was that, in both scenarios, as the sample size increased, the accuracy of estimating both $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ improved across all information criteria. Moreover, it is worth noting that no non-convergence problems were encountered during the estimation process of the LCMs. However, in 2.33% of the cases, the divergence between $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ was so large that the KL($\boldsymbol{\pi}||\hat{\boldsymbol{\pi}}$) tended to infinity.

Accuracy of $\hat{\boldsymbol{\pi}}$. Based on the results in Table 1, except for two conditions with $\boldsymbol{N = 2,000}$, and $\boldsymbol{J = 18}$, where aBIC was the best performing information criterion, BIC was the most accurate information criterion in all conditions. In addition, as the sample size increased the differences among information criteria in the accuracy of $\hat{\boldsymbol{\pi}}$ became smaller. Appendix A provides more detailed results regarding the distribution of the $\boldsymbol{KL}$ values for each condition.

Accuracy of $\hat{\boldsymbol{\pi}}_+$. Within each condition, the median values of KL($\boldsymbol{\pi}_+||\hat{\boldsymbol{\pi}}_+$) did not differ substantially across information criteria (Table 2). Differences among the median values were evident at the fourth decimal place, indicating that the investigated information criteria performed equally well.

Since the differences in the median values were negligible, the differences of $\boldsymbol{\pi}_+$ and $\hat{\boldsymbol{\pi}}_+$ were inspected at a finer-grained level (Figure 1). The black dotted curve represents the true total-score density $\boldsymbol{\pi}_+$ for $J = 18$, under nine different conditions. The blue area represents the 100 curves of the estimated total-score density $\hat{\boldsymbol{\pi}}_+$under the information criterion that yielded the smallest median KL value per condition (i.e., best scenario). In contrast, the red area represents the 100 curves of $\hat{\boldsymbol{\pi}}_+$ under the information criterion that yielded the largest median KL value (i.e., worst scenario). Figure 1 shows that as the sample size increased, both the worst and best scenarios were very close since the areas overlapped.

Figure 1
An Illustrated Example of the Actual Differences of $\boldsymbol{\pi}_+$ and $\hat{\boldsymbol{\pi}}_+$ in the Best and Worst Scenarios

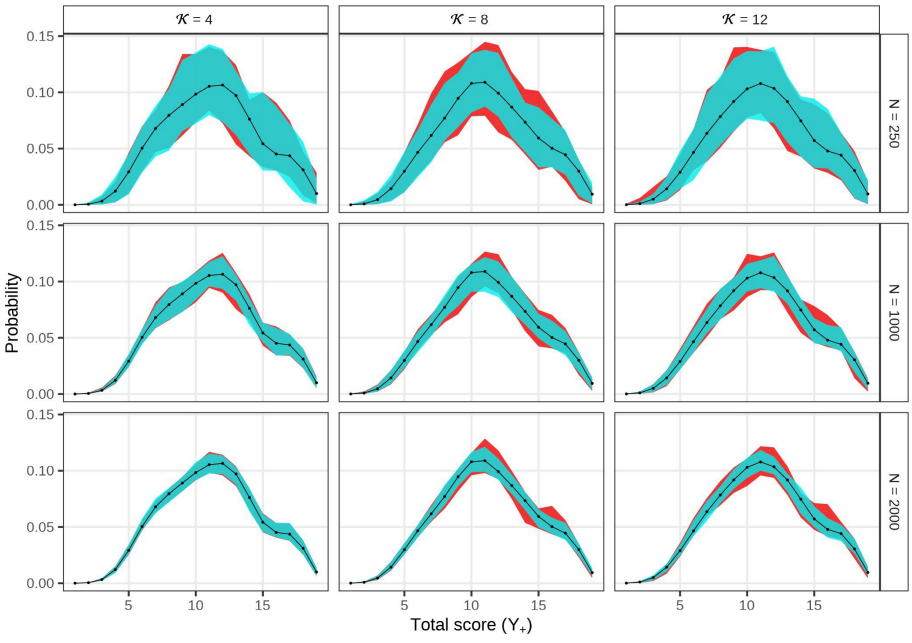


Table 1. Median Kullback-Leibler Divergence of $\boldsymbol{\pi}$ and $\hat{\boldsymbol{\pi}}$ Across 100 Replications

Simulation conditions			Information criterion			
N	**J**	$\mathcal{K}$	AIC	BIC	AIC3	aBIC
250	9	4	1.4241	**0.6768**	0.7352	0.8336
		8	1.5078	**0.5691**	0.8640	0.7185
		12	1.1299	**0.4646**	0.5225	0.9410
	18	4	0.8853	**0.3373**	0.4117	0.5970
		8	1.3556	**0.4053**	0.4984	0.5712
		12	2.2988	**0.4788**	0.5025	1.3766
1000	9	4	0.0555	**0.0349**	0.0507	0.0429
		8	0.2075	**0.0501**	0.0918	0.0614
		12	0.3044	**0.0685**	0.0854	0.0797
	18	4	0.1066	**0.0784**	0.0904	0.0891
		8	0.1828	**0.1208**	0.1393	0.1309
		12	0.3494	**0.1745**	0.2004	0.1790
2000	9	4	0.0248	**0.0182**	0.0223	0.0226
		8	0.1409	**0.0303**	0.0595	0.0445
		12	0.1990	**0.0503**	0.0567	0.0514
	18	4	0.0514	**0.0385**	0.0451	0.0442
		8	0.0923	0.0642	0.0696	**0.0591**
		12	0.1373	0.1241	0.1042	**0.0972**

Note. The values represent the median Kullback-Leibler divergence between the true and estimated $\boldsymbol{\pi}$ in a distribution of 100 replications. The smaller median per condition (i.e., per row) is denoted in bold.

Conclusions

Study 1 demonstrated that, in general, BIC outperformed other information criteria for finding the ULCM that resulted in the most accurate estimate of $\boldsymbol{\pi}$. In contrast, all competing information criteria (AIC, AIC3, BIC, aBIC) performed equally well in selecting the model that yielded the most accurate estimate of $\boldsymbol{\pi}_+$. In this study, the ULCM served as the data-generating mechanism for all conditions. The question that arose at this point was how well the ULCM could estimate the densities of $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ under different data-generating mechanisms.

Study 2

Study 2 investigated whether the ULCM could accurately estimate $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ even under a complex data-generating mechanism that included higher-order interaction effects. In addition to a complex data-generating model, described below, the ULCM and the 2PLM were also included, as benchmarks, as data-generating models. It was hypothesized that the ULCM engine would work best when the data were also generated under a ULCM, and the 2PLM engine would work best when the data were also generated under a 2PLM. It was expected that under the complex data-generating mechanism, the ULCM engine would outperform the 2PLM engine, because the ULCM has less restrictive assumptions than the 2PLM.

Table 2. Median Kullback-Leibler Divergence of $\boldsymbol{\pi}_+$ and $\widehat{\boldsymbol{\pi}}_+$ Across 100 Replications

Simulation conditions			Information criterion			
N	J	$\mathcal{K}$	AIC	BIC	AIC3	aBIC
250	9	4	0.0108	0.0105	**0.0099**	0.0101
		8	0.0103	0.0102	**0.0099**	0.0101
		12	0.0104	0.0107	**0.0102**	0.0105
	18	4	**0.0145**	0.0156	0.0174	0.0166
		8	**0.0149**	0.0189	0.0167	0.0185
		12	**0.0180**	0.0195	0.0199	0.0205
1000	9	4	0.0022	**0.0019**	0.0021	0.0022
		8	0.0024	**0.0021**	0.0024	0.0024
		12	**0.0026**	0.0032	0.0030	0.0032
	18	4	0.0029	0.0028	**0.0026**	0.0028
		8	**0.0037**	0.0050	0.0039	0.0039
		12	**0.0038**	0.0067	0.0043	0.0054
2000	9	4	0.0013	**0.0012**	**0.0012**	**0.0012**
		8	0.0011	0.0011	0.0012	**0.0010**
		12	**0.0014**	0.0017	0.0015	0.0017
	18	4	0.0013	**0.0012**	**0.0012**	0.0013
		8	**0.0013**	0.0016	**0.0013**	0.0014
		12	0.0014	0.0036	**0.0013**	0.0015

Note: The values represent the median Kullback-Leibler divergence between the true and estimated $\boldsymbol{\pi}_+$ in a distribution of 100 replications. The smaller median per condition (i.e., per row) is denoted in bold.

Method

Data generation. Scores for N respondents on 9 dichotomous items were generated using three data-generating models: a 2PLM, a ULCM($\mathcal{K} = 2$), and a loglinear model that included fifth-order interactions among the item scores (cf. Van der Palm et al., 2016a, who also used a loglinear model for a complex data-generating model). The parameters from the 2PLM were obtained by estimating a 2PLM on the MSAT-B data (see Study 1), and subsequently using the estimated item parameters as the population item-parameters and the estimated θ values as the population θ values in the data-generating model. Then $\boldsymbol{\pi}$ was computed from the data-generating model, and item response patterns were sampled from $\boldsymbol{\pi}$. Similarly, the parameters of the ULCM(2) were obtained by estimating a ULCM(2) on the MSAT-B data, and subsequently using the estimated parameters as the parameters in the data-generating model. As for the 2PLM, $\boldsymbol{\pi}$ was computed from the data-generating model, and item response patterns were sampled from $\boldsymbol{\pi}$. For the loglinear model, let λ_0 be the intercept, λ_i $(i = 1, \ldots, 9)$ be the parameters for the first-order effects, λ_{ij} $(i, j = 1, \ldots, 9; i \neq j)$ be the parameters for the second-order effects, λ_{ijk} $(i, j, k = 1, \ldots, 9; i \neq j \neq k)$ be the parameters for the third-order effects, λ_{ijkl} $(i, j, k, l = 1, \ldots, 9; i \neq j \neq k \neq l)$ be the parameters for the fourth-order effects, and λ_{ijklm} $(i, j, k, l, m = 1, \ldots, 9; i \neq j \neq k \neq l \neq m)$ be the parameters for the fifth-order effects. The data-generating model was

$$\log\left(P(\mathbf{Y}|\boldsymbol{\lambda})\right) = \lambda_o + \sum_i \lambda_i Y_i + \sum_{i,j} \lambda_{ij} Y_i Y_j + \sum_{i,j,k} \lambda_{ijk} Y_i Y_j Y_k + \sum_{i,j,k,l} \lambda_{ijkl} Y_i Y_j Y_k Y_l \sum_{i,j,k,l,m} \lambda_{ijklm} Y_i Y_j Y_k Y_l Y_m \quad (5)$$

As for the previous two data-generating models, $\boldsymbol{\pi}$ was computed from the data-generating model, and item response patterns were sampled from $\boldsymbol{\pi}$. Appendix B shows the parameter values of the three data-generating models.

Independent variables. The only within-subject independent variable, *Engine*, had two levels: "2PLM" as a representative of traditional CAT and "ULCM" as a representative of LSCAT. Study 2 had two between-subject independent variables. *Data-generating model* had three levels: "2PLM", "ULCM(2)", and loglinear model (see previous subsection). *Scenario* had two levels: Scenario 1, in which for each level of data-generating model $N = 100{,}000$ item response patterns were sampled without replications, and Scenario 2, in which $N = 5{,}000$ item response patterns were sampled in 50 replications. In Scenario 1, due to the large sample size, there was very little sampling error affecting the outcomes, rendering replications unnecessary. The results may be interpreted in terms of bias. In Scenario 2, sampling error can also be studied.

Dependent variables. As in Study 1, $\text{KL}(\boldsymbol{\pi}||\hat{\boldsymbol{\pi}})$ and $\text{KL}(\boldsymbol{\pi}_+||\hat{\boldsymbol{\pi}}_+)$ were the dependent variables. It was hypothesized that under the 2PLM data-generating model, the 2PLM engine would produce the lowest KL divergence, that under the ULCM(2) data-generation model, the ULCM engine would produce the lowest KL divergence, and that under the loglinear data-generating model, both engines would be less accurate, but the ULCM engine would produce the lowest KL divergence as the ULCM is less restrictive than the 2PLM.

Running the simulation. A full factorial design was conducted with 3 (data-generating model) × 2 (engine) × 2 (scenario) = 12 conditions; 6 conditions without replications (Scenario 1) and 6 with 50 replications (Scenario 2). The R package `MIRT` (Chalmers, 2012) was used to estimate the 2PLM and, as in Study 1, the `poLCA` package (Linzer & Lewis, 2011) for ULCM estimation. All other computations were done in personal R code available from OSF.

Results

Table 3 shows the KL values for Scenario 1, a sample of $N = 100,000$, without replications. For the 2PLM data-generation model, the accuracy of the estimated densities was perfect for both 2PLM and ULCM engines. When the data-generating model was a ULCM($\mathcal{K} = 2$), the estimated densities were more precise when the fitting model was also a ULCM, yet both estimates were close to zero indicating high accuracy. Lastly, in the conditions in which the data-generation model was the loglinear model, the fitted ULCM produced more accurate estimates of the total-score density ($\hat{\boldsymbol{\pi}}_+$) compared to the estimated 2PLM, and also more precise estimates of the item-score density ($\hat{\boldsymbol{\pi}}$).

Figures 2 and 3 show the results for Scenario 2 ($N = 5,000$, repeated 50 times). Consistent with expectations, the fitted ULCM provided more accurate estimated densities, as indicated by the smaller KL values. Specifically, for the 2PLM as the data-generation model, the ULCM engine provided results as accurate as the 2PLM engine for both $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_+$. For the ULCM as the data-generation model, the ULCM engine provided more accurate estimates of $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_+$.

Table 3. Results for Scenario 1: Values of KL($\boldsymbol{\pi}||\hat{\boldsymbol{\pi}}$) and KL($\boldsymbol{\pi}_+||\hat{\boldsymbol{\pi}}_+$) for Three Data-Generating Models and Two Engines

		Data-Generating Model		
Engine	Density	2PLM	ULCM($\mathcal{K}=2$)	Loglinear Model
2PLM	$\boldsymbol{\pi}$	0.000	0.024	0.370
	$\boldsymbol{\pi}_+$	0.000	0.000	0.014
ULCM	$\boldsymbol{\pi}$	0.000	0.018	0.135
	$\boldsymbol{\pi}_+$	0.000	0.000	0.000

Note. The numbers represent the Kullback-Leibler divergence between the true and the estimated densities, KL($\boldsymbol{\pi}||\hat{\boldsymbol{\pi}}$) and KL($\boldsymbol{\pi}_+||\hat{\boldsymbol{\pi}}_+$). All values are rounded to three decimals.

For $\hat{\boldsymbol{\pi}}_+$, though, the accuracy was almost equal and very close to perfect (i.e., KL approached zero). For the loglinear data-generation model, the results differed slightly for $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_+$. For $\hat{\boldsymbol{\pi}}$, the ULCM engine provided more accurate estimates on average, but with significantly more variance than the estimates of the 2PLM. In only 4 out of 50 cases (8%) were the accuracy of the 2PLM estimates more accurate than those of the ULCM. In one of these four cases, the estimated value tended to infinity (indicated in red color in Figure 2) which indicates an issue in the ULCM fit in this replication. For $\hat{\boldsymbol{\pi}}_+$, the ULCM engine provided almost perfect estimates and were substantially more accurate than the 2PLM engine, without any substantial variance as displayed for $\hat{\boldsymbol{\pi}}$.

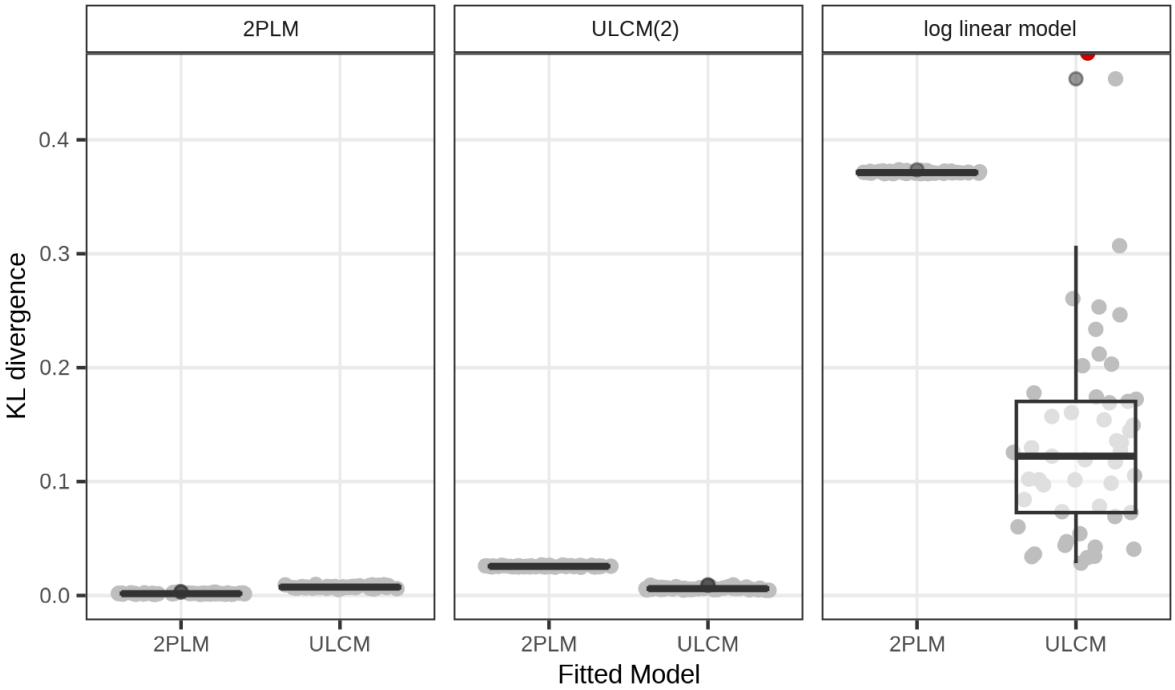
Conclusions

The ULCM engine showed strong flexibility in handling a data structure that involves higher-order interaction effects among item scores, as evidenced by the accurate estimations of $\hat{\boldsymbol{\pi}}$ and $\hat{\boldsymbol{\pi}}_+$ from simulated data generated using a loglinear model. The 2PLM engine showed precise results across all 50 replications of Scenario 2, yet were less accurate than ULCM engine on average. The results indicated that constructing a FlexCAT using the ULCM as an engine and the total score as a score, might provide unbiased results even if the data-generating mechanism is unknown or misspecified.

Study 3

Study 1 and Study 2 showed that under the simulated conditions the ULCM is a reliable density estimator ready for use as an engine, but the proposed CAT procedure has not yet been illustrated. Thus, the aim of Study 3 was to apply LSCAT and to compare it to an IRT-CAT to demonstrate their similarities and differences. More specifically, post-hoc simulations were conducted, wherein each CAT application was employed on existing empirical data as if responses were being collected adaptively (Finkelman et al., 2017; Forbey et al., 2000). The outcomes of the two CAT applications were then compared with those of the complete item score data.

Figure 2. Scenario 1: The Accuracy of $\hat{\boldsymbol{\pi}}$ With Sample Size $N = 5,000$, and $R = 50$ Replications



Note. The three panels represent the three data-generating models. Within each panel the boxplots represent the dispersion of the $\boldsymbol{\pi}$ estimates using the ULCM (right) and the 2PLM (left), respectively. The red-colored points indicate outliers that tend to infinity.

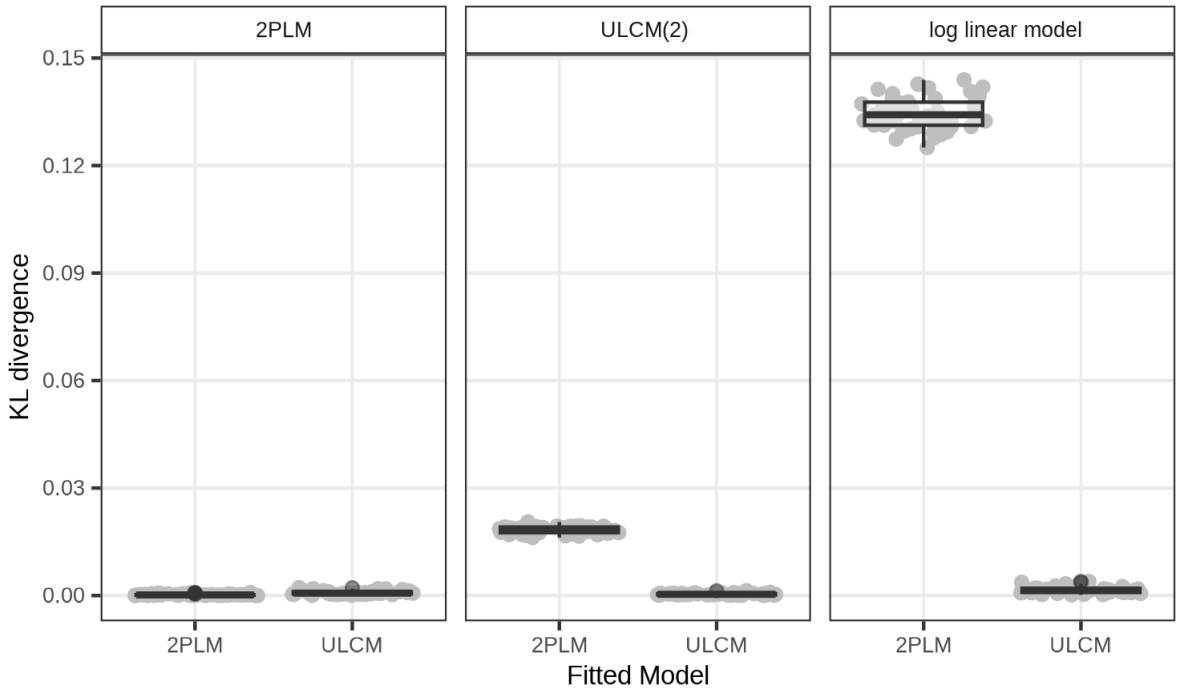
Method

Data. This study used the same empirical dataset as Van der Ark and Smits (2023), which consisted of 16 items from the ”learning task orientation” scale of the School Attitude Questionnaire – Internet (SAQI; Vorst, 2006). This scale focuses on identifying potential behavioral and educational challenges in students at school. The items, initially comprising three answer categories, were dichotomized for simplicity—following Van der Ark and Smits' suggestion—by merging two categories. This resulted in a final dataset of 16 dichotomized items coded as 1 for 'true' and 2 for 'false'. The final dataset was then randomly split into two parts: a "training" set (80% of the respondents, $N = 3,369$) and a "validation" set (20% of the respondents, $N = 842$). The training set was used for calibrating, and the validation set was used for the simulation process, which is described below.

Settings for the CAT algorithms. The IRT-CAT was calibrated using the training set by estimating a 2PLM as an engine and the latent trait θ as the score using three standard error (SE) values of $\hat{\theta}$ as stopping rules. Firstly, a value of 0.30, commonly employed as the default in the field (Wainer & Dorans, 2000), was used. Additionally, to demonstrate the effects of different stopping rule requirements, two more rules ($SE = 0.35$ and $SE = 0.40$) were applied. The IRT-CAT calibration was accomplished using the `MIRT` package (Chalmers, 2012) in R (R Core Team, 2023), with default settings. The adaptive procedure was simulated in the validation set using the `MIRTCAT` package (Chalmers, 2016) under the default settings for each of the three stopping rules.

LSCAT—in terms of the flexCAT framework—incorporated a ULCM as an engine and the total score for scoring. For calibrating LSCAT, the primary task was to estimate $\hat{\boldsymbol{\pi}}_+$ using the best fitting ULCM according to BIC, as Study 1 indicated. In setting up the adaptive procedure of LSCAT, the choices of Van der Ark and Smits (2023) for item selection, score estimation, and stopping rule were followed. Van der Ark and Smits proposed to stop the CAT once the modal

Figure 3. Scenario 2: The Accuracy of $\hat{\boldsymbol{\pi}}_+$ with Sample Size $N = 5,000$, and $R = 50$ Replications



Note. The three panels represent the three data-generating models. Within each panel the boxplots represent the dispersion of the $\boldsymbol{\pi}_+$ estimates using the ULCM (right) and the 2PLM (left), respectively.

value of $\hat{\boldsymbol{\pi}}$ exceeded some criterion c; that is, $max\ \hat{\boldsymbol{\pi}}_+ > c$. Similar to IRT-CAT, three versions of the stopping rule were employed, with $c = 0.97, 0.75,$ and 0.50, respectively (details for establishing these values are presented in the next section). LSCAT simulation was run in the validation set in R (R Core Team, 2023) using an adjustment of, and additions to the code of the `poLCA` package (Linzer & Lewis, 2011). For a detailed explanation of the item selection and stopping rules of the proposed CAT, the interested reader is referred to Van der Ark & Smits (2023).

Comparing complete and CAT outcomes. Because of the obvious trade-off between precision and efficiency (increasing the required precision increases the number of items used; decreasing the required number of items will lead to less precision; Thompson, 2019), all other things equal, an adaptive method using more items is expected to be more precise, therefore it is common to align either the precision or the efficiency among methods and evaluate the other outcome. It was chosen to do the latter, which was accomplished by adjusting the stopping rules in order to match the efficiency of the two types of CAT. As IRT-CAT is the standard in the field, its most common setup was used as a basis. The most popular stopping rule of IRT-CAT is halting the assessment once the standard error (SE) of the $\hat{\theta}$ is below a pre-specified value. Using this rule, for each SE-value, IRT-CAT was run in the validation set and the average number of required items was assessed. For LSCAT, testing stops if the modal value of $\hat{\boldsymbol{\pi}}_+$ exceeds a certain threshold c (Van der Ark & Smits, 2023). Subsequently, the precision criterion c of LSCAT was adjusted to match the SE of IRT-CAT by selecting values that resulted in an average number of administered items equal, to the nearest 0.5, to that of the IRT-CAT conditions. The outcomes under this value were recorded.

Having adjusted the stopping rules to align the CATs' efficiency, comparisons were then able to be made between each CAT and the full-item test. The evaluations were made based on the following outcomes: the *response burden* and *precision*. As measures of *response burden*, the

average (and SD) of the number of administered items was used, along with the percentage of items left to be administered (called *efficiency*).

Due to some challenges, adjustments needed to be made in the evaluation of precision. First, comparing the precision of LSCAT and IRT-CAT is hindered by the use of different *scores*. Moreover, as this was a real-data simulation no true (i.e., "data-generating") values of the respective scores were available. To deal with these challenges, the scores of both CAT applications were transformed into a common z-score metric. To compute the z-scores, the full-item mean and SD were used (for LSCAT these statistics were obtained using the full-item total-score density; for IRT-CAT these were obtained using the distribution of estimates of θ based on the full-item set). In addition, when examining precision, the estimate based on the full-item test was treated as a proxy for the true value, and the final score from each adaptive test was used as the estimated value (Smits et al., 2018). Given these adjustments, *precision* was studied using three outcomes: (1) the average difference between true and estimated z-values, (2) the correlation between true and estimated values, and (3) the difference between the true and estimated z-values as a function of the true z-value. The latter outcome was reported only if it provided information over and above the first two outcomes.

Results

Table 4 shows the response burden and precision as a function of the respective stopping rules for LSCAT and IRT-CAT. Columns three and four show the outcomes for the number of items administered. Table 4 shows that the mean difference in the number of administered items between the two methods was less than 0.5 in each condition. The efficiency measure corroborates this conclusion, showing very similar percentages of unadministered items. The SD, however, suggests that the distribution of the number of items used was different between the two methods; below this distribution is studied in more detail. With regard to the precision of the CAT applications, IRT-CATs estimated z-score and true z-score values were nearly identical, as evidenced by the mean difference of zero and almost perfect correlation. LSCAT also yielded near-perfect results, demonstrating negligible mean differences between the estimated and true z-scores.

Table 4. Results for Study 3:
Comparison of LSCAT and IRT-CAT with Full Test Scores

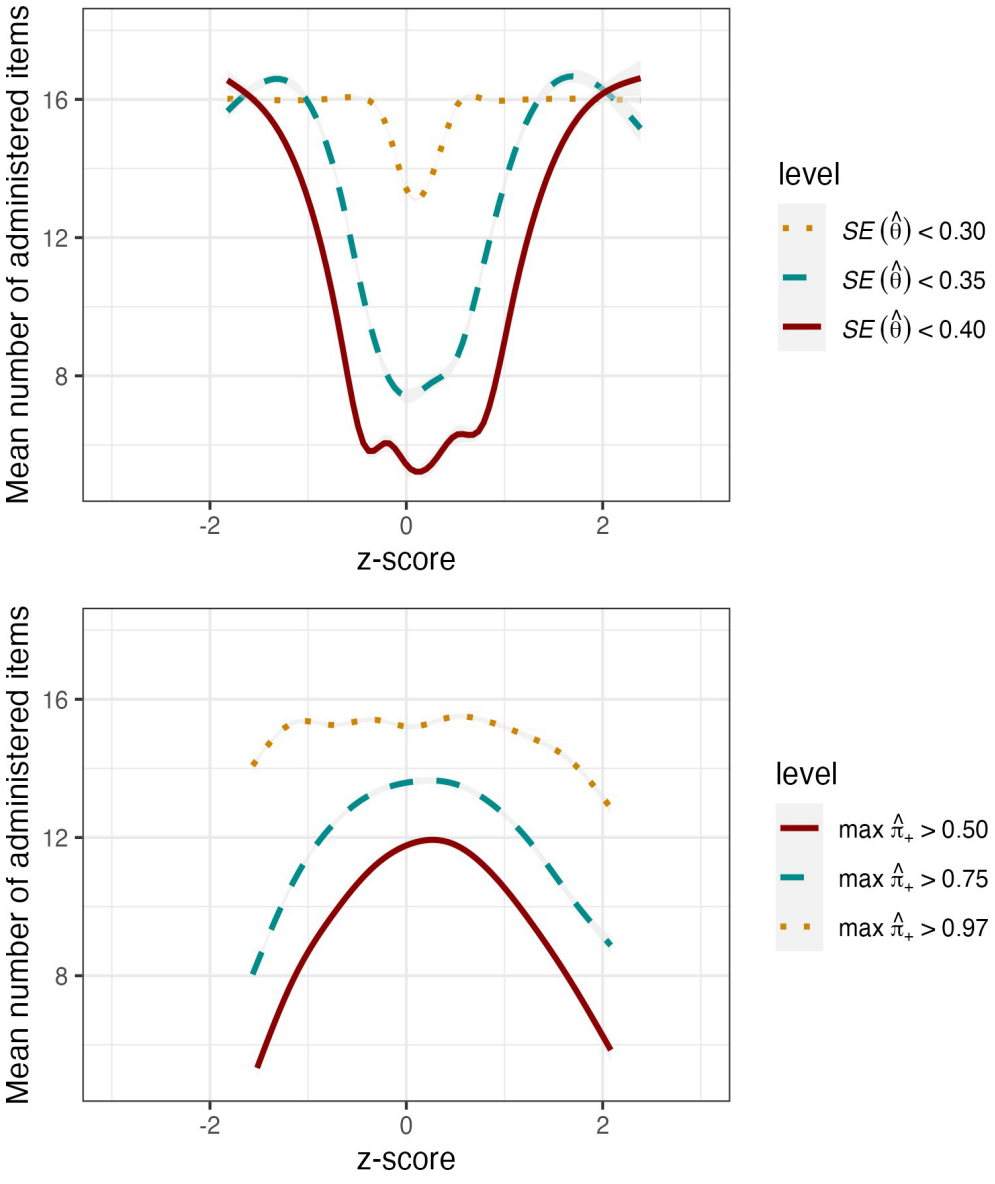
		Number of Items Used		CAT Vs. Full Test z-Score	
Method	Stopping Rule	$M(SD)$	Efficiency	Mean Difference	Correlation
IRT-CAT	$SE(\hat{\theta}) < 0.30$	15.53(0.93)	2.93%	0.00	0.999
	$SE(\hat{\theta}) < 0.35$	12.07(3.68)	24.56%	0.00	0.992
	$SE(\hat{\theta}) < 0.40$	9.50(4.36)	40.65%	0.00	0.980
LSCAT	$\max \hat{\boldsymbol{\pi}}_{+} > 0.97$	15.07(0.79)	5.81%	0.00	0.999
	$\max \hat{\boldsymbol{\pi}}_{+} > 0.75$	12.01(1.88)	24.96%	–0.01	0.993
	$\max \hat{\boldsymbol{\pi}}_{+} > 0.50$	9.68(2.36)	39.48%	0.00	0.978

Note. Efficiency = the percentage of unadministered items; SE = standard error.

To further inspect the difference in the distribution of the number of items used, the relationship between response burden and the true z-score was studied under the three stopping rules. Figure 4 shows smoothed trends across participants and, although the overall average response burden (i.e., number of

administered items) was aligned between CAT types, the shape of the distributions was very different. For LSCAT, more items were required in the center of the z-score scale, whereas for IRT-CAT more items were required in the extreme values. The different colors and line types in Figure 4 represent the distinct stopping rules that were considered in Study 3 within each CAT method. The same color (and line type) between CAT methods represent the matching rule that yielded a similar average number of administered items. The curves follow a locally estimated scatterplot smoothing (LOESS) function, smoothing the pattern over participants. Notably, due to the type of smoothing of the LOESS function, the curves in the top figure might exceed the available data points, indicating a potential trend beyond the range of the available number of items required.

Figure 4. Response Burden as a Function of Respondents' *z*-score



Note. The plot shows the average number of items required as a function of true z-value in the x-axis, comparing the results of IRT-CAT (top figure), and LSCAT (bottom figure).

Conclusions

This study compared a complete CAT application that uses LSCAT with a conventional IRT-CAT application. This comparison involved post-hoc simulations using empirical data, where the efficiency in terms of response burden and precision between CAT outcomes and the full-item test were found to be almost identical. Moreover, it was found that in IRT-CAT, respondents with true z-scores closer to extremes required more items on average, whereas LSCAT required more items for respondents closer to the center of the z-score scale (see Figure 4).

Discussion and Conclusions

The current research built upon the recently proposed FlexCAT framework (Van der Ark & Smits, 2023), in which CAT is decomposed into an *engine* and a *score*, and the choice of engine and score depends on the measurement or prediction problem at hand. In this framework, a flexible CAT application was explored using ULCM as the engine. It allows for either item scores (LRCAT) or total scores (LSCAT) as the scoring option. The paper demonstrates that the ULCM can accurately estimate both the joint item-score density ($\boldsymbol{\pi}$) and the total-score density ($\boldsymbol{\pi}_+$).

Based on the results of Study 1, BIC outperformed other competing information criteria AIC, aBIC, and AIC3 in selecting the number of latent classes that yield the most accurate estimation of $\boldsymbol{\pi}$. This result is partially in line with previous simulated studies that considered many latent classes and large sample sizes (e.g., Morgan, 2015). BIC, in principle, tends to impose a larger penalty for model complexity when compared to other criteria and it is consistent; if the true model exists within the set of considered models, the BIC will always select this model as the number of observations tends to infinity (Visser & Speekenbrink, 2022). This characteristic can be advantageous in scenarios involving less restrictive LCMs, as an increasing number of latent classes tends to make them more similar to each other. For estimating $\boldsymbol{\pi}_+$, all competing information criteria (AIC, AIC3, BIC, aBIC) performed equally well in selecting the model that produced the most accurate estimate. This result can be explained by the transformation of $\boldsymbol{\pi}$ to $\boldsymbol{\pi}_+$ using Equation 1, which effectively cancels out estimation errors at the response-pattern level, resulting in a more aggregated outcome. Future research might investigate whether selecting the number of latent classes using resampling methods, such as the bootstrapped likelihood ratio test (Peel & McLachlan, 2000) or the Vuong-Lo-Mendell-Rubin likelihood ratio test (Lo et al., 2001), leads to even more accurate estimate of $\boldsymbol{\pi}$ or $\boldsymbol{\pi}_+$. These resampling methods are computationally demanding and, therefore, were not considered in the current study.

In Study 2, it was demonstrated that the ULCM is able to estimate $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_+$ relatively accurately even if the data-generating mechanism involves higher-order interaction effects among item scores. Notably, the estimated densities were more accurate than the estimates obtained by the two-parameter logistic model, which is arguably the most popular engine in CAT. These findings suggest that in situations where the pattern underlying the data is unknown or complex, less restrictive models, such as the ULCM, might be needed to calibrate a possible CAT. It is important to emphasize, though, that the results of Study 2 do not disprove IRT-CAT. In situations where the data align with a unidimensional scale or exhibit a clear pattern of specific dimensions, IRT-CAT might indeed yield almost perfect accuracy. However, it is important to consider the stringent assumptions of IRT models for item calibration, especially in recent applications where different data sources are combined and might yield a multimodal pattern. When these assumptions are not met, LSCAT can be a valuable alternative with benefits that extend beyond those investigated in this paper.

Over and beyond the results of the first two studies, a proof-of-principle study (Study 3) was also conducted that demonstrated that the LSCAT precision in a CAT simulation using empirical data was nearly identical to an IRT-CAT simulation. This involved comparing LSCAT and a standard IRT-CAT method, focusing on their precision relative to the full test score from the post-hoc simulation. Adjusted for efficiency, the results showed that both methods achieved precision comparable to the full test score. These findings show that LSCAT is a useful alternative to IRT-CAT, offering also some benefits. For example, LSCAT was found to be more efficient than IRT-CAT close to the extremes of the respondents' abilities, demonstrating a potential advantage of the proposed method in situations where IRT-CAT has difficulty in accurately measuring respondents with either very low or very high ability traits. However, as this proof-of-principle study relied solely on the SAQI dataset, future research should further explore LSCAT's capabilities.

This paper is one of the first studies—along with Van der Ark and Smits (2023)—showing how effectively a ULCM can be used as a density estimation tool for item scores or the total score in a CAT application. While LSCAT offers several benefits, it is important to also consider its drawbacks. Although LSCAT offers the advantages of prediction and flexibility in handling different item categories, its computational demands raise concerns. The ULCM requires considering the entire dataset to estimate the probability of each item-response pattern, making it computationally intensive. The personal computer used in this study was limited to handling 20 items per simulation condition. Future research should aim to accommodate more items by exploring more efficient estimation techniques. Furthermore, while the current research focused on dichotomous items, introducing a third response category would result in an exponential increase in the number of potential item-response patterns, further exacerbating the computational burden. This issue, often referred to as "the curse of dimensionality," hinders the practical applications of LSCAT, necessitating further investigation.

Nevertheless, this paper's results add to research exploring non-IRT CAT methods (e.g., Rodrıguez-Cuadrado et al., 2020; Ueno & Songmuang, 2010; Van Buuren & Eggen, 2017; Yan et al., 1998, 2004), by investigating LSCAT. The unique features of LSCAT have the potential to contribute significantly to the CAT field, aligning with the ongoing advancements in technology as highlighted in recent research (Veldkamp, 2022; von Davier, Di Cerbo, et al., 2021).

References

Akaike, H. (1998). A new look at the statistical model identification. In E. Parzen,K. Tanabe, & G. Kitagawa (Eds.), *Selected papers of Hirotugu Akaike* (pp. 215–222). Springer. DOI

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. DOI

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-fisher information matrix. In O. Opitz, B. Lausen, & R. Klar (Eds.), *Information and classification* (pp. 40–54). Springer. DOI

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. DOI

Chalmers, R. P. (2016). Generating adaptive and non-adaptive test interfaces for multidimensional item response theory applications. *Journal of Statistical Software*, *71*(5), 1–39. DOI

Chen, Y., Moustaki, I., & Zhang, H. (2020). A note on Likelihood Ratio Tests for models with latent variables. *Psychometrika*, *85*(4), 996–1012. DOI

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619–632. DOI

Choi, S. W., Reise, S. P., Pilkonis, P. A., Hays, R. D., & Cella, D. (2010). Efficiency of static and computer adaptive short forms compared to full-length measures of depressive symptoms. *Quality of Life Research*, *19*(1), 125–136. DOI

Clark, C. L. (1976). *Proceedings of the first conference on computerized adaptive testing* (ED126110).WebLink

Conrad, L. (1977). *Graduate record examinations. Technical manual* (ED163085). ERIC. WebLink

Drabinová, A., & Martinková, P. (2017). Detection of differential item functioning with nonlinear regression: A non-irt approach accounting for guessing. *Journal of Educational Measurement*, *54*(4), 498–517. DOI

Ebesutani, C., Bernstein, A., Martinez, J. I., Chorpita, B. F., & Weisz, J. R. (2011). The youth self report: Applicability and validity across younger and older youths. *Journal of Clinical Child Adolescent Psychology*, *40*(2), 338–346. DOI

Eggen, T. J. H. M., & Verschoor, A. J. (2006). Optimal testing with easy or difficult items in computerized adaptive testing. *Applied Psychological Measurement*, *30* (5), 379–393. DOI

Finkelman, M. D., Lowe, S. R., Kim, W., Gruebner, O., Smits, N., & Galea, S. (2017). Customized computer-based administration of the PCL-5 for the efficient assessment of PTSD: A proof-of-principle study. *Psychological Trauma: Theory, Research, Practice, and Policy*, *9*(3), 379–389. DOI

Flens, G., Smits, N., Terwee, C. B., Dekker, J., Huijbrechts, I., & de Beurs, E. (2017). Development of a computer adaptive test for depression based on the dutch-flemish version of the PROMIS item bank. *Evaluation & the Health Professions*, *40*(1), 79–105. DOI

Forbey, J. D., Handel, R. W., & Ben-Porath, Y. S. (2000). A real data simulation of computerized adaptive administration of the MMPI-A. *Computers in Human Behavior*, *16*(1), 83–96. DOI

Gibbons, R. D., Weiss, D. J., Pilkonis, P. A., Frank, E., Moore, T., Kim, J. B., & Kupfer, D. J. (2012). Development of a computerized adaptive test for depression. *Archives of General Psychiatry*, *69*(11), 1104–1112. DOI

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

Holt, J. A., & Macready, G. B. (1989). A simulation study of the difference chi-square statistic for comparing latent class models under violation of regularity conditions. *Applied Psychological Measurement*, *13*(3), 221–231. DOI

Jelínek, M., Květon, P., Burešová, I., & Klimusová, H. (2021). Measuring depression in adolescence: Evaluation of a hierarchical factor model of the Children's Depression Inventory and measurement invariance across boys and girls. *PLoS One*, *16*(4). DOI

Kim, S. (2014). World Health Organization Quality of Life (WHOQOL) assessment. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research* (pp. 7260–7261). Springer. DOI

Koehler, K. J., & Larntz, K. (1980). An empirical investigation of goodness-of-fit statistics for sparse multinomials. *Journal of the American Statistical Association*, *75*(370), 336–344. DOI

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. DOI

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: implications for graduate student selection and performance. *Psychological Bulletin*, *127*(1), 162–181. DOI

Linzer, D. A. (2011). Reliable inference in highly stratified contingency tables: Using latent class models as density estimators. *Political Analysis*, *19*(2), 173–187. DOI

Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of Statistical Software*, *42*(10), 1–29. DOI

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767–778. DOI

Lord, F. M. (1969). Robbins-Monro procedures for tailored testing. *ETS Research Bulletin Series*, *1969*(1), i–29. DOI

Lukočienė, O., & Vermunt, J. K. (2010). Determining the number of components in mixture models for hierarchical data. In A. Fink, B. Lausen, W. Seidel, & A. Ultsch (Eds.), *Advances in data analysis, data handling and business intelligence* (pp. 241–249). Springer.

Magis, D., Yan, D., & von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R*. Springer. DOI

Morgan, G. B. (2015). Mixed mode latent class analysis: An examination of fit index performance for classification. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(1), 76–86. DOI

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. DOI

Peel, D., & McLachlan, G. (2000). Ml fitting of mixture models. In *Finite mixture models* (pp. 40–80). Wiley. DOI https://doi.org/10.1002/0471721182.ch

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. WebLink

Reckase, M. (2009). *Multidimensional item response theory*. Springer. DOI

Rodrıguez-Cuadrado, J., Delgado-Gómez, D., Laria, J. C., & Rodrıguez-Cuadrado, S. (2020). Merged Tree-CAT: A fast method for building precise computerized adaptive tests based on decision trees. *Expert Systems with Applications*, *143*(1), 113066. DOI

Şahin, A., & Weiss, D. J. (2015). Effects of calibration sample size and item bank size on ability estimation in computerized adaptive testing. *Educational Sciences: Theory and Practice*, *15*(6).

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

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343. DOI

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*(2), 331–354. DOI

Smits, N., Paap, M. C. S., & Böhnke, J. R. (2018). Some recommendations for developing multidimensional computerized adaptive tests for patient-reported outcomes. *Quality of Life Research*, *27*(4), 1055–1063. DOI

Sorrel, M. A., Nájera, P., & Abad, F. J. (2021). cdcatR: An R package for cognitive diagnostic computerized adaptive testing. *Psych*, *3*(3), 386–403. DOI

Thompson, N. A. (2019). Termination criteria for computerized classification testing. *Practical Assessment, Research Evaluation*, *16*(4), 1–7. DOI

Ueno, M., & Songmuang, P. (2010). Computerized adaptive testing based on decision tree. In *2010 10th IEEE International Conference on Advanced Learning Technologies* (pp. 191–193). IEEE. DOI

Van Buuren, N., & Eggen, T. H. J. M. (2017). Latent-class-based item selection for computerized adaptive progress tests. *Journal of Computerized Adaptive Testing*, *5*(2), 22–43. DOI

Van der Ark, L. A., & Smits, N. (2023). Computerized adaptive testing without IRT for flexible measurement and prediction. In L. A. Van der Ark, W. H. M. Emons, & R. R. Meijer (Eds.), *Essays on contemporary psychometrics* (pp. 369–388). Springer. DOI

Van der Ark, L. A., van der Palm, D. W., & Sijtsma, K. (2011). A latent class approach to estimating test-score reliability. *Applied Psychological Measurement*, *35*(5), 380–392. DOI

Van der Linden, W. J. (2018). *Handbook of item response theory*. CRC Press. DOI

Van der Linden, W. J., & Glas, C. A. W. (Eds.). (2010). *Elements of adaptive testing*. Springer. DOI

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016a). A comparison of incomplete-data methods for categorical data. *Statistical Methods in Medical Research*, *25*(2), 754–774. DOI

Van der Palm, D. W., Van der Ark, L. A., & Vermunt, J. K. (2016b). Divisive latent class modeling as a density estimation method for categorical data. *Journal of Classification*, *33*(1), 52–72. DOI

Veldkamp, B. P. (2022). *The double helix of adaptive measurement* [Keynote presentation]. 8th Conference of the International Association for Computerized Adaptive Testing, Frankfurt am Main, Germany. WebLink

Vermunt, J. K., & Magidson, J. (2004). Latent class analysis. In M. Lewis-Beck, A. Bryman, & T. Liao (Eds.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 549–553). Sage.

Vermunt, J. K., Van Ginkel, J. R., Van der Ark, L. A., & Sijtsma, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology*, *38*(1), 369–397. DOI

Visser, I., & Speekenbrink, M. (2022). Mixture and latent class models. In I. Visser & M. Speekenbrink (Eds.), *Mixture and hidden markov models with R* (pp. 45–93). Springer. DOI

von Davier, A. A., Di Cerbo, K., & Verhagen, J. (2021). Computational psychometrics: A framework for estimating learners' knowledge, skills and abilities from learning and assessments systems. In A. A. von Davier, R. J. Mislevy, & J. Hao (Eds.), *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python* (pp. 25–43). Springer. DOI

von Davier, A. A., Mislevy, R. J., & Hao, J. (Eds.). (2021). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in r and python*. Springer. DOI

von Davier, M., & Lee, Y.-S. (Eds.). (2019). *Handbook of diagnostic classification models: Models and model extensions, applications, software packages*. Springer. DOI

Vorst, H. C. M. (2006). School attitude questionnaire - internet (SAQI). *Uitgeverij Libbe Mulder*. WebLink

Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing: A primer* (2nd ed). Erlbaum.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 70–84. DOI

Whittaker, T. A., & Miller, J. E. (2021). Exploring the enumeration accuracy of cross-validation indices in latent class analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(3), 376–390. DOI

Yan, D., Lewis, C., & Stocking, M. (1998). *Adaptive testing without irt* (ED422359). ERIC. WebLink

Yan, D., Lewis, C., & Stocking, M. (2004). Adaptive testing with regression trees in the presence of multidimensionality. *Journal of Educational and Behavioral Statistics*, *29*(3), 293–316. DOI

Author Address

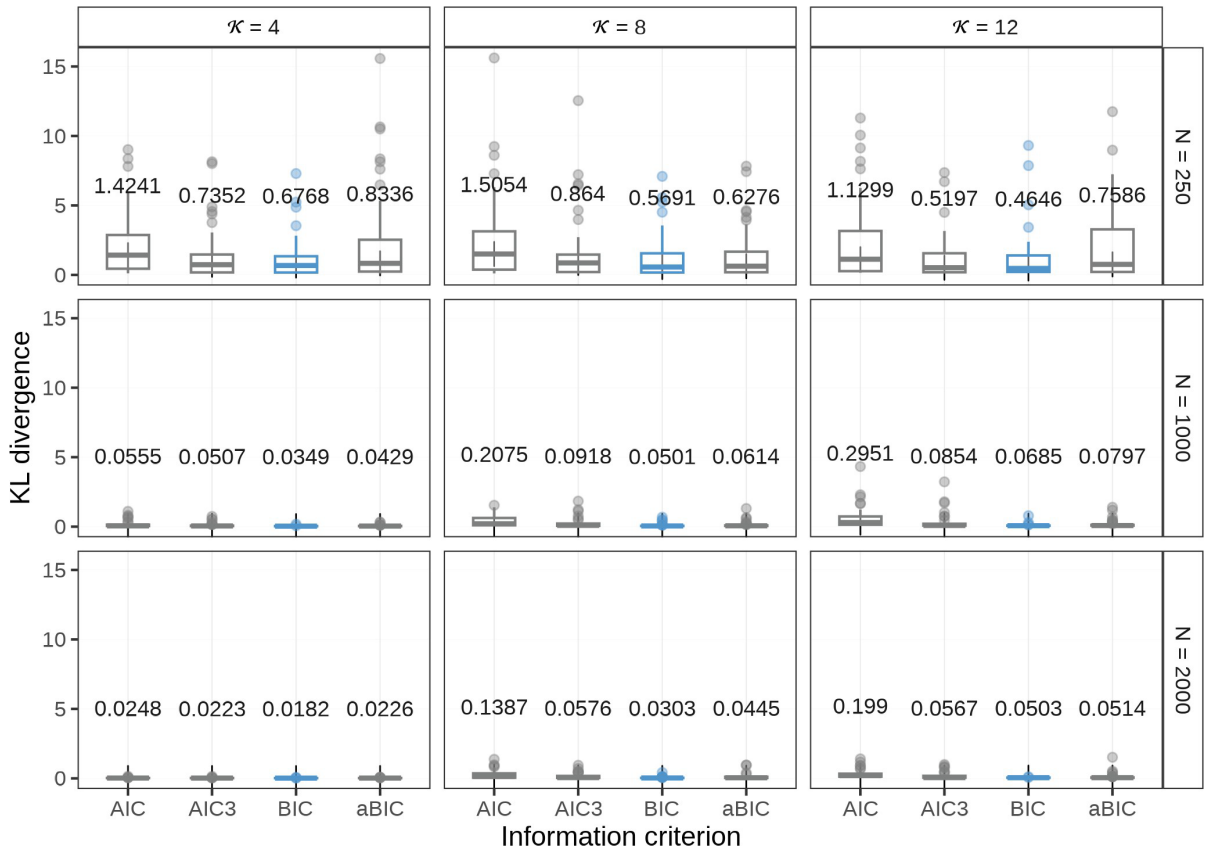
a.psychogyiopoulos@uva.nl

Citation

Psychogyiopoulos, A., Smits, N., & van der Ark, L. A. (2025).
Estimating the joint item-score density using an unrestricted latent class model:
Advancing flexibility in computerized adaptive testing.
Journal of Computerized Adaptive Testing, 12(3), 136-164. DOI 10.7333/2507-1203136

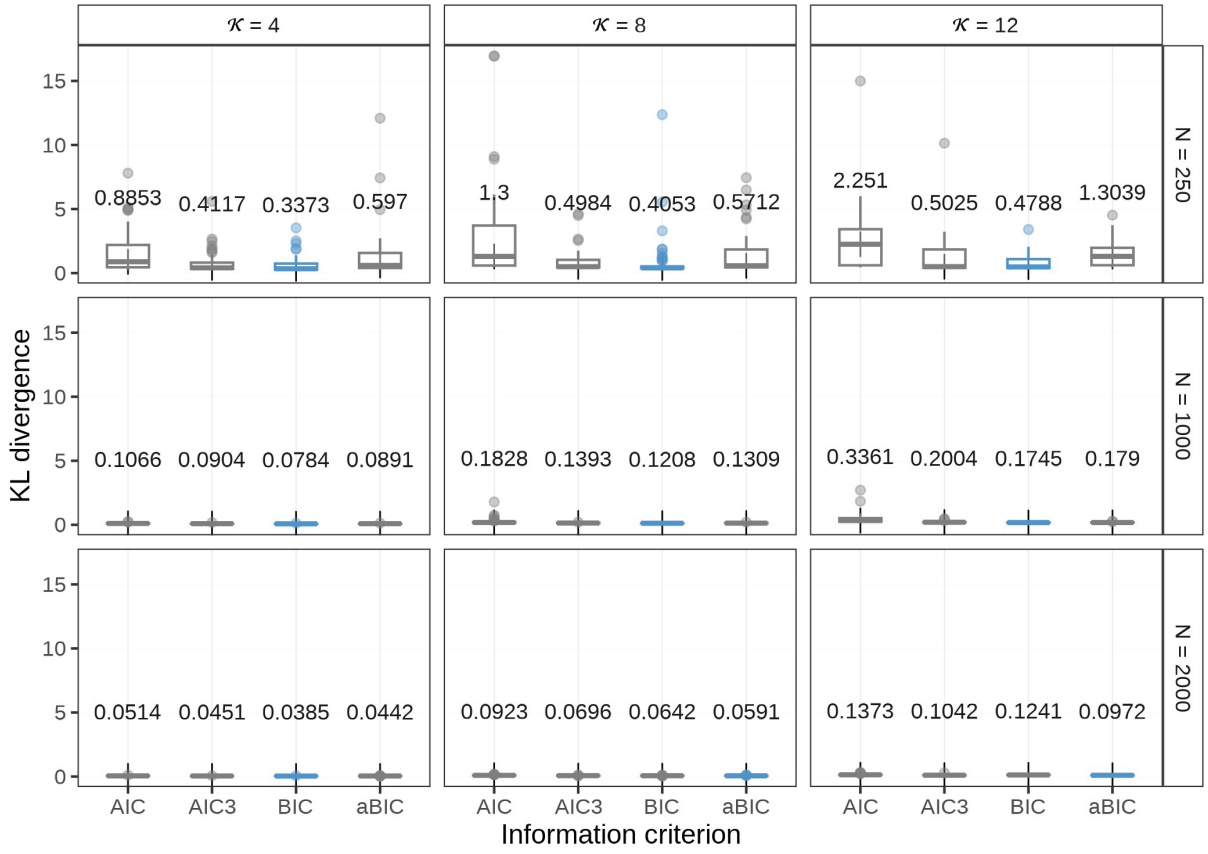
Appendix A: Study 1 Results

Figure A1. Boxplots of Kullback-Leibler Divergence for π Across Four Information Criteria (Within Cells), Three Sample Sizes (Rows), and Three Item Counts (Columns), for J = 9



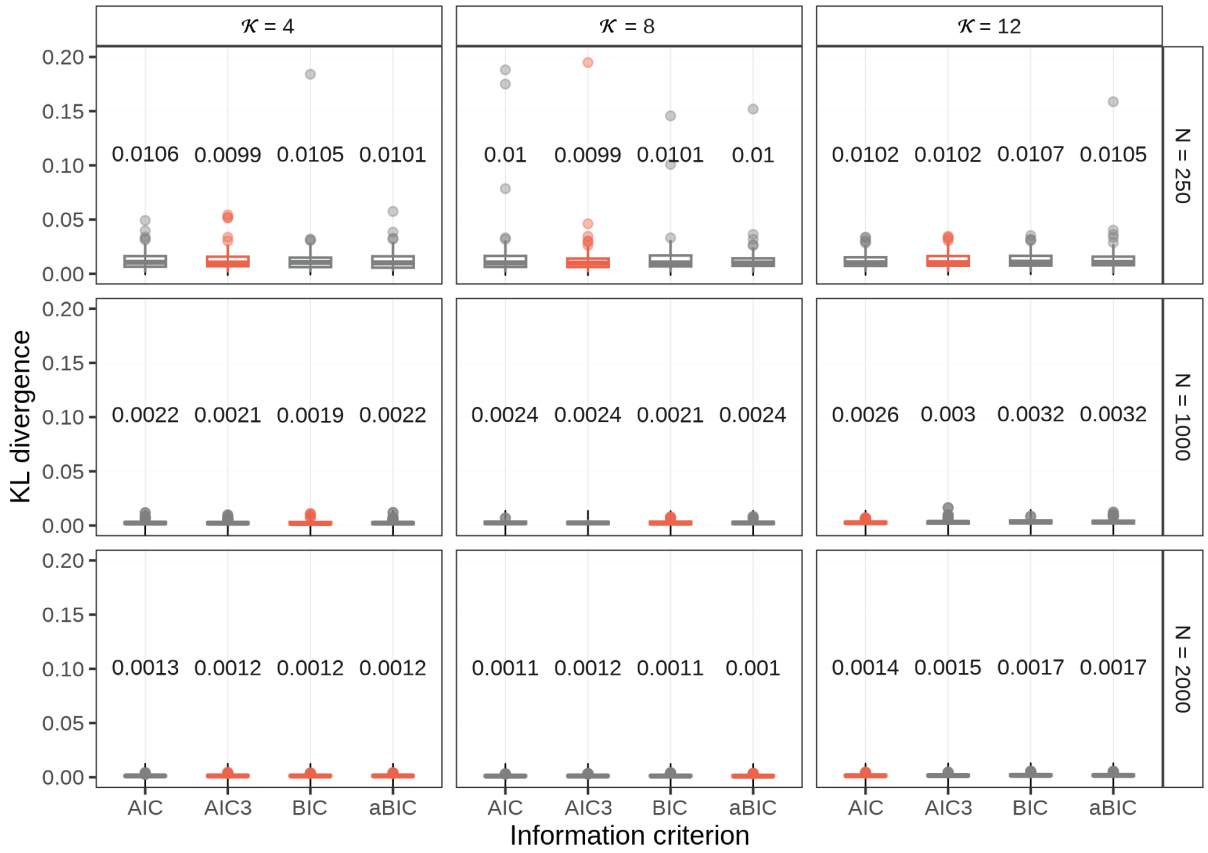
Note. The boxplots show the dispersion of KL divergence values between the simulation conditions of different true number of latent classes ($\mathcal{K}$) and the different sample sizes (N). Within each condition (cell), the spread of KL values is divided by each competing information criterion, indicating which showed on average better performance (i.e., lower KL values). The median within each condition for each information criterion is also shown numerically. The blue color indicates the information criterion that yielded the best estimate of **π**.

Figure A2. Boxplots of Kullback-Leibler Divergence for π Across Four Information Criteria (Within Cells), Three Sample Sizes (Rows), and Three Item Counts (Columns), for J = 18



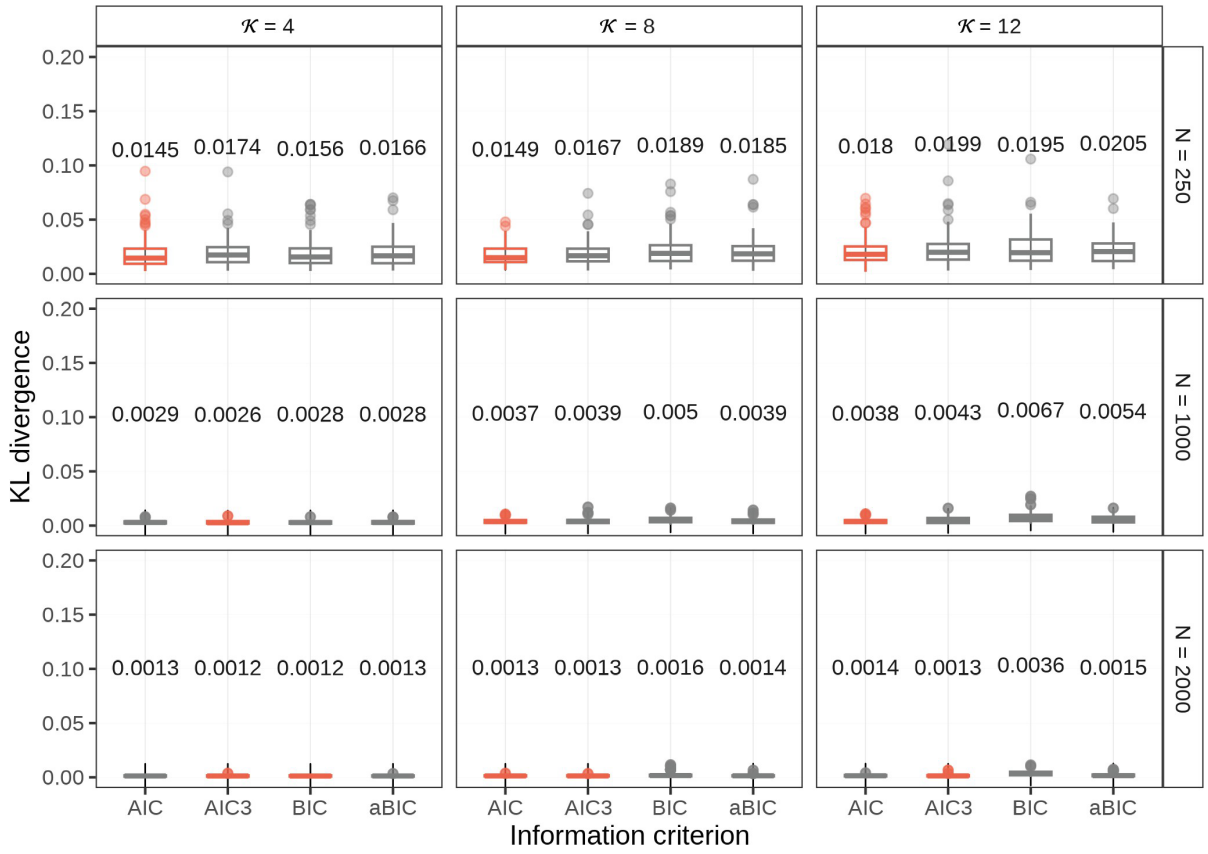
Note. The boxplots show the dispersion of KL divergence values between the simulation conditions of different true number of latent classes ($\mathcal{K}$) and the different sample sizes (N). Within each condition (cell), the spread of KL values is divided by each competing information criterion, indicating which showed on average better performance (i.e., lower KL values). The median within each condition for each information criterion is also shown numerically. The blue color indicates the information criterion that yielded the best estimate of .

Figure A3. Boxplots of Kullback-Leibler Divergence for $\boldsymbol{\pi}_+$ Across Four Information Criteria (Within Cells), Three Sample Sizes (Rows), and Three Item Counts (Columns), for J = 9



Note. The boxplots show the dispersion of KL divergence values between the simulation conditions of different true number of latent classe33s ($\mathcal{K}$) and the different sample sizes (N). Within each condition (cell), the spread of KL values is divided by each competing information criterion, indicating which showed on average better performance (i.e., lower KL values). The median within each condition for each information criterion is also shown numerically. The red color indicates the information criteria that yielded the best estimate of $\boldsymbol{\pi}_+$.

Figure A4. Boxplots of Kullback-Leibler Divergence for $\boldsymbol{\pi}_+$ Across Four Information Criteria (Within Cells), Three Sample Sizes (Rows), and Three Item Counts (Columns), for J = 18



Note. The boxplots show the dispersion of KL divergence values between the simulation conditions of different true number of latent classes ($\mathcal{K}$) and the different sample sizes (N). Within each condition (cell), the spread of KL values is divided by each competing information criterion, indicating which showed on average better performance (i.e., lower KL values). The median within each condition for each information criterion is also shown numerically. The red color indicates the information criteria that yielded the best estimate of $\boldsymbol{\pi}_+$.

Appendix B: Population Model Parameters

Table B1. MSAT-B Items and Item Parameters Used as Population Values to Generate Data Under a 2PLM in Study 2

Item	α	β	γ	δ
Item 49	1.06	-2.02	0	1
Item 27	1.08	1.29	0	1
Item 41	1.50	0.44	0	1
Item 7	1.04	0.38	0	1
Item 38	1.42	-0.80	0	1
Item 28	0.88	1.94	0	1
Item 9	0.96	-1.07	0	1
Item 47	0.95	-2.99	0	1
Item 75	0.92	-0.25	0	1

Note: The default settings in R package `MIRT` were used for the 2PLM parameterization. MSAT-B = Czech Medical School Admission Test – Biology; β = difficulty parameter; γ = lower asymptote; δ = upper asymptote.

Table B2. Response Probabilities Used as Population Parameters to Generate Data Under a ULCM(2) using the MSAT-B Dataset in Study 2

Class(ξ)	Response	
	$y = 0$	$y = 1$
	Item 49	
$\xi = 1$	0.22	0.78
$\xi = 2$	0.02	0.98
	Item 27	
$\xi = 1$	0.89	0.11
$\xi = 2$	0.55	0.45
	Item 41	
$\xi = 1$	0.82	0.18
$\xi = 2$	0.29	0.71
	Item 7	
$\xi = 1$	0.74	0.26
$\xi = 2$	0.32	0.68
	Item 38	
$\xi = 1$	0.46	0.54
$\xi = 2$	0.05	0.95
	Item 28	
$\xi = 1$	0.90	0.10
$\xi = 2$	0.68	0.32
	Item 9	
$\xi = 1$	0.42	0.58
$\xi = 2$	0.10	0.90
	Item 47	
$\xi = 1$	0.11	0.89
$\xi = 2$	0.02	0.98
	Item 75	
$\xi = 1$	0.60	0.40
$\xi = 2$	0.21	0.79

Note: The response probability refers to the probability of giving a certain response ($y \in (0, 1)$) conditional on the class membership. ULCM(2) = unrestricted latent class model with two latent classes.

Table B3. Class Weights Used as Population Parameters to Generate Data Under a ULCM(2) using the MSAT-B Dataset in Study 2

$P(\Xi = \xi)$
$P(\Xi = 1) = 0.62$
$P(\Xi = 2) = 0.38$

Note: The probabilities corresponding to class weights indicate the proportion of respondents belonging to each class.

Table B4. The Population Parameters of the Loglinear Model Used to Generate Data with Fifth-Order Interaction Effects in Study 2

λ_0									
$\lambda_0 = 0.32$									
λ_i									
$\lambda_1 = -2.64$	$\lambda_3 = -1.77$	$\lambda_6 = -1.37$	$\lambda_7 = -1.90$	$\lambda_9 = -1.33$					
$\lambda_{i,j}$									
$\lambda_{1,2} = -3.92$	$\lambda_{1,3} = -2.26$	$\lambda_{1,4} = -0.35$	$\lambda_{1,5} = -5.00$	$\lambda_{1,6} = -2.02$	$\lambda_{1,7} = -0.63$	$\lambda_{1,8} = -2.92$	$\lambda_{2,3} = -2.98$	$\lambda_{2,4} = -3.75$	$\lambda_{2,5} = -1.49$
$\lambda_{2,6} = -2.03$	$\lambda_{2,7} = -2.87$	$\lambda_{2,9} = -2.40$	$\lambda_{3,4} = -1.91$	$\lambda_{3,5} = -2.39$	$\lambda_{3,6} = -1.83$	$\lambda_{3,7} = -3.20$	$\lambda_{3,8} = -3.27$	$\lambda_{3,9} = -0.68$	$\lambda_{4,5} = -1.93$
$\lambda_{4,6} = -2.17$	$\lambda_{4,7} = -4.44$	$\lambda_{4,8} = -2.55$	$\lambda_{4,9} = -2.00$	$\lambda_{5,7} = -2.91$	$\lambda_{6,7} = -2.41$	$\lambda_{6,8} = -2.72$	$\lambda_{6,9} = -0.45$	$\lambda_{7,8} = -2.53$	$\lambda_{7,9} = -3.83$
$\lambda_{8,9} = -2.56$									
$\lambda_{i,j,k}$									
$\lambda_{1,2,5} = 2.95$	$\lambda_{1,3,5} = 2.44$	$\lambda_{1,3,7} = 2.50$	$\lambda_{1,4,5} = 2.9$	$\lambda_{1,5,7} = 2.89$	$\lambda_{1,5,8} = 3.40$	$\lambda_{1,5,9} = 3.36$	$\lambda_{1,6,8} = 2.16$	$\lambda_{1,6,9} = 1.04$	$\lambda_{1,7,8} = 2.81$
$\lambda_{1,7,9} = 2.11$	$\lambda_{2,3,5} = 2.33$	$\lambda_{2,3,9} = 4.27$	$\lambda_{2,4,5} = 1.14$	$\lambda_{2,4,9} = 3.25$	$\lambda_{2,5,7} = 2.51$	$\lambda_{2,5,8} = 2.92$	$\lambda_{2,6,9} = 1.65$	$\lambda_{2,7,9} = 1.70$	$\lambda_{3,4,7} = 1.95$
$\lambda_{3,4,9} = 3.29$	$\lambda_{3,5,8} = 3.12$	$\lambda_{3,7,8} = 1.46$	$\lambda_{3,8,9} = 2.89$	$\lambda_{4,5,6} = 2.33$	$\lambda_{4,5,7} = 3.48$	$\lambda_{4,6,7} = 3.50$	$\lambda_{4,6,8} = 3.00$	$\lambda_{4,6,9} = 0.74$	$\lambda_{4,7,8} = 3.59$
$\lambda_{4,7,9} = 0.98$	$\lambda_{5,6,7} = 2.66$	$\lambda_{5,7,8} = 3.36$	$\lambda_{5,8,9} = 1.81$	$\lambda_{6,7,8} = 2.72$	$\lambda_{6,7,9} = 1.82$	$\lambda_{6,8,9} = 1.52$	$\lambda_{7,8,9} = 2.62$		
$\lambda_{i,j,k,l}$									
$\lambda_{1,2,3,4} = 0.52$	$\lambda_{1,2,3,5} = 0.10$	$\lambda_{1,2,3,6} = 0.57$	$\lambda_{1,2,3,8} = 1.17$	$\lambda_{1,2,3,9} = 1.36$	$\lambda_{1,2,4,5} = -1.96$	$\lambda_{1,2,4,6} = 0.74$	$\lambda_{1,2,4,7} = -0.59$	$\lambda_{1,2,4,8} = 1.52$	$\lambda_{1,2,4,9} = 0.24$
$\lambda_{1,2,5,6} = 0.24$	$\lambda_{1,2,5,7} = -0.09$	$\lambda_{1,2,5,8} = -1.05$	$\lambda_{1,2,6,7} = 0.19$	$\lambda_{1,2,7,8} = 0.78$	$\lambda_{1,2,7,9} = -1.62$	$\lambda_{1,3,4,5} = -0.54$	$\lambda_{1,3,4,6} = 0.69$	$\lambda_{1,3,4,7} = 0.53$	$\lambda_{1,3,4,9} = -0.49$
$\lambda_{1,3,5,6} = -1.09$	$\lambda_{1,3,5,7} = 0.33$	$\lambda_{1,3,5,8} = -0.63$	$\lambda_{1,3,5,9} = -0.04$	$\lambda_{1,3,6,7} = 0.55$	$\lambda_{1,3,6,8} = 1.37$	$\lambda_{1,3,6,9} = -0.17$	$\lambda_{1,3,8,9} = -0.24$	$\lambda_{1,4,5,7} = -1.14$	$\lambda_{1,4,6,7} = -1.11$
$\lambda_{1,4,6,8} = -0.46$	$\lambda_{1,4,6,9} = -0.12$	$\lambda_{1,4,7,8} = -0.69$	$\lambda_{1,4,7,9} = 0.86$	$\lambda_{1,4,8,9} = -1.53$	$\lambda_{1,5,6,7} = -1.80$	$\lambda_{1,5,6,8} = 0.04$	$\lambda_{1,5,7,8} = 0.54$	$\lambda_{1,5,7,9} = -0.31$	$\lambda_{1,6,7,8} = 1.35$
$\lambda_{1,6,8,9} = -1.34$	$\lambda_{1,7,8,9} = -1.08$	$\lambda_{2,3,4,5} = 0.38$	$\lambda_{2,3,4,7} = -0.23$	$\lambda_{2,3,4,8} = 1.44$	$\lambda_{2,3,4,9} = -0.86$	$\lambda_{2,3,5,6} = 1.00$	$\lambda_{2,3,5,7} = -0.87$	$\lambda_{2,3,5,9} = 0.36$	$\lambda_{2,3,6,7} = -0.91$
$\lambda_{2,3,6,9} = 1.84$	$\lambda_{2,3,7,8} = 1.23$	$\lambda_{2,3,8,9} = 1.30$	$\lambda_{2,4,5,7} = 0.20$	$\lambda_{2,4,5,8} = 0.32$	$\lambda_{2,4,5,9} = -0.94$	$\lambda_{2,4,6,7} = -1.47$	$\lambda_{2,4,6,9} = -0.70$	$\lambda_{2,5,6,7} = -0.89$	$\lambda_{2,5,6,8} = -0.52$
$\lambda_{2,5,6,9} = -0.73$	$\lambda_{2,5,7,8} = 0.80$	$\lambda_{2,5,7,9} = 0.90$	$\lambda_{2,5,8,9} = -0.86$	$\lambda_{2,6,7,8} = 0.60$	$\lambda_{2,6,7,9} = -0.36$	$\lambda_{2,6,8,9} = -0.26$	$\lambda_{2,7,8,9} = -0.49$	$\lambda_{3,4,5,7} = -0.22$	$\lambda_{3,4,5,8} = 0.20$
$\lambda_{3,4,5,9} = -0.31$	$\lambda_{3,4,6,7} = 0.02$	$\lambda_{3,4,6,8} = -0.65$	$\lambda_{3,4,6,9} = -0.27$	$\lambda_{3,4,7,8} = -0.63$	$\lambda_{3,4,7,9} = -0.62$	$\lambda_{3,4,8,9} = -0.09$	$\lambda_{3,5,6,7} = -3.63$	$\lambda_{3,5,6,8} = -1.14$	$\lambda_{3,5,6,9} = -1.72$
$\lambda_{3,5,7,8} = -0.28$	$\lambda_{3,5,7,9} = 0.41$	$\lambda_{3,5,8,9} = 0.69$	$\lambda_{3,6,7,8} = -0.19$	$\lambda_{3,7,8,9} = 0.48$	$\lambda_{4,5,6,8} = -0.14$	$\lambda_{4,5,6,9} = -1.30$	$\lambda_{4,5,7,8} = -1.57$	$\lambda_{4,5,7,9} = 0.74$	$\lambda_{4,5,8,9} = 0.38$
$\lambda_{4,6,7,8} = -1.64$	$\lambda_{4,6,7,9} = -1.5$	$\lambda_{4,6,8,9} = -0.80$	$\lambda_{4,7,8,9} = -0.08$	$\lambda_{5,6,7,8} = 1.53$	$\lambda_{5,6,7,9} = 0.89$	$\lambda_{5,6,8,9} = -2.18$	$\lambda_{6,7,8,9} = 0.26$		
$\lambda_{i,j,k,l,m}$									
$\lambda_{1,2,3,4,8} = -2.00$	$\lambda_{1,2,3,5,6} = -2.00$	$\lambda_{1,2,3,5,7} = -2.00$	$\lambda_{1,2,3,5,8} = -2.00$	$\lambda_{1,2,3,6,8} = -2.00$	$\lambda_{1,2,3,6,9} = -2.00$	$\lambda_{1,2,3,8,9} = -2.00$	$\lambda_{1,2,4,5,6} = -2.00$	$\lambda_{1,2,4,5,7} = -2.00$	$\lambda_{1,2,4,5,8} = -2.00$
$\lambda_{1,2,4,6,7} = -2.00$	$\lambda_{1,2,4,6,8} = -2.00$	$\lambda_{1,2,4,6,9} = -2.00$	$\lambda_{1,2,4,8,9} = -2.00$	$\lambda_{1,2,5,6,7} = -2.00$	$\lambda_{1,2,5,6,8} = -2.00$	$\lambda_{1,2,5,8,9} = -2.00$	$\lambda_{1,2,6,7,8} = -2.00$	$\lambda_{1,2,7,8,9} = -2.00$	$\lambda_{1,3,4,5,8} = -2.00$
$\lambda_{1,3,4,5,9} = -2.00$	$\lambda_{1,3,4,6,7} = -2.00$	$\lambda_{1,3,4,6,8} = -2.00$	$\lambda_{1,3,5,6,7} = -2.00$	$\lambda_{1,3,5,6,9} = -2.00$	$\lambda_{1,3,6,7,9} = -2.00$	$\lambda_{1,3,7,8,9} = -2.00$	$\lambda_{1,4,5,6,7} = -2.00$	$\lambda_{1,4,5,7,8} = -2.00$	$\lambda_{1,4,5,7,9} = -2.00$
$\lambda_{1,4,5,8,9} = -2.00$	$\lambda_{1,4,6,8,9} = -2.00$	$\lambda_{1,4,7,8,9} = -2.00$	$\lambda_{2,3,4,5,6} = -2.00$	$\lambda_{2,3,4,5,7} = -2.00$	$\lambda_{2,3,5,6,7} = -2.00$	$\lambda_{2,3,5,6,8} = -2.00$	$\lambda_{2,3,7,8,9} = -2.00$	$\lambda_{2,4,5,7,8} = -2.00$	$\lambda_{2,4,5,8,9} = -2.00$
$\lambda_{2,5,6,7,9} = -2.00$	$\lambda_{2,6,7,8,9} = -2.00$	$\lambda_{3,4,5,6,8} = -2.00$	$\lambda_{3,4,5,7,9} = -2.00$	$\lambda_{3,4,5,8,9} = -2.00$	$\lambda_{3,5,6,8,9} = -2.00$	$\lambda_{4,5,6,7,9} = -2.00$	$\lambda_{4,5,6,8,9} = -2.00$	$\lambda_{4,5,7,8,9} = -2.00$	$\lambda_{4,6,7,8,9} = -2.00$

Note. See Study 2 for details.