DOI: 10.1017/psy.2026.10086

A Likelihood-Based Profile Shrinkage Algorithm for Efficient Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT)

Xiuxiu Tang

Ying Cheng

University of Notre Dame

Corresponding author: Ying Cheng; Email: ycheng4@nd.edu; 390 Corbett Hall, Notre Dame, IN 46556.

Abstract

Various item selection algorithms have been proposed for Cognitive Diagnostic Computerized Adaptive Testing (CD-CAT), with the goal of efficiently diagnosing examinees' strengths and weaknesses. However, these algorithms often come with significant computational costs, which can hinder their practical implementation. A Likelihood-Based Profile Shrinkage (LBPS) Algorithm is proposed to simplify the item selection process and reduce the computational costs in CD-CAT. Our simulation results indicate that incorporating LBPS into existing item selection methods yields substantial computational efficiency gains, with greater reductions in computation time as the number of attributes and test length increase. Additionally, LBPS maintains estimation accuracy at both the attribute and pattern levels. These findings suggest that LBPS is a scalable and effective solution for the item selection of CD-CAT in complex scenarios.

Keywords: cognitive diagnosis, computerized adaptive testing, item selection method, computational efficiency

1. Introduction

There has been an increasing emphasis in educational assessment on formative evaluation and diagnostic feedback (Black & Wiliam, 2009; Morris et al., 2021). Cognitive diagnostic assessment (CDA) addresses this need by providing detailed information about examinees' mastery of specific skills or attributes (Leighton & Gierl, 2007; Rupp et al., 2010). To improve the efficiency of CDA administration, cognitive diagnostic computerized adaptive testing (CD-CAT) has emerged as a powerful approach that combines the benefits of cognitive diagnosis with the efficiency of adaptive testing (Cheng, 2009).

A critical component in CD-CAT is the item selection algorithm, which determines the items and the sequence in which they are administered to each examinee. Various item selection methods have been proposed, most of which fundamentally connected through information theory principles as demonstrated by Cheng (2009) and Wang et al. (2020). These methods include the original Kullback-Leibler (KL) index (Xu et al., 2003), likelihood-weighted KL and posterior-weighted KL (PWKL) index (Cheng, 2009), modified PWKL (MPWKL) index (Kaplan et al., 2015), the Shannon entropy (SHE) procedure (Tatsuoka, 2002) and mutual information methods (Wang, 2013). While these approaches may appear distinct, they all derive from related information-theoretic concepts, with KL divergence, Shannon entropy, and mutual information sharing deep mathematical connections in quantifying information gain and uncertainty reduction (Cover & Thomas, 1991). More recently, the generalized deterministic inputs, noisy ‘‘and’’ gate (G-DINA) model discrimination index (GDI) was introduced as an efficient alternative (Kaplan et al., 2015). GDI quantifies the weighted variance in item success probabilities given a specific attribute distribution.

While these methods have demonstrated effectiveness in attribute-level classification accuracy, they face a significant computational challenge: as the number of attributes (K) increases, the computational burden grows exponentially. This occurs because most existing methods require evaluating all possible attribute patterns (2^K) for each item in the bank before selecting the most suitable one as the next item to administer. For instance, with K = 10 attributes, algorithms must evaluate 1,024 possible patterns for each candidate item. This computational intensity can make real-time implementation challenging, particularly in settings requiring rapid item selection decisions.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

The only existing method that attempts to reduce the computational burden is the GDI method, which partially addresses this issue by working with reduced attribute patterns. Although GDI is more computationally efficient than the PWKL method (Kaplan et al., 2015), which is known to be the most computationally intensive, its efficiency relative to KL and SHE—two other widely discussed methods in CD-CAT—remains unclear. This study aims to address this gap by evaluating GDI's computational efficiency relative to KL and SHE. Moreover, the primary objective is to propose a novel and flexible approach that not only substantially reduces computational demands but also maintains the theoretical foundations and measurement precision of existing methods.

This paper introduces the Likelihood-Based Profile Shrinkage (LBPS) algorithm as a solution to this challenge. The key insight of LBPS is that as testing proceeds, the set of plausible attribute patterns for an examinee rapidly shrinks based on their response patterns. By focusing on only the most likely attribute patterns, LBPS achieves substantial efficiency gains while preserving measurement accuracy. Importantly, LBPS can be integrated with any existing item selection method, making it a flexible enhancement to current CD-CAT implementations. In addition, LBPS can be implemented without requiring changes to existing item banks or cognitive diagnostic models. Through simulation studies, we demonstrate that LBPS achieves comparable attribute classification accuracy to traditional methods while greatly reducing computation time, particularly for long tests measuring larger numbers of attributes. The remainder of this paper is organized as follows. Section 2 reviews the theoretical framework of cognitive diagnostic models and existing item selection methods. Section 3 introduces the LBPS algorithm and establishes its theoretical properties. Section 4 presents simulation studies comparing LBPS with existing methods across various conditions. Section 5 discusses practical implications and future research directions.

2. Background

2.1 CDM Framework

2.1.1 Basic Setup

Cognitive diagnostic models (CDMs) aim to provide detailed information about examinees' mastery of specific skills or attributes underlying test performance. In CDA, the goal is to measure examinees' mastery of K discrete attributes or skills. Each examinee's mastery

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

profile is represented by an attribute pattern $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_K)$ where $\alpha_k = 1$ indicates mastery of attribute *k* and $\alpha_k = 0$ indicates non-mastery for *k* = 1, 2, …, *K* attributes. Note that the terms 'pattern' and 'profile' are used interchangeably in the paper to refer to **α**. For *K* attributes, there are 2^K possible attribute patterns, representing all possible combinations of mastery and non-mastery across the measured attributes (de la Torre, 2011). The relationship between items and attributes is specified through a J × K Q-matrix (Tatsuoka, 1995), where entry $q_{jk} = 1$ if item *j* requires attribute *k* and $q_{jk} = 0$ otherwise. The Q-matrix represents the cognitive specifications of the test by mapping each item to its required attributes.

2.1.2 Types of CDMs

CDMs can be categorized based on how they model the relationship between attributes and item responses (Ravand & Baghaei, 2020). In conjunctive models, examinees must master all required attributes to have a high probability of correctly answering an item. The DINA model (Junker & Sijtsma, 2001) and the Noisy Inputs, Deterministic "And" Gate (NIDA) model (Maris, 1999) are prominent examples of conjunctive models. These models are particularly appropriate when skills build upon each other in a non-compensatory way. Disjunctive models assume that mastery of any one of the required attributes is sufficient for a high probability of success. The Deterministic Input, Noisy "Or" Gate (DINO) model (Templin & Henson, 2006) exemplifies this approach. Such models are suitable when multiple solution strategies can lead to correct answers and mastery of one attribute can compensate for non-mastery of others.

Additive models, such as the Additive CDM (ACDM; de la Torre, 2011) and the Linear Logistic Model (LLM; Maris, 1999), take a different approach where each mastered attribute contributes independently to the probability of a correct response. These models are appropriate when attributes have cumulative but independent effects on performance.

More recently, general diagnostic models have been developed that can accommodate multiple types of attribute relationships within the same assessment. The Generalized DINA model (G-DINA; de la Torre, 2011), the Log-linear CDM (LCDM; Henson et al., 2009), and the General Diagnostic Model (GDM; von Davier, 2005) allow different items to exhibit different attribute relationships. These general models provide greater flexibility but typically require larger sample sizes for stable parameter estimation.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

2.1.3 The DINA Model

While the methods developed in this paper apply to any CDM, we use the DINA model for illustration due to its parsimony and wide use in diagnostic testing applications (de la Torre, 2009; Junker & Sijtsma, 2001). Under the DINA model, an examinee must master all required attributes to have a high probability of answering an item correctly, making it a conjunctive model. For an examinee *i* with attribute pattern **α** responding to item *j*, the ideal response is:

$$\eta_{ij}(\boldsymbol{\alpha}) = \prod_{k=1}^{K} {\alpha_{ik}}^{q_{jk}} , \tag{1}$$

where $\eta_{ij}(\boldsymbol{\alpha}) = 1$ indicates mastery of all required attributes and $\eta_{ij}(\boldsymbol{\alpha}) = 0$ indicates lack of at least one required attribute (de la Torre, 2009).

The probability of a correct response of examinee *i* on item *j* is given by:

$$P\left(X_{ij} = 1|\boldsymbol{\alpha}\right) = \left(1 - s_j\right)^{\eta_{ij}(\boldsymbol{\alpha})} {g_j}^{1-\eta_{ij}(\boldsymbol{\alpha})}, \tag{2}$$

where s_j is the slipping parameter (probability of incorrect response despite mastery) and g_j is the guessing parameter (probability of correct response despite non-mastery). These item parameters account for the probabilistic nature of the response process, where examinees who have mastered all required attributes may still make mistakes (slips) and those who lack required attributes may still answer correctly through guessing (de la Torre & Douglas, 2004). The DINA model's simple form makes it particularly useful for understanding the fundamental principles of cognitive diagnosis while still capturing essential features of the response process. Its parsimony in parameter estimation and clear interpretation of results has made it a popular choice in diagnostic testing applications.

2.2 Item Selection Methods in CD-CAT

Item selection methods in CD-CAT can be broadly categorized into parametric and nonparametric approaches (Chang et al., 2019). While nonparametric methods have emerged recently to address certain limitations of parametric approaches (Chiu & Chang, 2021), parametric methods remain fundamental to CD-CAT implementation. These parametric methods can be further classified as single-purpose or dual-purpose (Wang et al., 2012). Single-purpose methods focus solely on optimizing the measurement of attribute profiles, while dual-purpose

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

methods simultaneously measure both attribute profiles and general ability (Dai et al., 2016; Kang et al., 2017; Wang et al., 2014). This paper proposes a new algorithm within the framework of parametric single-purpose item selection methods. Therefore, we focus our review on existing methods in this category, which form the foundation for CD-CAT item selection and remain the most widely used in practice.

2.2.1 Basic Framework

In CD-CAT, parametric single-purpose item selection methods aim to optimize the measurement of examinees' attribute mastery profiles. These methods utilize item parameters and probability models within a cognitive diagnostic framework to select items that maximize information about attribute patterns. After J items have been administered to an examinee, let $\boldsymbol{x_J} = (x_1, \ldots, x_J)$ denote the vector of observed responses, where $x_j \in \{0,1\}$. Following Bayes' theorem, the posterior probability of attribute pattern is:

$$\pi(\boldsymbol{\alpha}|\boldsymbol{x_J}) \propto \pi_0(\boldsymbol{\alpha})L(\boldsymbol{x_J}|\boldsymbol{\alpha}), \tag{3}$$

where $\pi_0(\boldsymbol{\alpha})$ is the prior probability and $L(\boldsymbol{x_J}|\boldsymbol{\alpha}) = \prod_{j=1}^{J} P(X_j = x_j|\boldsymbol{\alpha})$ is the likelihood function under the specified CDM. Define the item h as a candidate item in the pool of available items, from which the (J+1)-th item is to be selected based on a specified item selection method.

2.2.2 Information-Theoretic Methods

Kullback-Leibler Based Approaches

The Kullback-Leibler (KL) information (Cover & Thomas, 1991; Kullback & Leibler, 1951) provides a foundation for measuring the distance between probability distributions under different attribute patterns. For item j and two attribute patterns $\boldsymbol{\alpha}, \boldsymbol{\alpha}' \in \{0,1\}^K$, the KL information is defined as:

$$D_j(\boldsymbol{\alpha}||\boldsymbol{\alpha}') = \sum_{x=0}^{1} P(X_j = x|\boldsymbol{\alpha}) \, log \left[\frac{P(X_j = x|\boldsymbol{\alpha})}{P(X_j = x|\boldsymbol{\alpha}')}\right], \tag{4}$$

where $P(X_j = x|\boldsymbol{\alpha})$ denotes the probability of response x given attribute pattern $\boldsymbol{\alpha}$ under the specified CDM. Building on this framework, the KL index was proposed by Xu et al. (2003) to select the next item maximizing:

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

$$KL_h(\hat{\boldsymbol{\alpha}}) = \sum_{c=1}^{2^K} D_h(\hat{\boldsymbol{\alpha}}||\boldsymbol{\alpha}_c)\,, \tag{5}$$

where h represents the candidate item in the pool of available items, $\hat{\boldsymbol{\alpha}}$ is the current estimate of the examinee's attribute pattern, and c indexes attribute patterns (c = 1, 2, …, 2^K). This index measures the total divergence between the response distributions under the estimated pattern and all other possible patterns. Cheng (2009) enhanced this approach by incorporating posterior probabilities through the posterior-weighted KL (PWKL) index:

$$PWKL_h(\hat{\boldsymbol{\alpha}}) = \sum_{c=1}^{2^K} D_h(\hat{\boldsymbol{\alpha}}||\boldsymbol{\alpha}_c)\pi\left(\boldsymbol{\alpha}_c|\boldsymbol{x}_J\right)\,, \tag{6}$$

where $\pi(\boldsymbol{\alpha}_c|\boldsymbol{x}_J)$ is the posterior probability after J items have been answered, and $\boldsymbol{x}_J$ is the response vector.

Other KL information-based item selection methods include the modified PWKL (MPWKL) method (Kaplan et al., 2015), and posterior-weighted CDM discrimination index (PWCDI) method (Zheng & Chang, 2016). This study employs only PWKL for comparison, as MPWKL, while achieving comparable performance to GDI, incurs substantial computational costs (Kaplan et al., 2015). Additionally, PWCDI demonstrates inferior performance to PWKL with small calibration samples (Chang et al., 2019), making PWKL the most suitable KL-based comparator for this investigation.

Shannon Entropy Based Approaches

Shannon entropy (Cover & Thomas, 1991; Shannon, 1948) provides an alternative framework for quantifying uncertainty in the posterior distribution of attribute patterns. In the context of CD-CAT, after J items have been selected, the entropy of a distribution π is defined as:

$$H(\pi) = -\sum_{c=1}^{2^K} \pi\left(\boldsymbol{\alpha}_c|\boldsymbol{x}_J\right) log\left[\pi\left(\boldsymbol{\alpha}_c|\boldsymbol{x}_J\right)\right]. \tag{7}$$

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

Lower entropy values indicate greater certainty about the true attribute pattern. Building on information theory principles (Cover & Thomas, 1991), Tatsuoka (2002) proposed selecting items by minimizing the expected posterior entropy:

$$SHE_h = \sum_{x=0}^{1} H\left(\pi | X_h = x, \boldsymbol{x}_J\right) P\left(X_h = x | \boldsymbol{x}_J\right), \tag{8}$$

where $\pi | X_h = x, \boldsymbol{x}_{\boldsymbol{J}}$ denotes the posterior distribution after observing response *x* to the candidate item *h* in the pool of available items, $P\left(X_h = x \middle| \boldsymbol{x}_{\boldsymbol{J}}\right)$ denotes the predicted probability of observing response *x* conditional on the response vector $\boldsymbol{x}_J$, and:

$$P\left(X_h = x | \boldsymbol{x}_J\right) = \sum_{c=1}^{2^K} P(X_h = x | \boldsymbol{\alpha}_c) \pi\left(\boldsymbol{\alpha}_c | \boldsymbol{x}_J\right). \tag{9}$$

Recent methodological advances have extended this framework through the expected mutual information index (Wang, 2013) and the Jensen–Shannon divergence (JSD) index (Minchen & de la Torre, 2016; Yigit et al., 2019). Theoretical investigations have established that JSD is mathematically equivalent to mutual information in quantifying the information gain about an examinee's attribute pattern (Yigit et al., 2019). Furthermore, a linear relationship was found between SHE and JSD (Wang et al., 2020), indicating that these methods yield equivalent item selection decisions in CD-CAT applications. Given these theoretical equivalences, the present study employs the SHE method (Tatsuoka, 2002) as the representative entropy-based approach in our comparative analyses.

2.2.3 GDI Approach

Building on the generalized DINA framework (de la Torre, 2011), Kaplan et al. (2015) introduced the G-DINA discrimination index (GDI). This index offers computational advantages by working with reduced attribute patterns, which are made up by K_h^* attributes required by item *h*. For example, if a $\boldsymbol{q}$-vector is defined as (1, 0, 1, 0, 1), $K_h^* = 3$ attributes since this item only requires the first, third, and fifth attributes. Consequently, there are 8 reduced attribute patterns based on the three required attributes. The GDI for item *h* is defined as

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

$$GDI_h = \sum_{c=1}^{2^{K_h^*}} \pi(\boldsymbol{\alpha}_{ch}^*)[P(X_h = 1|\ \boldsymbol{\alpha}_{ch}^*) -\ \bar{P}_h]^2\ , \tag{10}$$

where $\boldsymbol{\alpha}_{ch}^*$ represents the c-th reduced attribute pattern for item h $(c = 1, 2, \dots, 2^{K_h^*})$, $\pi(\boldsymbol{\alpha}_{ch}^*)$ is the posterior probability of the reduced attribute pattern after J items have been selected, and $\bar{P}_h = \sum_{c=1}^{2^{K_h^*}} \pi(\boldsymbol{\alpha}_{ch}^*) P(X_h = 1|\boldsymbol{\alpha}_{ch}^*)$ is the mean success probability. The GDI measures an item's ability to differentiate between reduced attribute vectors, emphasizing those with higher success probabilities. The item with the highest GDI in the pool is selected.

2.2.4 Comparative Properties

These methods offer distinct advantages for single-purpose CD-CAT. The KL-based methods directly measure discrimination between attribute patterns, with PWKL improving upon the original KL index by incorporating posterior information (Cheng, 2009). The entropy-based methods approach item selection through uncertainty reduction in the posterior distribution. While SHE directly minimizes expected uncertainty, the mutual information method provides a theoretically equivalent formulation through information gain (Wang, 2013). The GDI achieves computational efficiency through dimension reduction while maintaining measurement precision, particularly advantageous for assessments with many attributes (Kaplan et al., 2015). Despite their demonstrated effectiveness, opportunities remain for improving attribute pattern estimation efficiency in CD-CAT. The following section introduces a likelihood-based profile shrinkage (LBPS) algorithm that builds upon these theoretical foundations while addressing certain limitations in existing approaches.

2.3 Computational Considerations in CD-CAT

The practical implementation of CD-CAT item selection methods faces significant computational challenges, primarily arising from the need to evaluate large numbers of attribute patterns during the item selection process. For KL-based methods, item selection entails computing and summing up the KL information between the current attribute pattern estimate and all 2^K possible patterns, and this is done for all eligible items in the bank. When using PWKL, additional computational burden comes from calculating posterior probabilities for each pattern combination.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

Consider a test measuring $K = 5$ attributes with an item bank of 300 items as an example. Even in this relatively simple case, 32 possible attribute patterns must be evaluated for each item selection decision. The PWKL method requires computing and summing KL divergence values across all 32 patterns for each item under consideration. This computation must be performed for all eligible items in the bank to select the next item. The Shannon entropy method involves similar computational intensity, requiring the calculation of expected entropy by evaluating posterior distributions for possible responses across all attribute patterns, again repeated for each item in the bank.

The GDI method introduced by Kaplan et al. (2015) offers some computational advantages by working only with the attributes required for each item. This reduces the pattern space from 2^K to $2^{K_h^*}$, where K_h^* is typically much smaller than K, and requires fewer posterior probability calculations. However, even with these improvements, significant computational challenges remain. These computational demands become particularly acute when tests measure many attributes and as item bank expands.

The practical implications of these computational demands are substantial. They can affect response time between items, overall test administration efficiency, and the system resources required to implement CD-CAT. When multiple examinees are tested simultaneously, as is common in educational settings, these computational requirements become even more demanding. While the GDI method has made progress in reducing computational burden through reduced attribute patterns, there remains a clear need for more efficient approaches that can maintain measurement precision while reducing computation time and scaling effectively with the number of attributes.

3. The Likelihood-Based Profile Shrinkage Algorithm

3.1 Key Ideas

The computational burden of traditional CD-CAT item selection methods grows exponentially with the number of attributes K, as each method evaluates all 2^K possible attribute patterns for each eligible item in the bank at every item selection decision. However, as testing proceeds, the set of plausible, or most likely, attribute patterns for an examinee typically shrinks based on their response pattern. This observation motivates the key insight of LBPS: by focusing on the most likely attribute patterns for item selection while maintaining full pattern space for estimation, substantial computational savings can be achieved without sacrificing measurement precision.

The likelihood function provides a natural mechanism for identifying these plausible patterns. After each response, patterns with maximum likelihood represent the most probable true states given the observed data. Figure 1 illustrates changes in attribute profiles' likelihoods using LBPS with KL when K = 5. Early on, multiple profiles may have similar likelihoods, but as the test proceeds, the number of likely profiles shrinks (see Figure 1). By restricting item selection calculations to these patterns while using all patterns for estimation, LBPS balances computational efficiency with measurement accuracy.

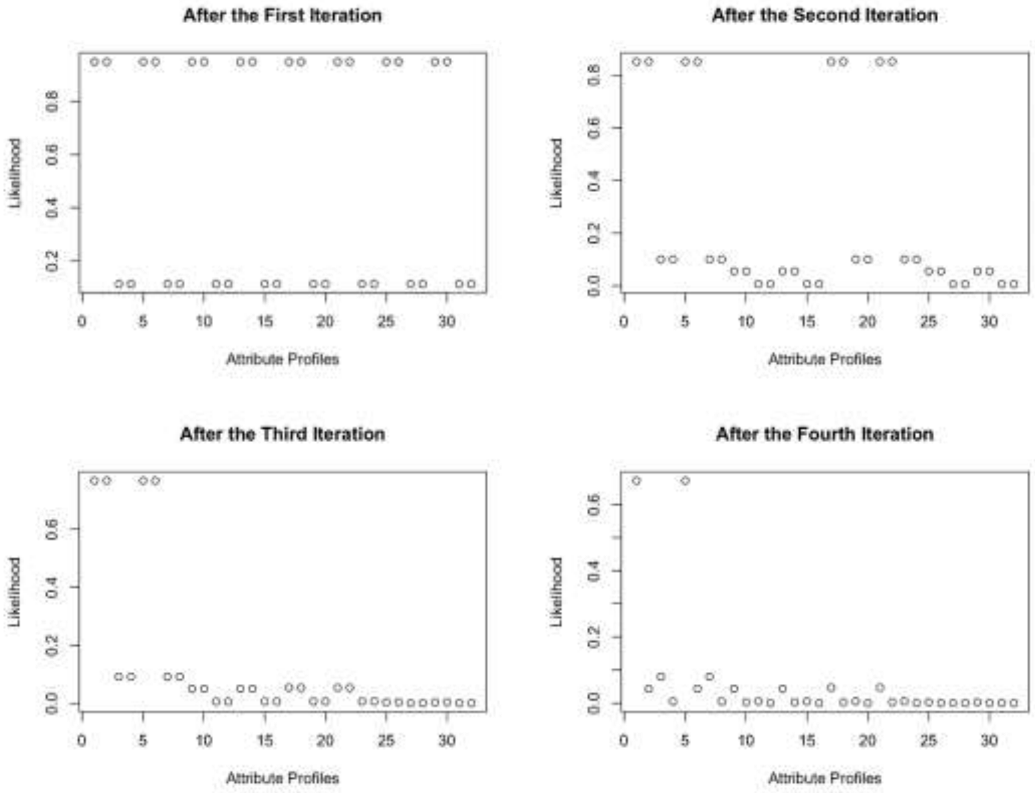


Figure 1. An illustration of changes in attribute profiles' likelihoods using LBPS with KL when K=5

Note: An iteration refers to a single cycle of the adaptive testing process: selecting the next item, collecting the examinee's response, and updating the likelihoods of all attribute profiles and the examinee's estimated attribute profile based on the accumulated responses.

3.2 Theoretical Framework

Let $\boldsymbol{x}_t = (x_1, \ldots, x_t)$ denote the response vector after *t* items have been administered. For any attribute pattern $\boldsymbol{\alpha} \in A = \{0,1\}^K$, the likelihood under a cognitive diagnostic model is:

$$L(\boldsymbol{\alpha}|\boldsymbol{x}_t) = \prod_{j=1}^{t} P\left(X_j = x_j\middle|\boldsymbol{\alpha}\right). \tag{11}$$

Define the set of attribute patterns with the largest likelihood after *t* items have been answered as:

$$M(\boldsymbol{x}_t) = \arg\max_{\boldsymbol{\alpha}\in A} L(\boldsymbol{\alpha}|\boldsymbol{x}_t). \tag{12}$$

For response pattern $\boldsymbol{x}_t$ and a new item *j* that is answered with response x_j, $L\left(\boldsymbol{\alpha}\middle|\boldsymbol{x_t}, x_j\right) = L(\boldsymbol{\alpha}|\boldsymbol{x_t})P\left(X_j = x_j\middle|\boldsymbol{\alpha}\right)$.

Theorem 1 (Pattern Set Size After First Item).

Under the DINA model with $s_j, g_j < 0.5$, for an item requiring *k* attributes:

i. $If\ x_1 = 1{:}\ |M(x_1)| = \ 2^{K-k}$
ii. $If\ x_1 = 0{:}\ |M(x_1)| = 2^K - 2^{K-k}$

where |M| represents the size of the set M, that is, the number of unique attributes patterns in the set M.

Proof:

a) Under the DINA model, for item 1, $L(\boldsymbol{\alpha}|x_1) = P(x_1|\boldsymbol{\alpha}) = (1 - s_1)^{\eta_1(\boldsymbol{\alpha})} {g_1}^{1-\eta_1(\boldsymbol{\alpha})}$, where $\eta_1(\boldsymbol{\alpha}) = \prod_{k=1}^{K} {\alpha_k}^{q_{1k}}$.

b) For $x_1 = 1$, $L(\boldsymbol{\alpha}|x_1 = 1)$ is maximized when $\eta_1(\boldsymbol{\alpha}) = 1$ since $(1 - s_1) > g_1$. This requires all *k* attributes specified by item 1 to be mastered. The remaining $K - k$ attributes can be 0 or 1. Therefore, $|M(x_1)| \ = \ 2^{K-k}$.

c) For $x_1 = 0$, $L(\boldsymbol{\alpha}|x_1 = 0)$ is maximized when $\eta_1(\boldsymbol{\alpha}) = 0$ since $(1 - g_1) > s_1$. This occurs when any required attribute is 0. So in this case $|M(x_1)|= 2^K - 2^{K-k}$.

Theorem 2 (Pattern Set Change)

After *t* items have been answered $(t \geq 1)$, response pattern $\boldsymbol{x}_t$ and any new response x_j from item *j*, the set of attribute patterns with the largest likelihood is

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

$$M(\boldsymbol{x_t}, x_j) = \arg\max_{\boldsymbol{\alpha}\in A} L(\boldsymbol{\alpha}|\boldsymbol{x}_t, x_j) = \arg\max_{\boldsymbol{\alpha}\in A}[L(\boldsymbol{\alpha}|\boldsymbol{x}_t)\, P(X_j = x_j|\boldsymbol{\alpha})].$$

Let $M_{1t} = \{\boldsymbol{\alpha} \in M(\boldsymbol{x}_t): \eta_j(\boldsymbol{\alpha}) = 1\}$ (i.e., patterns within $M(\boldsymbol{x}_t)$ that mastered all attributes required item *j*), and $M_{0t} = \{\boldsymbol{\alpha} \in M(\boldsymbol{x}_t): \eta_j(\boldsymbol{\alpha}) = 0\}$ (patterns within $M(\boldsymbol{x}_t)$ that miss one or more attributes required by item *j*). The updated pattern set $M(\boldsymbol{x_t}, x_j)$ follows one of the three cases:

Case 1 (Shrinkage): If $M_{1t} \neq \emptyset$, and $M_{0t} \neq \emptyset$ (which implies $|M(\boldsymbol{x_t})| \geq 2$), then $|M(\boldsymbol{x_t}, x_j)| < |M(\boldsymbol{x_t})|$. This occurs because item *j* separates patterns within $M(\boldsymbol{x}_t)$, with at least one pattern mastering all required attributes while others miss at least one attribute. If $x_j = 1, M(\boldsymbol{x_t}, x_j) = M_{1t}$; if $x_j = 0, M(\boldsymbol{x_t}, x_j) = M_{0t}$. This is the most common case when $|M(\boldsymbol{x_t})|$ is large, which tends to be the case at the beginning of the test, especially when *K* is large.

Case 2 (Stability): If either (a) $M_{1t} \neq \emptyset, M_{0t} = \emptyset$ and $x_j = 1$, or (b) $M_{1t} = \emptyset, M_{0t} \neq \emptyset$ and $x_j = 0$, then $|M(\boldsymbol{x_t}, x_j)| = |M(\boldsymbol{x_t})|$. This occurs when all patterns within $M(\boldsymbol{x}_t)$ lead to the same η_j and the observed response x_j matches η_j.

Case 3: If (1) $M_{1t} = \emptyset$ and $x_j = 1$, or (2) $M_{0t} = \emptyset$ and $x_j = 0$:

a) **Growth** occurs when external patterns, i.e., patterns outside of $M(\boldsymbol{x}_t)$, have precisely the threshold likelihood: $L(\boldsymbol{\alpha}'|\boldsymbol{x}_t) = L^* \cdot \frac{g_j}{1-s_j}$ for $x_j = 1$, or $L(\boldsymbol{\alpha}'|\boldsymbol{x}_t) = L^* \cdot \frac{s_j}{1-g_j}$ for $x_j = 0$. Here, $L^* = L(\boldsymbol{\alpha}|\boldsymbol{x}_t)$ for any $\boldsymbol{\alpha} \in M(\boldsymbol{x}_t)$. This exact equality is mathematically possible but rare in practice.

b) **Replacement** occurs when there exists at least one external pattern $\boldsymbol{\alpha}'$ leading to $L(\boldsymbol{\alpha}'|\boldsymbol{x}_t)$ that exceeds the threshold likelihood ratio. This can result in $|M(\boldsymbol{x_t}, x_j)|$ being larger, smaller, or equal to $|M(\boldsymbol{x_t})|$ depending on the number of qualifying external patterns.

c) **Stability** occurs when no external patterns meets the threshold.

Proof:

Define $A_{1t} = \{\boldsymbol{\alpha} \in \boldsymbol{A}\backslash M(\boldsymbol{x}_t): \eta_j(\boldsymbol{\alpha}) = 1\}$ (i.e., patterns outside of $M(\boldsymbol{x}_t)$ that should lead to a correct answer to item *j*), and $A_{0t} = \{\boldsymbol{\alpha} \in \boldsymbol{A}\backslash M(\boldsymbol{x}_t): \eta_j(\boldsymbol{\alpha}) = 0\}$ (i.e., patterns outside of $M(\boldsymbol{x}_t)$

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

that should lead to an incorrect answer to item *j*). Let $L^{*} = L(\boldsymbol{\alpha}|\boldsymbol{x}_t)$ for any $\boldsymbol{\alpha} \in M(\boldsymbol{x}_t)$, and $L^{*\prime} = L(\boldsymbol{\alpha}'|\boldsymbol{x}_t)$ for any $\boldsymbol{\alpha}' \notin M(\boldsymbol{x}_t)$. Note that $L^{*} > L^{*\prime}$ by the definition of $M(\boldsymbol{x}_t)$.

(1) When response $x_j = 1$:

Under DINA, if $\eta_j(\boldsymbol{\alpha}) = 1, P\left(X_j = 1\middle|\boldsymbol{\alpha}\right) = 1 - s_j$; if $\eta_j(\boldsymbol{\alpha}) = 0, P\left(X_j = 1\middle|\boldsymbol{\alpha}\right) = g_j$. Since $s_j, g_j < 0.5$, we have $\left(1 - s_j\right) > \ g_j$. After observing $x_j = 1$:

- For $\boldsymbol{\alpha} \in M_{1t}$: $L\left(\boldsymbol{\alpha}\middle|\boldsymbol{x}_t, x_j\right) = L^{*} \cdot \left(1 - s_j\right)$

- For $\boldsymbol{\alpha} \in M_{0t}$: $L\left(\boldsymbol{\alpha}\middle|\boldsymbol{x}_t, x_j\right) = L^{*} \cdot g_j$

- For $\boldsymbol{\alpha}' \in A_{1t}$: $L\left(\boldsymbol{\alpha}'\middle|\boldsymbol{x}_t, x_j\right) = L^{*\prime} \cdot \left(1 - s_j\right)$

- For $\boldsymbol{\alpha}' \in A_{0t}$: $L\left(\boldsymbol{\alpha}'\middle|\boldsymbol{x}_t, x_j\right) = L^{*\prime} \cdot g_j$

Case 1: If $M_{1t} \neq \emptyset$ and $M_{0t} \neq \emptyset$, then the maximum likelihood after observing x_j is $L^{*} \cdot \left(1 - s_j\right)$, since $\left(1 - s_j\right) > g_j$ and $L^{*} > L^{*\prime}$. This only happens for attribute patterns in M_{1t}. Therefore, $M\left(\boldsymbol{x_t}, x_j\right)$ shrinks to M_{1t} when $M_{0t} \neq \emptyset$.

Case 2: If $M_{1t} \neq \emptyset$ and $M_{0t} = \emptyset$ (i.e., $M(\boldsymbol{x_t}) = M_{1t}$), all patterns in $M(\boldsymbol{x_t})$ achieve $L^{*} \cdot \left(1 - s_j\right)$. Therefore, $M\left(\boldsymbol{x_t}, x_j\right) = M(\boldsymbol{x_t})$ (Stability).

Case 3: If $M_{1t} = \emptyset$ and $M_{0t} \neq \emptyset$ (i.e., $M(\boldsymbol{x_t}) = M_{0t}$), all patterns in $M(\boldsymbol{x_t})$ achieve $L^{*} \cdot g_j$.

- For $\boldsymbol{\alpha}' \in A_0$, $L\left(\boldsymbol{\alpha}'\middle|\boldsymbol{x}_t, x_j\right) = L^{*\prime} \cdot g_j < L^{*} \cdot g_j$, so those patterns do not enter $M\left(\boldsymbol{x_t}, x_j\right)$. Therefore, we can focus on $\boldsymbol{\alpha}' \in A_{1t}$.

- For $\boldsymbol{\alpha}' \in A_{1t}$:

- if $L^{*\prime} \cdot \left(1 - s_j\right) = L^{*} \cdot g_j$, that is, $\frac{L^{*\prime}}{L^{*}} = \frac{g_j}{1-s_j}$, then pattern $\boldsymbol{\alpha}'$ ties with patterns in M_{0t} and becomes a part of $M\left(\boldsymbol{x_t}, x_j\right)$ (growth);
- if $L^{*\prime} \cdot \left(1 - s_j\right) > L^{*} \cdot g_j$, that is, $\frac{L^{*\prime}}{L^{*}} > \frac{g_j}{1-s_j}$, then pattern $\boldsymbol{\alpha}'$ achieves higher likelihood than all patterns in M_{0t}, so $M\left(\boldsymbol{x_t}, x_j\right) = \{\boldsymbol{\alpha}' \in A_{1t}: L(\boldsymbol{\alpha}'|\boldsymbol{x_t}) > L^{*} \cdot \frac{g_j}{1-s_j}\}$ (replacement), and the size of $M\left(\boldsymbol{x_t}, x_j\right)$ may increase, decrease, or remain unchanged relative to the size of $M(\boldsymbol{x_t})$;

- if $L^{*\prime} \cdot (1 - s_j) < L^* \cdot g_j$, that is, $\frac{L^{*\prime}}{L^*} < \frac{g_j}{1-s_j}$, then pattern $\boldsymbol{\alpha}'$ doesn't become a member of $M(\boldsymbol{x_t}, x_j)$, so $M(\boldsymbol{x_t}, x_j) = M_{0t} = M(\boldsymbol{x_t})$ (stability).

Therefore, when $x_j = 1$, for $M(\boldsymbol{x_t}, x_j)$ to expand beyond $M(\boldsymbol{x_t})$, $\frac{L^{*\prime}}{L^*}$ must be at least as large as $\frac{g_j}{1-s_j}$. If equality holds for some patterns $\boldsymbol{\alpha}' \in A_{1t}$, growth occurs; if inequality holds for some patterns $\boldsymbol{\alpha}' \in A_{1t}$, replacement occurs; if no external patterns meets threshold, stability occurs.

(2) When response $x_j = 0$, the same logic applies. Whether the size of $M(\boldsymbol{x_t}, x_j)$ increases or shrinks compared to $M(\boldsymbol{x_t})$ depends on $\frac{L^{*\prime}}{L^*}$ and $\frac{s_j}{1-g_j}$. If they are equal for some patterns $\boldsymbol{\alpha}' \in A_{0t}$, the size increases; if $\frac{L^{*\prime}}{L^*} > \frac{s_j}{1-g_j}$ holds for some patterns $\boldsymbol{\alpha}' \in A_{0t}$, the size may increase, remain stable, or decrease.

A Special Case: When $|\boldsymbol{M}(\boldsymbol{x_t})| = 1$

A critical scenario arises when the pattern set contains only a single pattern: $M(\boldsymbol{x_t}) = \{\boldsymbol{\alpha}^*\}$. This situation may arise at the late stage of a test. With only one pattern, Case 1 (i.e., mixed mastery) in Theorem 2 is impossible. The single pattern either meets item *j*'s requirements ($M_{1t} = \{\boldsymbol{\alpha}^*\}, M_{0t} = \emptyset$) or doesn't ($M_{1t} = \emptyset, M_{0t} = \{\boldsymbol{\alpha}^*\}$). This is a special case of either Case 2 or Case 3 discussed in Theorem 2. According to Theorem 2, $|M(\boldsymbol{x_t})|$ either stays at 1 or possibly expands. There is no chance of further shrinkage in terms of the size of the set.

Conditions Driving Shrinkage in LBPS. The probability of shrinkage (Case 1) in Theorem 2 depends critically on heterogeneity within $M(\boldsymbol{x_t})$ — that is, whether some patterns meet all requirements of item *j*, while others do not. This heterogeneity is likely to occur when *K* is large and testing is in early stages, due to the combinatorial structure of the pattern space. After *t* items which have tested *m* < *K* attributes, $M(\boldsymbol{x_t})$ must contain all 2^{K-m} variants on untested attributes for each viable tested configuration. This ensures a diverse set of attribute patterns that item *j* can potentially differentiate.

While early stages benefit from guaranteed shrinkage, repeated shrinkage often drives $|M(\boldsymbol{x_t})|$ to a small number of patterns at the later stage of testing. This results in a computational advantage: with LBPS, item selection methods like KL, PWKL, or SHE need to evaluate items

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

against only a handful of patterns remaining in $M(\boldsymbol{x_t})$; while without LBPS, they must evaluate all 2^K patterns at every stage. Thus, LBPS effectively leverages the structure of high-dimensional attribute spaces to mitigate the computational burden of exhaustive pattern evaluation. Its advantage is pronounced when K is large, offering the greatest benefit precisely when traditional methods become computationally prohibitive.

Theorem 3 (Pattern Set Size Reduction).

General reduction: When Case 1 (shrinkage) occurs, the reduction in pattern set size of $M\left(\boldsymbol{x_t}, x_j\right)$ is :

$$\left|M\left(\boldsymbol{x_t}, x_j\right)\right| = \begin{cases} |M_{1t}|, & if\ x_j = 1 \\ |M_{0t}|, & if\ x_j = 0 \end{cases}$$

The reduction ratio is $p_j = \frac{|M(\boldsymbol{x_t}, x_j)|}{|M(\boldsymbol{x_t})|} = \begin{cases} \frac{|M_{1t}|}{|M(\boldsymbol{x_t})|}, & if\ x_j = 1 \\ \frac{|M_{0t}|}{|M(\boldsymbol{x_t})|}, & if\ x_j = 0 \end{cases}$

This ratio depends on the proportion of patterns in $M(\boldsymbol{x_t})$ that would ideally yield each response. As a special case, when item *j* measures only untested attributes:

i. $If\ x_j = 1$: $\left|M\left(\boldsymbol{x_t}, x_j\right)\right| = |M(\boldsymbol{x_t})| \cdot \frac{1}{2^{k_{new}}}$

ii. $If\ x_j = 0$: $\left|M\left(\boldsymbol{x_t}, x_j\right)\right| = |M(\boldsymbol{x_t})| \cdot \left(1 - \frac{1}{2^{k_{new}}}\right)$

Here, k_{new} refers to the number of newly introduced attributes—that is, attributes required by item *j* but not yet assessed by any of the first *t* items. When $k_{new} > 1$, the reduction of the most likely pattern space is sharper for a correct response ($x_j = 1$), and less sharp but still substantial for an incorrect response ($x_j = 0$). When $k_{new} = 1$, the reduction is by half for a correct response or an incorrect response.

Proof: This follows directly from Case 1 of Theorem 2, where we showed that $M\left(\boldsymbol{x_t}, x_j\right) = M_{1t}$ when $x_j = 1$ and $M\left(\boldsymbol{x_t}, x_j\right) = M_{0t}$ when $x_j = 0$.

For the special case:

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

a) When $x_j = 1$, $L\left(\boldsymbol{\alpha}\middle|\boldsymbol{x_t}, x_j\right)$ is maximized when $\eta_j(\boldsymbol{\alpha}) = 1$. This means that each new required attribute must be mastered. Only $\frac{1}{2^{k_{new}}}$ of all patterns in $M(\boldsymbol{x_t})$ can have 1's on all the new attributes. Therefore, $\left|M\left(\boldsymbol{x_t}, x_j\right)\right| = \frac{|M(\boldsymbol{x_t})|}{2^{k_{new}}}$.

b) When $x_j = 0$, $L\left(\boldsymbol{\alpha}\middle|\boldsymbol{x_t}, x_j\right)$ is maximized when $\eta_j(\boldsymbol{\alpha}) = 0$. This means that at least one of the new required attributes is not mastered, and the proportion of such patterns in $M(\boldsymbol{x_t})$ is $1 - \frac{1}{2^{k_{new}}}$. Therefore, $\left|M\left(\boldsymbol{x_t}, x_j\right)\right| = |M(\boldsymbol{x_t})| \cdot \left(1 - \frac{1}{2^{k_{new}}}\right)$.

When $k_{new} = 1$, following a) and b), the size of the set of attribute patterns that maximize the likelihood is reduced by half.

The aforementioned theorems suggest that the efficiency of LBPS is influenced by the **q**-vectors of the selected items. The sequential selection of items that assess new attributes leads to a reduced pattern space for item selection in LBPS. This insight coincides with Xu et al.'s (2016) optimal initial item selection theory for CD-CAT. Specifically, Xu et al. (2016) demonstrated that to achieve minimum test length, the first administered item must assess exactly one attribute, followed by items that sequentially introduce single, previously unmeasured attributes. If this condition is not met, identifying all attribute profiles within *K* items becomes infeasible, resulting in test lengths exceeding *K*. If the condition is met, following Theorem 2 and 3, LBPS should help shrink the search space by half at each step during the early stage of the test when *K* is large, thereby achieving substantial computational gains.

Note that the above theorems are built on the DINA model, but they can be extended to other CDMs with ideal response functions. For example, for the DINO model with $\omega_j(\boldsymbol{\alpha}) = 1 - \prod_{k=1}^{K}(1 - {\alpha_k}^{q_{jk}})$, the theorems hold with ω_j replacing η_j. For CDMs without ideal response functions (e.g., general CDMs such as LCDM and G-DINA), LBPS can still be implemented as it operates directly on likelihood values and relies on likelihood updating. Shrinkage of $M(\boldsymbol{x}_t)$ may occur when patterns within $M(\boldsymbol{x}_t)$ yield different response probabilities, because likelihood updating favors patterns whose predicted probabilities better align with the observed outcome. Such behavior is more likely when $|M(x_t)|$ is large, as characterized under the DINA model. That said, the specific theoretical properties established in this paper (e.g., the characterization of shrinkage via M_{1t} and M_{0t}) do not directly apply, and further investigation of LBPS performance under these models is needed.

While the above theorems demonstrate how the set of potential likely attribute patterns changes with each item response, practical implementation requires a concrete algorithm. The following section outlines the step-by-step LBPS procedure.

3.3 Algorithm Description

The LBPS algorithm maintains two key sets: (1) $M(\boldsymbol{x}_t)$: most likely patterns or patterns that lead to the maximum likelihood, and (2) $M^*(\boldsymbol{x}_t)$: working pattern set used for item selection. Note that the estimation of each examinee's attribute profile uses the full pattern space (all possible 2^K attribute profiles); the working pattern set is only used for item selection. In terms of the working set size, at any stage *t*, $2 \leq |M^*(\boldsymbol{x}_t)| \leq 2^K$. This ensures minimally two distinct patterns for item selection decisions. The algorithm proceeds as follows:

Step 1: First Item Selection

a) Use full pattern space **A**

b) Select an item using traditional method (or randomization)

c) Obtain response x_1

d) Calculate $L(\boldsymbol{\alpha}|x_1)$ for all $\boldsymbol{\alpha} \in \mathbf{A}$

e) Define initial $M(x_1)$ and $M^*(x_1)$: $M(x_1)$ is the set of attribute patterns with the largest likelihood after the first item has been answered. $M^*(x_1)$ is the working pattern set used for item selection defined as follows:

$$M^*(x_1) = \begin{cases} \mathrm{M}(x_1), & \text{if } |\mathrm{M}(x_1)| \geq 2 \\ \{\boldsymbol{\alpha}^1_{(1)}, \boldsymbol{\alpha}^1_{(2)}\}, & \text{if } |\mathrm{M}(x_1)| = 1 \end{cases}$$

where $\boldsymbol{\alpha}^1_{(1)}$ has maximum likelihood and $\boldsymbol{\alpha}^1_{(2)}$ has the second-highest likelihood at stage 1 (*t* = 1).

f) Estimate the examinee's attribute profile (using full pattern space **A**) based on response x_1 using the maximum likelihood estimation (MLE).

Step 2: Subsequent Items (t > 1)

For each eligible item in the pool:

a) Pattern Space Update

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

- Calculate $L(\boldsymbol{\alpha}|\boldsymbol{x}_t)$ for all $\boldsymbol{\alpha} \in \mathbf{A}$
- Identify $M(\boldsymbol{x}_t) = \{\boldsymbol{\alpha}: L(\boldsymbol{x}_t) = L^*\}$, L^* is the current maximum likelihood value among all profiles' likelihoods
- Define working set:

$$M^*(\boldsymbol{x}_t) = \begin{cases} \mathrm{M}(\boldsymbol{x}_t), & \text{if } |\mathrm{M}(\boldsymbol{x}_t)| \geq 2 \\ \{\boldsymbol{\alpha}_{(1)}^t, \boldsymbol{\alpha}_{(2)}^t\}, & \text{if } |\mathrm{M}(\boldsymbol{x}_t)| = 1 \end{cases}$$

where $\boldsymbol{\alpha}_{(1)}^t$ has maximum likelihood and $\boldsymbol{\alpha}_{(2)}^t$ has the second-highest likelihood at stage t.

b) Item Selection

- Couple an existing item selection method with $M^*(\boldsymbol{x}_t)$ instead of **A** to select the next item. For example, when KL is used for item selection, the summation in (5) is not over all 2^K patterns in A, but only over the patterns in $M^*(\boldsymbol{x}_t)$.

c) Response Processing

- Obtain response x_{t+1} and iterate steps a) and b)

Repeat the whole process of step 2 until the desired number of items have been administered or a prefixed termination criterion has been reached.

Logically, the key efficiency gain of LBPS comes from restricting the item selection computations to within the working set, which shrinks quickly over time, while maintaining estimation accuracy through full pattern space calculations. Practically, the extent to which LBPS helps improve computational efficiency and maintains classification accuracy needs to be evaluated in light of many factors, such as the number of attributes *K*, the test length, and the underlying CDM model. Therefore, a simulation study was conducted manipulating these factors to evaluate the practical impact of the LBPS.

4. Simulation Design

A simulation study was conducted to evaluate the measurement efficiency of the proposed LBPS algorithm in selecting items for CD-CAT. Specifically, LBPS was coupled with

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

the well-known KL, PWKL, SHE, and GDI methods and compared to the performance of these methods in their original forms.

(1) CDM: DINA was used in the study to model item parameters and simulate examinees' responses to items. To assess the generalizability of LBPS, we also conducted a simulation using the DINO model under conditions described below. Due to space limitations, results from the DINO-based simulation are presented in the Appendix.

(2) Item bank:

   (a) Number of assessed attributes (*K*): *K* = 3, 5, and 7.

   (b) Item bank size (*J*) and item quality: For each *K*, two sizes of item banks were generated: 300 and 500. For each of the two bank sizes, two levels of item quality banks were generated: one item bank consisted of high-quality items, with guessing and slipping parameters randomly drawn from U(0.05, 0.25); the other item bank contained lower-quality items, with guessing and slipping parameters randomly drawn from U(0.25, 0.50). In total, 12 item banks were generated for the simulation.

   (c) Q-matrix: Corresponding to the combinations of *J* and *K*, 6 different Q-matrices were generated. The Q-matrix used in this study was generated item by item and attribute by attribute. Each item has a 30% chance of measuring each attribute. This mechanism was employed to ensure that every attribute is adequately and equally represented in the item pool. Details of the Q-matrices are summarized in Tables A1 and A2 in the Appendix.

(3) Test length: *T* = 5, 10, 15, 20, 25, and 30 items

(4) Examinees: the attribute profiles of 1,000 examinees were randomly generated from the set of all possible attribute profiles for each condition. Examinees' responses to each item was generated from the DINA model.

(5) Item selection methods: a) traditional approaches: KL, PWKL, and SHE; b) GDI, selected for its known computational advantage; and c) LBPS-enhanced variants: LBPS-KL, LBPS-PWKL, LBPS-SHE, and LBPS-GDI. The uniform prior was used for each

method when applicable to select items, that is, each attribute profile was assumed to have equal prior probability, $\frac{1}{2^K}$, before the start of the test.

(6) Estimation: The initial attribute profile estimate, $\widehat{\boldsymbol{\alpha}}(0)$, was randomly drawn from all possible attribute profiles (2^K profiles). Then the maximum likelihood estimation (MLE) method was used to update $\widehat{\boldsymbol{\alpha}}(t)$. The final estimates were $\widehat{\boldsymbol{\alpha}}(T)$, where T is the test length.

(7) Evaluation criteria:

(a) profile estimation accuracy: average attribute-wise agreement rate (AAR), pattern-wise agreement rate (PAR)

$$AAR = \sum_{i=1}^{N} \sum_{k=1}^{K} \frac{I[\widehat{\alpha}_{ik} = \alpha_{ik}]}{NK}, \tag{13}$$

$$PAR = \sum_{i=1}^{N} \frac{I[\widehat{\boldsymbol{\alpha}}_i = \boldsymbol{\alpha}_i]}{N}, \tag{14}$$

where $I[\cdot]$ is an indicator function, N is the number of examinees, $\widehat{\boldsymbol{\alpha}}_i$ and $\boldsymbol{\alpha}_i$ denotes the estimated and true attribute profile estimate for examinee i, and $\hat{\alpha}_{ik}$ and α_{ik} denotes the estimated and true attribute k for examinee i.

(b) computation efficiency: average computation time (seconds) per examinee on the test

(c) test security: mean of test overlap rates ($tor_{ii'}$) between all possible pairs of examinees (Chen et al., 2003; Choe et al., 2018):

$$\overline{tor} = \left(\frac{2}{n}\right)^{-1} \sum_{i=1}^{n-1} \sum_{i'=i+1}^{n} tor_{ii'} \; = \frac{n}{T(n-1)} \sum_{j=1}^{m} er_j^2 - \frac{1}{n-1}. \tag{15}$$

Here, m denotes the size of the item pool, and T is the fixed test length. The index $tor_{ii'}$ represents the proportion of common items between a pair of examinees, calculated as the number of shared items divided by T. The observed exposure rate for item j, denoted er_j, is computed as the number of times item j was administered divided by the total number of examinees n.

The simulation study was conducted using R, and run on a computer system with 48 Cores for computing.

4. Results

Below are the results for the 300-item banks using DINA, including profile estimation accuracy (AAR and PAR), computational efficiency, and mean test-overlap rates. Due to page limitations and their similarity to the 300-item bank results, the corresponding results for the 500-item banks using DINA are presented in the Appendix. As noted in the Methods section, DINO-based results are also included in the Appendix.

4.1 Attribute-wise Agreement Rate (AAR)

Table 1 and Figure 2 present the AARs of various methods across item banks with different characteristics, including the number of assessed attributes and item bank quality. In Table 1, D denotes the difference in AAR (LBPS − Original), where positive values indicate that LBPS increased AARs, and negative values indicate decreases. Overall, the AARs of methods incorporating the proposed LBPS algorithm are largely comparable to those of their original counterparts, indicating that LBPS does not compromise estimation accuracy at the attribute level.

For PWKL, SHE, and GDI, the differences between the LBPS-integrated and original versions are minimal, with AAR differences typically below 0.02. In contrast, LBPS-KL demonstrates modest improvements, particularly with high-quality item banks and larger attribute spaces. Under these conditions, LBPS-KL consistently yields higher AARs at shorter test lengths, with improvements reaching up to 0.14 (e.g., when $K = 7$ and $T = 10$ with a high-quality item bank, $D = 0.14$). For low-quality item banks, the differences between LBPS-KL and KL remain small, ranging from 0.01 to 0.07. These findings suggest that integrating LBPS into traditional item selection methods maintains attribute-level estimation accuracy across a wide range of testing scenarios.

4.2 Pattern-wise Agreement Rate (PAR)

Table 2 and Figure 3 display the PARs for the same set of methods and testing conditions. In Table 2, D denotes the difference in PAR (LBPS − Original), where positive values indicate that LBPS increased PARs and negative values indicate decreases. The results show that LBPS-based methods maintain classification accuracy at the pattern level comparable to that of their original versions. For PWKL, SHE, and GDI, the PAR differences are

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

consistently small—typically below 0.06—regardless of item bank quality, number of attributes, or test length. In contrast, the KL method benefits more substantially from the inclusion of LBPS. The improvements in PAR become more pronounced as the number of attributes increases. Consistent with the AAR findings, the largest gains in PAR are observed when LBPS is combined with KL under high-quality item banks. For example, when T = 15, K = 7, and the item bank is of high quality, LBPS-KL achieves a PAR approximately 0.6 higher than that of KL (D = 0.59). However, under low-quality item banks, the performance gains of LBPS-KL are more modest.

Overall, the AAR and PAR results suggest that incorporating LBPS into traditional item selection methods generally maintains comparable classification accuracy, with a couple of exceptions when K is large or test length is short. For PWKL, SHE, and GDI, the LBPS-integrated versions closely match the performance of their original forms. For KL, LBPS offers modest but consistent improvements, particularly in scenarios involving high item quality, a larger number of attributes, and shorter test lengths.

Table 1. Attribute-Wise Agreement Rates (AAR) under DINA (J = 300)

			KL			PWKL			SHE			GDI		
B	K	T	Original	LBPS	D	Original	LBPS	D	Original	LBPS	D	Original	LBPS	D
H	3	5	0.89	0.94	0.05	0.94	0.94	0.00	0.94	0.94	0.00	0.95	0.95	0.00
		10	0.94	0.99	0.05	0.99	0.99	0.00	0.99	0.98	0.00	0.99	1.00	0.00
		15	0.99	1.00	0.01	1.00	1.00	0.00	0.99	1.00	0.01	1.00	1.00	0.00
		20	0.99	1.00	0.01	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
		25	1.00	1.00	0.01	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
		30	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
	5	5	0.74	0.87	0.13	0.88	0.88	0.00	0.92	0.91	-0.01	0.91	0.91	0.00
		10	0.84	0.97	0.12	0.96	0.97	0.00	0.96	0.95	-0.01	0.97	0.97	0.00
		15	0.90	0.99	0.09	0.99	0.99	0.00	0.99	0.98	-0.01	0.99	0.99	0.00
		20	0.94	1.00	0.05	1.00	1.00	0.00	0.99	0.99	0.00	1.00	1.00	0.00
		25	0.97	1.00	0.03	1.00	1.00	0.00	1.00	0.99	0.00	1.00	1.00	0.00
		30	0.98	1.00	0.02	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00

	7	5	0.69	0.77	0.08	0.78	0.77	-0.01	0.78	0.78	-0.01	0.78	0.78	-0.01
		10	0.78	0.92	0.14	0.92	0.91	-0.01	0.93	0.91	-0.02	0.93	0.92	-0.01
		15	0.83	0.96	0.13	0.97	0.96	-0.02	0.97	0.95	-0.02	0.98	0.97	-0.01
		20	0.87	0.99	0.12	0.99	0.98	0.00	0.98	0.97	-0.01	0.99	0.99	-0.01
		25	0.89	0.99	0.11	1.00	0.99	0.00	0.99	0.98	-0.01	0.99	0.99	0.00
		30	0.91	1.00	0.09	1.00	1.00	0.00	0.99	0.99	0.00	1.00	1.00	0.00
L	3	5	0.73	0.75	0.01	0.75	0.74	-0.01	0.75	0.75	-0.01	0.75	0.75	0.00
		10	0.80	0.84	0.04	0.82	0.83	0.00	0.82	0.82	0.00	0.83	0.84	0.01
		15	0.85	0.89	0.04	0.89	0.87	-0.01	0.86	0.85	-0.01	0.89	0.88	-0.01
		20	0.89	0.91	0.02	0.91	0.93	0.02	0.89	0.87	-0.02	0.91	0.92	0.00
		25	0.90	0.95	0.05	0.94	0.94	0.00	0.90	0.91	0.01	0.94	0.94	0.00
		30	0.92	0.95	0.03	0.95	0.95	0.00	0.91	0.91	-0.01	0.96	0.95	0.00
	5	5	0.64	0.70	0.06	0.68	0.70	0.02	0.69	0.71	0.02	0.70	0.71	0.02
		10	0.70	0.75	0.06	0.76	0.75	-0.01	0.75	0.74	-0.01	0.77	0.76	0.00
		15	0.73	0.80	0.07	0.80	0.80	0.00	0.80	0.79	-0.01	0.81	0.80	0.00
		20	0.78	0.83	0.05	0.83	0.83	0.00	0.82	0.81	-0.01	0.85	0.84	-0.01
		25	0.82	0.88	0.06	0.87	0.86	-0.01	0.85	0.83	-0.02	0.88	0.86	-0.02
		30	0.84	0.88	0.04	0.89	0.89	0.00	0.86	0.85	-0.02	0.90	0.89	-0.01
	7	5	0.63	0.67	0.04	0.66	0.66	0.00	0.65	0.67	0.02	0.65	0.67	0.02
		10	0.69	0.72	0.03	0.72	0.72	0.00	0.74	0.73	0.00	0.72	0.73	0.01
		15	0.73	0.77	0.04	0.77	0.76	-0.01	0.76	0.75	-0.01	0.76	0.77	0.00
		20	0.76	0.79	0.04	0.79	0.79	0.00	0.78	0.78	0.00	0.80	0.79	0.00
		25	0.78	0.80	0.02	0.82	0.81	-0.01	0.80	0.78	-0.02	0.81	0.81	0.00
		30	0.80	0.83	0.03	0.84	0.82	-0.01	0.81	0.80	-0.01	0.84	0.83	-0.02

Note. B = bank quality; K = number of attributes measured by the test; T = test length; KL = Kullback–Leibler Index method; PWKL = posterior weighted Kullback–Leibler information method; SHE = Shannon entropy method; GDI = the generalized deterministic inputs, noisy ‘‘and’’ gate (G-DINA) discrimination index; original = original methods (KL, PWKL, SHE, GDI); LBPS = methods adding LBPS (LBPS-KL, LBPS-PWKL, LBPS-SHE, LBPS-

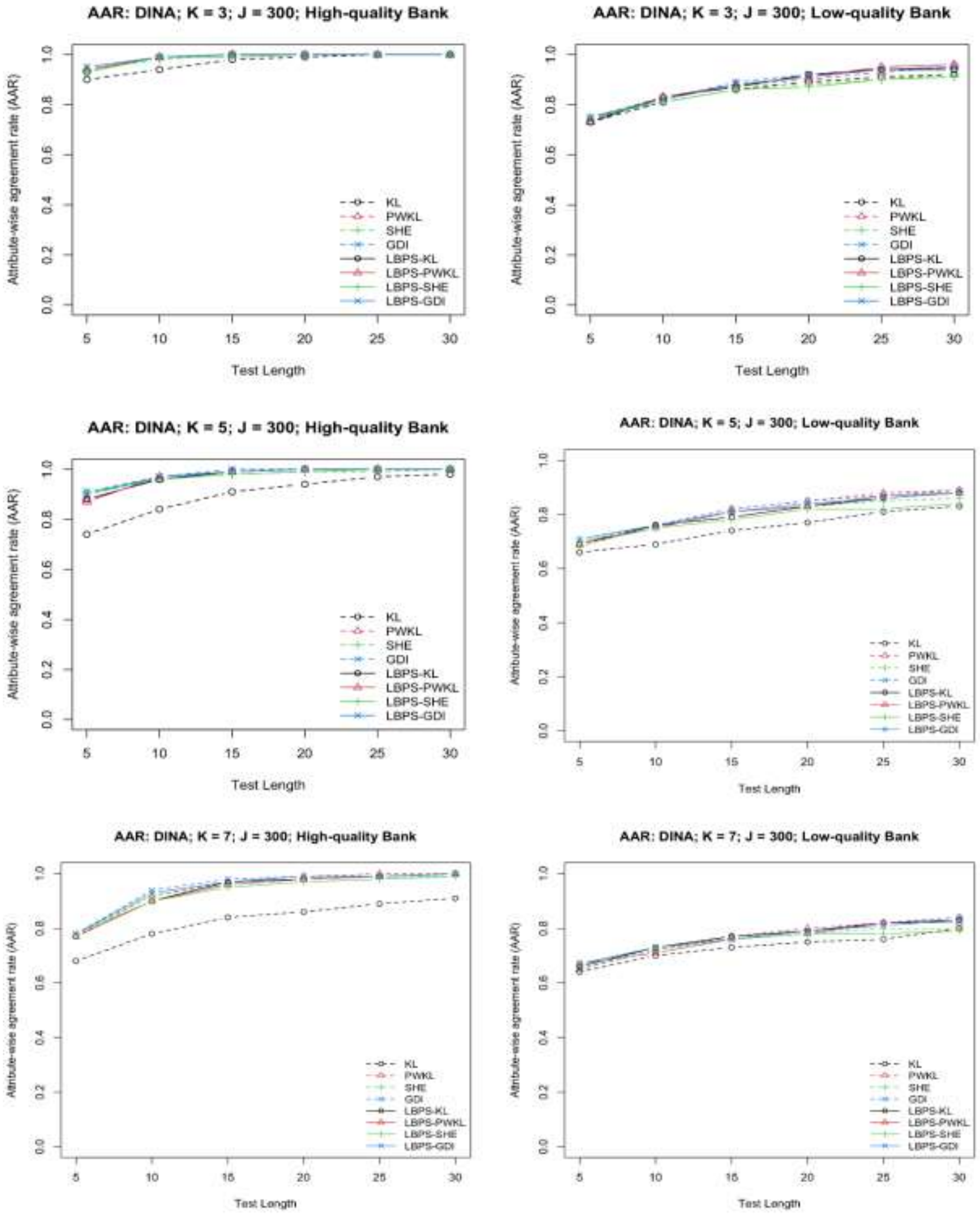
GDI); D = the AAR difference between LBPS-incorporated methods and the original methods (LBPS – Original); positive D values indicate that LBPS increased AARs; negative D values indicate decreases.

Table 2. Pattern-Wise Agreement Rates (PAR) under DINA (J = 300)

			KL			PWKL			SHE			GDI		
B	K	T	Original	LBPS	D	Original	LBPS	D	Original	LBPS	D	Original	LBPS	D
H	3	5	0.71	0.85	0.14	0.85	0.84	-0.01	0.85	0.82	-0.03	0.86	0.87	0.01
		10	0.83	0.98	0.15	0.99	0.98	-0.01	0.96	0.96	-0.01	0.99	0.99	0.00
		15	0.96	1.00	0.04	1.00	1.00	0.00	0.98	0.99	0.01	1.00	1.00	0.00
		20	0.98	1.00	0.02	1.00	1.00	0.00	0.99	0.99	0.00	1.00	1.00	0.00
		25	0.99	1.00	0.01	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
		30	0.99	1.00	0.01	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00
	5	5	0.18	0.55	0.37	0.54	0.57	0.04	0.69	0.68	-0.01	0.70	0.69	-0.01
		10	0.40	0.86	0.46	0.87	0.85	-0.02	0.83	0.82	-0.02	0.89	0.88	-0.01
		15	0.60	0.97	0.37	0.96	0.97	0.02	0.94	0.91	-0.03	0.97	0.97	0.01
		20	0.74	0.99	0.25	0.99	0.99	0.01	0.96	0.96	0.01	1.00	0.99	0.00
		25	0.84	1.00	0.16	1.00	1.00	0.00	0.98	0.97	-0.01	1.00	1.00	0.00
		30	0.89	1.00	0.11	1.00	1.00	0.00	0.98	0.98	0.00	1.00	1.00	0.00
	7	5	0.06	0.18	0.11	0.15	0.17	0.02	0.14	0.17	0.02	0.16	0.15	-0.01
		10	0.14	0.63	0.49	0.61	0.62	0.00	0.62	0.61	-0.01	0.66	0.66	0.00
		15	0.23	0.82	0.59	0.85	0.79	-0.06	0.81	0.75	-0.06	0.88	0.85	-0.03
		20	0.31	0.92	0.62	0.94	0.92	-0.02	0.89	0.85	-0.04	0.95	0.92	-0.03
		25	0.39	0.97	0.58	0.97	0.96	-0.01	0.94	0.90	-0.03	0.97	0.97	0.00
		30	0.47	0.99	0.52	0.99	0.98	-0.01	0.96	0.95	-0.01	0.99	0.99	0.00
L	3	5	0.41	0.45	0.04	0.44	0.44	0.00	0.45	0.46	0.02	0.44	0.47	0.03
		10	0.54	0.63	0.09	0.58	0.59	0.01	0.58	0.59	0.01	0.60	0.61	0.01
		15	0.64	0.74	0.10	0.72	0.69	-0.03	0.67	0.64	-0.02	0.71	0.70	0.00
		20	0.71	0.76	0.06	0.78	0.82	0.04	0.71	0.68	-0.03	0.78	0.78	0.00

	25	0.75	0.86	0.11	0.84	0.83	-0.01	0.76	0.76	0.00	0.84	0.84	0.01
	30	0.79	0.88	0.09	0.87	0.88	0.00	0.77	0.76	-0.01	0.88	0.88	0.00
5	5	0.09	0.19	0.09	0.16	0.18	0.02	0.15	0.20	0.05	0.18	0.23	0.05
	10	0.16	0.30	0.15	0.31	0.30	-0.01	0.28	0.29	0.01	0.30	0.31	0.01
	15	0.20	0.39	0.20	0.38	0.39	0.01	0.36	0.36	0.00	0.38	0.38	0.00
	20	0.29	0.44	0.15	0.45	0.47	0.01	0.41	0.42	0.01	0.50	0.49	0.00
	25	0.37	0.57	0.20	0.52	0.55	0.03	0.48	0.46	-0.02	0.60	0.54	-0.05
	30	0.42	0.61	0.19	0.61	0.61	0.00	0.51	0.48	-0.03	0.62	0.59	-0.03
7	5	0.04	0.06	0.02	0.05	0.07	0.02	0.05	0.05	0.00	0.05	0.05	0.00
	10	0.08	0.13	0.06	0.11	0.13	0.01	0.13	0.14	0.01	0.12	0.14	0.02
	15	0.12	0.20	0.08	0.19	0.19	-0.01	0.17	0.15	-0.02	0.17	0.19	0.02
	20	0.15	0.27	0.12	0.24	0.27	0.03	0.21	0.22	0.01	0.27	0.28	0.01
	25	0.18	0.29	0.11	0.33	0.30	-0.03	0.26	0.24	-0.03	0.30	0.32	0.02
	30	0.24	0.39	0.15	0.36	0.34	-0.02	0.28	0.28	0.00	0.38	0.35	-0.02

Note. B = bank quality; K = number of attributes measured by the test; T = test length; KL = Kullback–Leibler Index method; PWKL = posterior weighted Kullback–Leibler information method; SHE = Shannon entropy method; GDI = the generalized deterministic inputs, noisy ‘‘and’’ gate (G-DINA) discrimination index; original = original methods (KL, PWKL, SHE, GDI); LBPS = methods adding LBPS (LBPS-KL, LBPS-PWKL, LBPS-SHE, LBPS-GDI); D = the PAR difference between LBPS-incorporated methods and the original methods (LBPS – Original); positive D values indicate that LBPS increased PARs; negative D values indicate decreases.



AAR: DINA; K = 3; J = 300; High-quality Bank

AAR: DINA; K = 3; J = 300; Low-quality Bank

AAR: DINA; K = 5; J = 300; High-quality Bank

AAR: DINA; K = 5; J = 300; Low-quality Bank

AAR: DINA; K = 7; J = 300; High-quality Bank

AAR: DINA; K = 7; J = 300; Low-quality Bank

Figure 2. AARs under DINA when J = 300 items

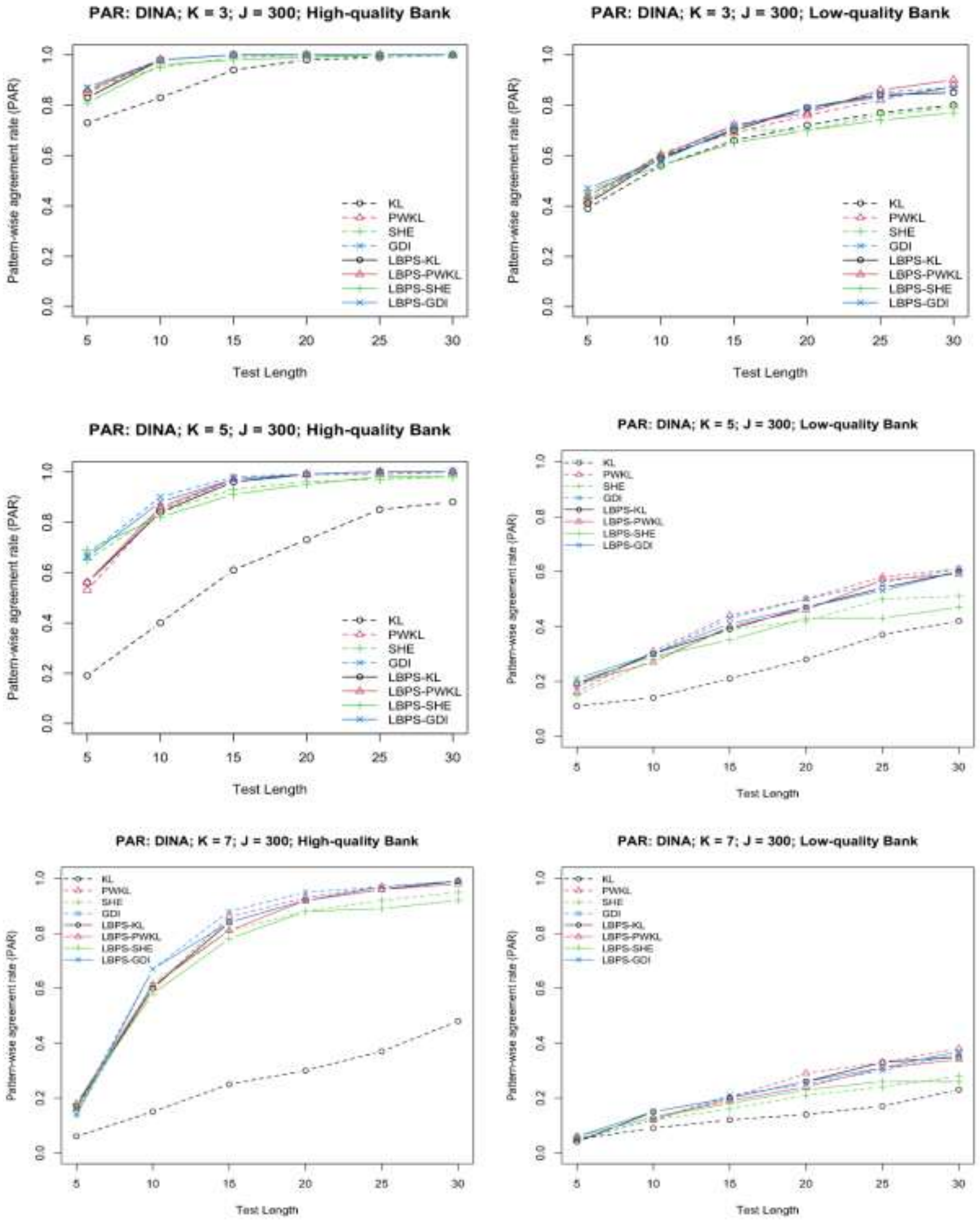


Figure 3. PARs under DINA when J = 300 items

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

4.3 Computation Efficiency

Table 3 and Figure 4 present the average computation time (in seconds) per person for the compared methods under varying conditions. The results show that while integrating LBPS into GDI yields modest to moderate efficiency gains, its integration into KL, PWKL, and SHE consistently leads to substantially lower computation times.

For KL, PWKL, and SHE, the efficiency improvements from LBPS integration are consistent, ranging from 45% in the simplest scenario (K = 3 and T = 5) to nearly 90% in the most demanding case (K = 7, T = 30). These gains become more pronounced as K and T increase. Moreover, LBPS-integrated versions achieves computation times comparable to or lower than GDI, a method known for its relatively higher efficiency than PWKL (Kaplan et al., 2015).

LBPS was also incorporated into GDI to further improve computational efficiency. While gains are modest for small K, they become more substantial as K and T increase. For instance, integrating LBPS into GDI reduced computation time by 37% for T = 30 and K = 7. This demonstrates that LBPS is a highly flexible algorithm that can be integrated with item selection methods beyond information-theoretic approaches like SHE, KL or PWKL.

In sum, these findings show that LBPS not only maintains measurement accuracy but also provides substantial computational efficiency gains, especially for assessments involving many attributes or requiring rapid item selection in longer tests.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

Table 3. Average Computation Time Per Person under DINA (J = 300)

			KL			PWKL			SHE			GDI		
B	K	T	Original	LBPS	PR	Original	LBPS	PR	Original	LBPS	PR	Original	LBPS	PR
H	3	5	0.04	0.02	47%	0.04	0.02	45%	0.04	0.02	54%	0.03	0.03	1%
		10	0.08	0.04	56%	0.09	0.04	56%	0.08	0.03	63%	0.06	0.05	3%
		15	0.12	0.05	59%	0.13	0.05	59%	0.11	0.04	67%	0.09	0.08	5%
		20	0.16	0.06	61%	0.18	0.07	61%	0.15	0.05	68%	0.11	0.11	4%
		25	0.20	0.08	62%	0.22	0.08	62%	0.19	0.06	69%	0.14	0.14	5%
		30	0.24	0.09	62%	0.27	0.10	63%	0.23	0.07	69%	0.17	0.17	2%
	5	5	0.18	0.09	53%	0.19	0.09	51%	0.18	0.07	60%	0.04	0.04	5%
		10	0.36	0.11	70%	0.38	0.11	72%	0.36	0.09	76%	0.08	0.07	8%
		15	0.55	0.12	77%	0.57	0.13	78%	0.55	0.10	82%	0.12	0.11	9%
		20	0.72	0.14	80%	0.76	0.15	80%	0.73	0.11	84%	0.16	0.14	11%
		25	0.89	0.16	82%	0.95	0.17	82%	0.93	0.13	86%	0.21	0.18	13%
		30	1.07	0.18	83%	1.15	0.19	84%	1.11	0.15	87%	0.26	0.22	15%
	7	5	1.18	0.54	55%	0.85	0.41	51%	0.87	0.34	61%	0.06	0.05	12%
		10	2.08	0.62	70%	1.71	0.45	74%	1.73	0.37	79%	0.13	0.10	21%
		15	3.10	0.57	81%	2.58	0.49	81%	2.62	0.39	85%	0.21	0.15	26%
		20	4.09	0.62	85%	3.40	0.53	85%	3.49	0.43	88%	0.31	0.21	30%
		25	5.06	0.74	85%	4.27	0.56	87%	4.37	0.46	89%	0.42	0.28	34%
		30	5.81	0.82	86%	5.11	0.60	88%	5.26	0.50	90%	0.54	0.34	37%
L	3	5	0.04	0.02	46%	0.04	0.02	45%	0.04	0.02	53%	0.03	0.03	0%
		10	0.08	0.04	56%	0.08	0.04	56%	0.08	0.03	63%	0.06	0.06	4%
		15	0.12	0.05	59%	0.12	0.05	60%	0.12	0.04	68%	0.09	0.08	2%
		20	0.16	0.06	60%	0.16	0.06	61%	0.15	0.05	68%	0.11	0.11	3%
		25	0.20	0.08	62%	0.21	0.08	61%	0.19	0.06	69%	0.14	0.14	4%
		30	0.24	0.09	62%	0.25	0.10	60%	0.23	0.07	70%	0.17	0.16	3%

5	5	0.19	0.08	56%	0.18	0.09	54%	0.18	0.08	57%	0.04	0.04	5%
	10	0.36	0.10	73%	0.38	0.10	73%	0.37	0.09	77%	0.08	0.07	8%
	15	0.54	0.12	78%	0.56	0.12	78%	0.54	0.10	81%	0.12	0.11	9%
	20	0.72	0.14	81%	0.75	0.14	81%	0.72	0.11	84%	0.17	0.15	12%
	25	0.91	0.15	83%	0.94	0.16	83%	0.92	0.13	86%	0.22	0.18	15%
	30	1.05	0.18	83%	1.13	0.19	83%	1.12	0.15	87%	0.26	0.22	15%
7	5	0.81	0.38	53%	0.85	0.39	54%	0.84	0.33	61%	0.06	0.05	8%
	10	1.71	0.40	76%	1.71	0.41	76%	1.66	0.35	79%	0.13	0.10	21%
	15	2.43	0.44	82%	2.55	0.45	83%	2.52	0.38	85%	0.21	0.16	27%
	20	3.24	0.47	85%	3.43	0.49	86%	3.36	0.42	88%	0.31	0.22	31%
	25	4.00	0.50	87%	4.29	0.54	88%	4.22	0.45	89%	0.44	0.28	35%
	30	4.73	0.57	88%	5.11	0.57	89%	5.07	0.50	90%	0.55	0.35	37%

Note. B = bank quality; K = number of attributes measured by the test; T = test length; KL = Kullback–Leibler Index method; PWKL = posterior weighted Kullback–Leibler information method; SHE = Shannon entropy method; GDI = the generalized deterministic inputs, noisy ‘‘and’’ gate (G-DINA) discrimination index; original = original methods (KL, PWKL, SHE, GDI); LBPS = methods adding LBPS (LBPS-KL, LBPS-PWKL, LBPS-SHE, LBPS-GDI); PR = the percentage of reduction in computation time after adding LBPS.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

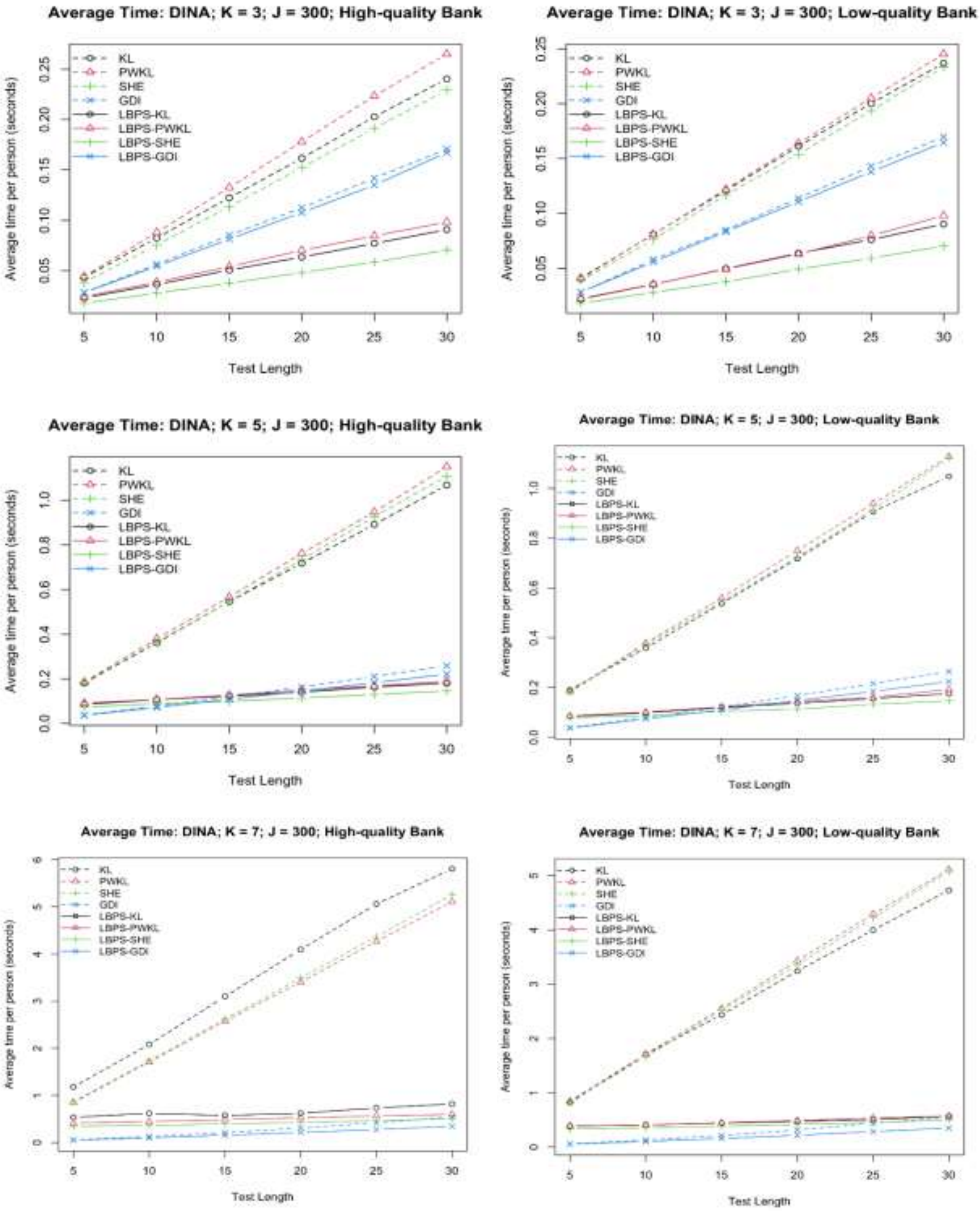


Figure 4. Average computation time per person under DINA when J = 300 items

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

4.4 Mean Test-Overlap Rates

Table 4 and Figure 5 summarize the test-overlap rates for LBPS-integrated and original methods. In Table 4, *D* represents the difference in overlap rates (Original − LBPS), where positive values indicate that LBPS reduced overlap and negative values indicate increases. Under high-quality item banks, LBPS increased overlap rates when attribute dimensionality was low (K = 3), particularly for GDI (e.g., D = −0.24 at T = 30) and to a lesser extent for KL and PWKL at longer test lengths (e.g., D = −0.13 to −0.14). SHE remained largely unaffected. At K = 5, differences were mixed but small (mostly within ±0.05). At K = 7, LBPS reduced overlap at shorter test lengths (e.g., D = 0.15–0.17 at T = 5 for KL, PWKL, and SHE), with diminishing effects as test length increased. Under low-quality item banks, LBPS generally reduced overlap for K = 5 and 7 across all methods. The largest differences appeared for KL (e.g., D = 0.18 at T = 15 and 20 for K = 5), with modest and consistent reductions for PWKL, SHE, and GDI. At K = 3, differences were smaller and mixed, but slightly favored LBPS (e.g., D = 0.09 at T = 5 for KL).

Overall, LBPS affected test-overlap rates differently across conditions: Overlap increased for high-quality banks with small *K* (particularly for GDI), but decreased under more challenging conditions—low-quality banks or large *K*. Even though the benefit of LBPS in reducing test overlap rate does not universally apply to all conditions, it is clearly effective when LBPS is needed the most: i.e., when *K* is large. Moreover, given the goal of cognitive diagnosis assessments is to support formative assessment and immediate feedback (Leighton & Gierl, 2007; Rupp et al., 2010), CD-CAT is typically considered in a low-stakes context, with classification accuracy and computational efficiency being the primary concerns rather than exposure control.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

Table 4. Mean Test Overlap Rate Differences under DINA (J = 300)

			KL			PWKL			SHE			GDI		
B	K	T	Original	LBPS	D	Original	LBPS	D	Original	LBPS	D	Original	LBPS	D
H	3	5	0.55	0.56	-0.01	0.62	0.56	0.06	0.57	0.59	-0.02	0.67	0.70	-0.03
		10	0.49	0.56	-0.07	0.53	0.56	-0.03	0.44	0.46	-0.02	0.60	0.75	-0.15
		15	0.48	0.61	-0.13	0.50	0.61	-0.11	0.47	0.49	-0.02	0.57	0.76	-0.19
		20	0.50	0.62	-0.12	0.49	0.62	-0.13	0.49	0.51	-0.02	0.55	0.77	-0.22
		25	0.50	0.63	-0.13	0.49	0.63	-0.14	0.52	0.53	-0.01	0.55	0.77	-0.22
		30	0.52	0.64	-0.12	0.50	0.64	-0.14	0.53	0.54	-0.01	0.54	0.78	-0.24
	5	5	0.56	0.46	0.10	0.59	0.46	0.13	0.61	0.59	0.02	0.60	0.59	0.01
		10	0.52	0.49	0.03	0.51	0.48	0.03	0.39	0.39	0.00	0.53	0.59	-0.06
		15	0.51	0.50	0.01	0.48	0.50	-0.02	0.36	0.38	-0.02	0.50	0.60	-0.10
		20	0.51	0.52	-0.01	0.48	0.52	-0.04	0.40	0.41	-0.01	0.49	0.61	-0.12
		25	0.50	0.53	-0.03	0.47	0.53	-0.06	0.45	0.47	-0.02	0.48	0.63	-0.15
		30	0.50	0.54	-0.04	0.48	0.55	-0.07	0.49	0.51	-0.02	0.48	0.63	-0.15
	7	5	0.47	0.32	0.15	0.50	0.33	0.17	0.68	0.53	0.15	0.53	0.53	0.00
		10	0.46	0.38	0.08	0.48	0.38	0.11	0.45	0.37	0.08	0.45	0.44	0.01
		15	0.44	0.42	0.02	0.45	0.42	0.03	0.37	0.32	0.05	0.45	0.46	-0.01
		20	0.43	0.44	-0.01	0.44	0.44	0.01	0.36	0.34	0.01	0.44	0.48	-0.04
		25	0.44	0.45	-0.02	0.44	0.45	-0.01	0.39	0.39	0.00	0.44	0.50	-0.06
		30	0.44	0.46	-0.02	0.44	0.46	-0.02	0.44	0.45	-0.01	0.44	0.51	-0.08
L	3	5	0.68	0.59	0.09	0.59	0.59	0.00	0.70	0.55	0.15	0.68	0.66	0.02
		10	0.66	0.63	0.03	0.61	0.63	-0.02	0.46	0.40	0.06	0.65	0.63	0.02
		15	0.66	0.63	0.03	0.60	0.63	-0.03	0.39	0.34	0.05	0.62	0.63	-0.01
		20	0.68	0.65	0.03	0.59	0.64	-0.05	0.35	0.32	0.03	0.61	0.65	-0.04
		25	0.70	0.66	0.04	0.60	0.65	-0.05	0.33	0.30	0.03	0.60	0.66	-0.06
		30	0.72	0.67	0.05	0.61	0.66	-0.05	0.33	0.30	0.03	0.61	0.66	-0.05

5	5	0.67	0.62	0.05	0.65	0.62	0.03	0.72	0.67	0.05	0.76	0.67	0.09
	10	0.68	0.55	0.13	0.54	0.55	-0.01	0.49	0.43	0.06	0.61	0.54	0.07
	15	0.71	0.53	0.18	0.53	0.52	0.01	0.41	0.36	0.05	0.57	0.52	0.05
	20	0.71	0.53	0.18	0.53	0.53	0.00	0.38	0.32	0.06	0.56	0.52	0.04
	25	0.70	0.53	0.17	0.53	0.53	0.00	0.35	0.30	0.05	0.56	0.53	0.03
	30	0.71	0.54	0.17	0.54	0.54	0.00	0.33	0.29	0.04	0.56	0.53	0.03
7	5	0.69	0.64	0.05	0.67	0.64	0.03	0.89	1.00	-0.11	0.93	1.00	-0.07
	10	0.64	0.60	0.04	0.63	0.61	0.02	0.68	0.59	0.09	0.69	0.64	0.05
	15	0.62	0.55	0.07	0.55	0.55	0.01	0.55	0.45	0.10	0.62	0.56	0.06
	20	0.62	0.52	0.10	0.54	0.53	0.01	0.47	0.38	0.09	0.59	0.52	0.08
	25	0.62	0.51	0.11	0.52	0.51	0.02	0.42	0.35	0.07	0.58	0.50	0.08
	30	0.63	0.49	0.14	0.52	0.50	0.03	0.39	0.36	0.03	0.57	0.48	0.08

Note. B = bank quality; K = number of attributes measured by the test; T = test length; KL = Kullback–Leibler Index method; PWKL = posterior weighted Kullback–Leibler information method; SHE = Shannon entropy method; GDI = the generalized deterministic inputs, noisy ‘‘and’’ gate (G-DINA) discrimination index; original = original methods (KL, PWKL, SHE, GDI); LBPS = methods adding LBPS (LBPS-KL, LBPS-PWKL, LBPS-SHE, LBPS-GDI); D = Original – LBPS; positive D values indicate LBPS reduced overlap rates; negative D values indicate increases.

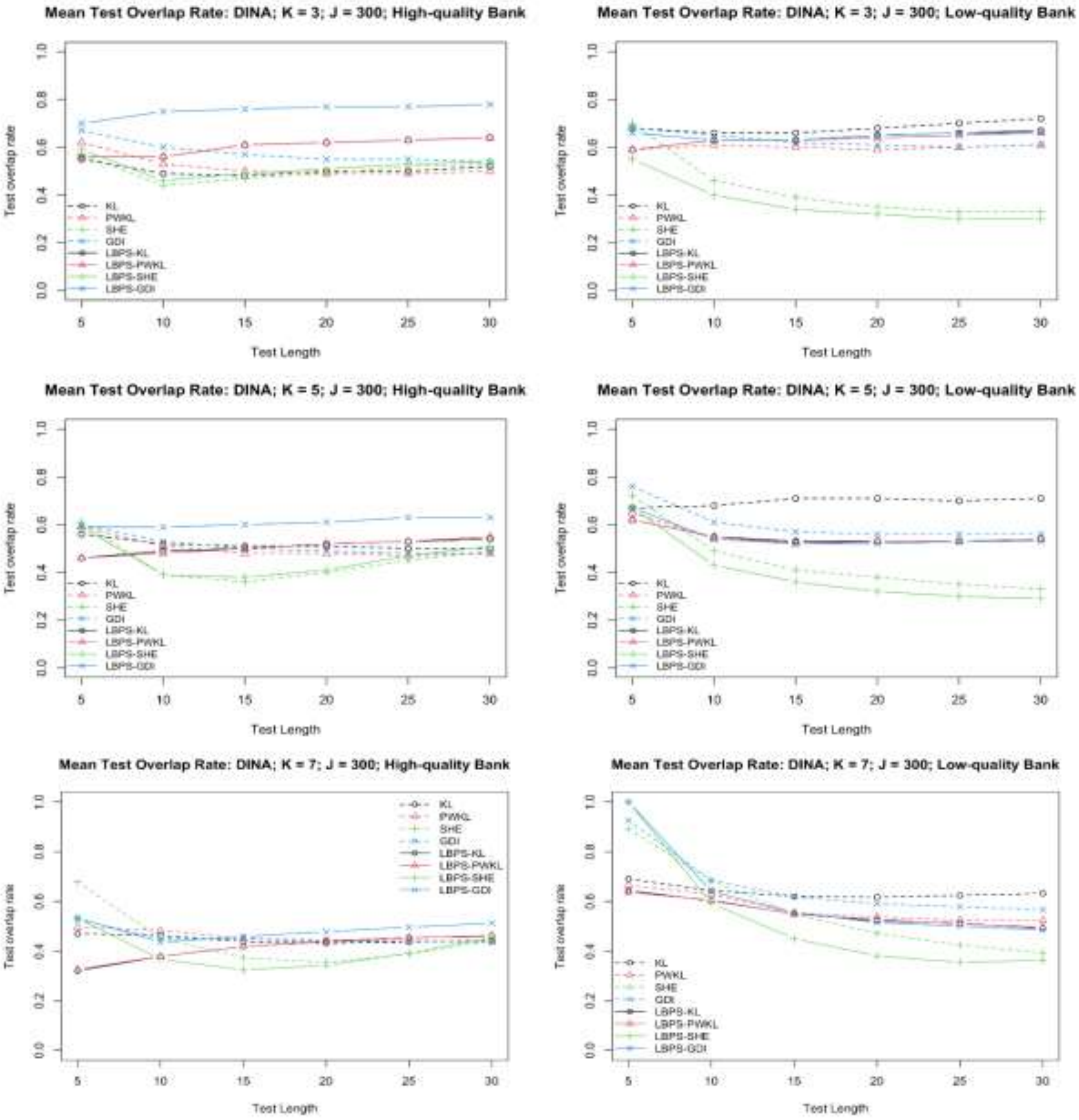


Figure 5. Mean test overlap rate results under DINA when J = 300 items

Additional simulations with DINA (500-item banks) and DINO (both 300-item and 500-item banks) (see Appendix) also found substantial computational gains in conditions with large *K* and/or *T*, without compromising classification accuracy or test security, lending further support of the general applicability of LBPS. That said, there are some nuanced differences. For example, under the DINA model, there is a substantial gain in classification accuracy when

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

LBPS is coupled with KL, compared to the original KL. The improvement in PAR is not nearly substantial under the DINO model.

In summary, results demonstrate that LBPS, when combined with a variety of item selection methods, successfully accelerates CD-CAT in the most computationally demanding scenarios (large *K* and *T*), without compromising classification accuracy or test overlap rates, making it particularly well-suited for complex diagnostic assessments.

5. Discussion

This study introduced the Likelihood-Based Profile Shrinkage (LBPS) algorithm to improve computational efficiency in CD-CAT. LBPS works by focusing item selection on the most probable attribute profiles at each test stage. As more items are administered, the posterior distribution over profiles becomes concentrated, allowing LBPS to exclude highly improbable profiles from consideration. While traditional methods evaluate every possible profile, LBPS uses a reduced working set, leading to faster computations with minimal trade-offs in accuracy. Simulations confirmed that LBPS maintains comparable AAR and PAR values while substantially reducing computational time, particularly as the number of attributes or test length increases. LBPS had mixed effects on test-overlap rates, but generally maintained test security when *K* is large.

Note that although LBPS begins after the first item response, it does not constrain attribute estimation. The full set of profiles is always used to compute likelihoods and update mastery estimates. LBPS simply filters out low-likelihood profiles during item selection, without narrowing the estimation space. This preserves diagnostic accuracy, even in early stages when estimation is less stable.

Moreover, the algorithm is highly flexible and can be effectively integrated with existing CD-CAT item selection methods (e.g., KL, PWKL, SHE, GDI). This extends prior work on CD-CAT efficiency (Kaplan et al., 2015) by offering a generalizable framework applicable across selection strategies. However, our simulations reveal differential benefits across methods: computation speed gains are more pronounced for SHE, KL, and PWKL than for GDI. This is mainly because GDI already includes procedures that reduce computational burden by operating on a reduced set of attribute patterns, thereby diminishing the marginal benefit of LBPS.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

Although developed under the DINA model, LBPS is model-agnostic. Because it ranks attribute patterns based on likelihoods, it can be extended to any CDM by modifying the response probability function. While different CDMs create different partition structures—for example, DINA's conjunctive rule (requiring all attributes) versus DINO's disjunctive rule (requiring at least one attribute)—the core logic of LBPS remains applicable.

Adapting LBPS to other CDMs involves substituting the response model in the likelihood calculation and pairing the reduced pattern set with an item selection index. Preliminary results under DINO demonstrate substantial computation time reductions with negligible loss in accuracy (see Appendix), comparable to those achieved under DINA. Thus, LBPS provides a robust and scalable strategy for accelerating CD-CAT across different CDMs.

On the other hand, this study has limitations. While we demonstrated LBPS under DINA and DINO, future work should test its performance under other models and item types. Our simulations assumed uniform priors and did not explore correlated attributes or alternative Q-matrix structures. In educational contexts, hierarchical attributes—where one skill is a prerequisite for another—are common. Incorporating such hierarchies may further improve efficiency. Future work should also examine higher-dimensional scenarios (e.g., $K > 7$) to assess scalability.

For operational implementation, research should examine three practical aspects: (1) exposure control mechanisms to prevent item overuse, (2) attribute balancing strategies when using reduced profile sets, and (3) variable-length termination criteria for adaptive test length. For example, test developers can implement LBPS as a first-stage filter to identify promising items, then apply exposure and content control constraints as subsequent selection criteria (Cheng, 2010; Li et al., 2021; Lin & Chang, 2019; Wang et al., 2011). Validation using real item banks will further assess robustness.

Moreover, because the shrinkage of the maximum likelihood pattern set (Case 1 of Theorem 2) occurs predominantly in the early stages of an assessment, a hybrid approach may be advantageous: apply LBPS during the initial phase to reduce $|M(\boldsymbol{x}_t)|$ from exponentially large to a manageable size, then maintain this reduced pattern set for all subsequent item selections. Once $|M(\boldsymbol{x}_t)|$ becomes sufficiently small (e.g., 20 patterns), the computational overhead of updating likelihoods and maintaining $M(\boldsymbol{x}_t)$ after each response may outweigh the benefits of further

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

reduction. By switching to a fixed pattern set at this point, we eliminate update costs while preserving the efficiency gained from evaluating items against only a small subset of patterns rather than the full 2^K space. The optimal switching point—when to transition from dynamic LBPS to a static pattern set—likely depends on multiple factors including the number of attributes K, the computational cost of likelihood calculations, and the specific item selection method employed, warranting future investigation.

Computational efficiency in CD-CAT can also be improved through programming optimizations. For instance, PWKL and SHE require updating likelihood functions and posterior probabilities after each item response. Caching likelihood values from previous steps, rather than recalculating them entirely, can reduce redundant computations. Notably, LBPS and programming optimizations operate at different levels: programming optimizations reduce redundant calculations within a fixed computational framework, whereas LBPS reduces the search space itself from 2^K patterns to a smaller working set. These approaches are complementary, and practitioners can combine LBPS with strategies such as likelihood caching to achieve additional efficiency gains.

In sum, LBPS provides a computationally efficient enhancement to CD-CAT that maintains diagnostic precision. Its flexibility, scalability, and compatibility with existing methods make it well suited for modern adaptive assessments. We recommend using LBPS in CD-CAT when K is large, as this is when computational efficiency becomes a primary concern and its benefits are most pronounced.

Code Availability

The example materials and code implementing the LBPS algorithm under the DINA model can be found at https://osf.io/pnavk/files.

Funding statement: The research is supported by NSF OAC #2321054.

Competing interests: The author(s) declare none.

References

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Chen, S. Y., Ankenmann, R. D., & Spray, J. A. (2003). The relationship between item exposure and test overlap in computerized adaptive testing. *Journal of Educational Measurement*, *40*(2), 129-145. https://doi.org/10.1111/j.1745-3984.2003.tb01100.x

Choe, E. M., Kern, J. L., & Chang, H. H. (2018). Optimizing the use of response times for item selection in computerized adaptive testing. *Journal of Educational and Behavioral Statistics*, *43*(2), 135-158. https://doi.org/10.3102/1076998617723642

Chang, Y.-P., Chiu, C.-Y., & Tsai, R.-C. (2019). Nonparametric CAT for CD in educational settings with small samples. *Applied Psychological Measurement*, *43*(7), 543–561. https://doi.org/10.1177/0146621618813113

Cheng, Y. (2010). Improving cognitive diagnostic computerized adaptive testing by balancing attribute coverage: The modified maximum global discrimination index method. *Educational and Psychological Measurement*, *70*(6), 902-913. https://doi.org/10.1177/0013164410366693

Cheng, Y. (2009). When cognitive diagnosis meets computerized adaptive testing: CD-CAT. *Psychometrika*, *74*(4), 619-632. https://doi.org/10.1007/s11336-009-9123-2

Chiu, C. Y., & Chang, Y. P. (2021). Advances in CD-CAT: The general nonparametric item selection method. *psychometrika*, *86*(4), 1039-1057. https://doi.org/10.1007/s11336-021-09792-z

Cover, T. M., & Thomas, J. A. (1991) Elements of information theory. Wiley & Sons, New York. https://doi.org/10.1002/0471200611

Dai, B., Zhang, M., & Li, G. (2016). Exploration of item selection in dual-purpose cognitive diagnostic computerized adaptive testing: Based on the RRUM. *Applied Psychological Measurement*, *40*(8), 625-640. https://doi.org/10.1177/0146621616666008

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

de la Torre, J. (2009). DINA model and parameter estimation: A didactic. *Journal of educational and behavioral statistics*, *34*(1), 115-130. https://doi.org/10.3102/1076998607309474

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*(3), 333–353. https://doi.org/10.1007/BF02295640

de la Torre, J. (2011). The generalized DINA model framework. *Psychometrika*, *76*(2), 179–199. https://doi.org/10.1007/s11336-011-9207-7

Henson, R. A., Templin, J. L., & Willse, J. T. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, *74*(2), 191-210. https://doi.org/10.1007/s11336-008-9089-5

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement*, *25*(3), 258–272. https://doi.org/10.1177/01466210122032064

Kang, H., Zhang, S., & Chang, H. (2017). Dual-objective item selection criteria in cognitive diagnostic computerized adaptive testing. *Journal of Educational Measurement*, *54*(2), 165–183. https://doi.org/10.1111/jedm.12139

Kaplan, M., De La Torre, J., & Barrada, J. R. (2015). New item selection methods for cognitive diagnosis computerized adaptive testing. *Applied Psychological Measurement*, *39*(3), 167–188. https://doi.org/10.1177/0146621614554650

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86. https://doi.org/10.1214/aoms/1177729694

Leighton, J. P., & Gierl, M. J. (2007). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, *26*(2), 3–16. https://doi.org/10.1111/j.1745-3992.2007.00090.x

Leighton, J., & Gierl, M. (2007). Cognitive diagnostic assessment for education: Theory and applications. Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

Li, J., Ma, L., Zeng, P., & Kang, C. (2021). New item selection method accommodating practical constraints in cognitive diagnostic computerized adaptive testing: Maximum deviation

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

and maximum limitation global discrimination indexes. *Frontiers in Psychology*, *12*, 619771. https://doi.org/10.3389/fpsyg.2021.619771

Lin, C.-J., & Chang, H.-H. (2019). Item selection criteria with practical constraints in cognitive diagnostic computerized adaptive testing. *Educational and Psychological Measurement*, *79*(2), 335–357. https://doi.org/10.1177/0013164418790634

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika*, *64*(2), 187–212. https://doi.org/10.1007/BF02294535

Minchen, N. D., & de la Torre, J. (2016). *The continuous G-DINA model and the Jensen-Shannon divergence.* Paper presented at the *International Meeting of the Psychometric Society*, Asheville, NC.

Morris, R., Perry, T., & Wardle, L. (2021). Formative assessment and feedback for learning in higher education: A systematic review. *Review of Education*, *9*(3), e3292. https://doi.org/10.1002/rev3.3292

Ravand, H., & Baghaei, P. (2020). Diagnostic classification models: Recent developments, practical issues, and prospects. *International Journal of Testing*, *20*(1), 24–56. https://doi.org/10.1080/15305058.2019.1588278

Rupp, A. A., Templin, J., & Henson, R. A. (2010). *Diagnostic measurement: Theory, methods, and applications*. Guilford Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Tatsuoka, C. (2002). Data analytic methods for latent partially ordered classification models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *51*(3), 337–350. https://doi.org/10.1111/1467-9876.00272

Tatsuoka, K. K. (1995). Architecture of knowledge structure and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), Cognitively diagnostic assessment (pp. 327-361). Hillsdale, NJ: Lawrence Erlbaum.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

Templin, J. L., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods*, *11*(3), 287–305. https://doi.org/10.1037/1082-989X.11.3.287

von Davier, M. (2005). A general diagnostic model applied to language testing data. *ETS Research Report Series*, *2005*(2), i–35. https://doi.org/10.1002/j.2333-8504.2005.tb01993.x

Wang, C., Chang, H.-H., & Huebner, A. (2011). Restrictive stochastic item selection methods in cognitive diagnostic computerized adaptive testing: Restrictive stochastic methods in CD-CAT. *Journal of Educational Measurement*, *48*(3), 255–273. https://doi.org/10.1111/j.1745-3984.2011.00145.x

Wang, C. (2013). Mutual information item selection method in cognitive diagnostic computerized adaptive testing with short test length. *Educational and Psychological Measurement*, *73*(6), 1017–1035. https://doi.org/10.1177/0013164413498256

Wang, C., Chang, H.-H., & Douglas, J. (2012). Combining CAT with cognitive diagnosis: A weighted item selection approach. *Behavior Research Methods*, *44*(1), 95–109. https://doi.org/10.3758/s13428-011-0143-3

Wang, C., Zheng, C., & Chang, H. (2014). An enhanced approach to combine item response theory with cognitive diagnosis in adaptive testing. *Journal of Educational Measurement*, *51*(4), 358–380. https://doi.org/10.1111/jedm.12057

Wang, W., Song, L., Wang, T., Gao, P., & Xiong, J. (2020). A note on the relationship of the Shannon entropy procedure and the Jensen–Shannon divergence in cognitive diagnostic computerized adaptive testing. *Sage Open*, *10*(1), 2158244019899046. https://doi.org/10.1177/2158244019899046

Xu, G., Wang, C., & Shang, Z. (2016). On initial item selection in cognitive diagnostic computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *69*(3), 291–315. https://doi.org/10.1111/bmsp.12072

Xu, X., Chang, H.-H., & Douglas, J. (2003). *Computerized adaptive testing strategies for cognitive diagnosis.* Paper presented at *the annual meeting of National Council on Measurement in Education*, Montreal, Canada.

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press

Yigit, H. D., Sorrel, M. A., & De La Torre, J. (2019). Computerized adaptive testing for cognitively based multiple-choice data. *Applied Psychological Measurement*, *43*(5), 388–401. https://doi.org/10.1177/0146621618798665

Zheng, C., & Chang, H.-H. (2016). High-efficiency response distribution–based item selection algorithms for short-length cognitive diagnostic computerized adaptive testing. *Applied Psychological Measurement*, *40*(8), 608–624. https://doi.org/10.1177/0146621616665196

https://doi.org/10.1017/psy.2026.10086 Published online by Cambridge University Press