Computational Statistics (2026) 41:57
https://doi.org/10.1007/s00180-026-01739-1

ORIGINAL PAPER

A new method for cheating detection during computerized adaptive testing

Luca Bungaro[1,3] · **Mariagiulia Matteucci**[1] · **Stefania Mignani**[1] · **Bernard P. Veldkamp**[2]

Received: 26 February 2025 / Accepted: 5 March 2026

Abstract

In the field of educational and psychological measurement, computerized adaptive testing (CAT) is flexible and convenient, but its reliance on repeatedly administered, pre-calibrated items makes it vulnerable to item exposure and pre-knowledge. We propose a method called CHeater Identification using Interim Person fit Statistic (CHIPS) and a slight modification of it, called Modified CHIPS (M-CHIPS), both designed to identify and limit cheaters during test administration. The methodological novelty lies in redefining a likelihood-based person-fit statistic for response times so that it becomes computable at each adaptive step. CHIPS replaces parameters that traditionally require full-test MCMC estimation with interim maximum-likelihood estimators of speed and expected log-response times, yielding a statistic (IPS) with an analytically tractable asymptotic χ^2 distribution. This allows the IPS to be embedded as a constraint within the Shadow Test Approach, producing a dynamic item-selection algorithm that switches between databases based on real-time evidence of item pre-knowledge. M-CHIPS further introduces an early-stage speed-based intervention to improve detectability under extreme cheating scenarios. A simulation study evaluates estimation accuracy, error rates, and computational performance under varying pre-knowledge levels, ability–speed correlations, and test-length settings. Results show that the proposed methods substantially improve ability estimation for cheaters without affecting non-cheaters, demonstrating the statistical and algorithmic effectiveness of incorporating interim fit statistics into adaptive testing.

Keywords Computerized adaptive testing · Response time · Cheating · Item pre-knowledge · Interim person fit statistic

Extended author information available on the last page of the article

Published online: 17 March 2026

1 Introduction

For a long time, educational assessment primarily relied on traditional paper-and-pencil exams. However, the emergence of personal computers in the late 1980s revolutionized testing by facilitating computer-based testing (CBT). Another historical moment that led to a massive use of CBT was the Covid-19 pandemic, which, along with distance learning (DL), also created the need to assess students through online platforms (Iannario et al. 2024). Among its many advantages, CBT allows for flexible scheduling, immediate feedback, and the use of extensive item banks, significantly improving the efficiency of test administration (van Krimpen-Stoop and Meijer 2000; Wainer et al. 2000). Moreover, digital testing enables the collection of process data such as response times (RTs), response sequences, and revisit behaviors, offering deeper insights into test-taker performance beyond mere correctness.

One of the most innovative developments in CBT is Computerized Adaptive Testing (CAT), which personalizes test difficulty based on a test-taker's responses. This adaptive approach optimizes measurement accuracy while reducing the number of items required to estimate ability levels. However, despite its many benefits, CAT is vulnerable to a significant threat: cheating. In high-stakes testing, test security is paramount, as test-takers may attempt to exploit vulnerabilities in CAT to gain unfair advantages. For example, in 2002, there was a massive use of information from some sites known as "brain dumb sites" in China, Taiwan, and South Korea to cheat on the graduate record exam (GRE). The spread was so extensive that the authorities had to stop the CAT administration and return to pen-and-paper tests (Cizek and Wollack 2016).

Cheating in computerized testing can take various forms, from traditional misconduct like copying from peers to more sophisticated methods such as item pre-knowledge. In the context of CAT, due to its adaptive nature, many traditional cheating methods do not work, such as copying from a nearby test taker, as their tests will almost certainly contain different questions. However, item pre-knowledge poses a severe risk. Pre-knowledge refers to a situation in which a test taker has obtained the correct answers to some or all of the items in the item bank used to generate the tests (Cizek and Wollack 2016).

Typically, such information is extracted directly from item bankfollowing hacker attacks. In this case, it is referred to as an information leak, and the item bank is considered compromised (Stocking and Lewis 1998; Sympson and Hetter 1985). However, this is not the only way pre-knowledge can occur. The repeated use of the same items to generate tests can lead to a "natural" compromise of the item bank, for example, through word-of-mouth among test takers. This represents a major challenge for CAT, because when individuals access this information before the test, they can respond correctly regardless of their true ability, leading to artificially inflated scores and invalid assessment outcomes.

Multiple approaches have been developed to detect cheaters. Some rely on response patterns to identify inconsistencies indicative of misconduct (Wollack and Maynes 2016), while others specifically target item pre-knowledge using statistical methods (Sinharay 2017). Additionally, response time-based detection techniques

have gained popularity, as cheaters tend to answer known items significantly faster than expected (Fox and Marianti 2017; Marianti et al. 2014).

Machine learning techniques have also been explored for their potential in identifying anomalous behaviors indicative of cheating (Ranger et al. 2023; Zhou and Jiao 2023).

Despite these advances, most existing methods share a fundamental limitation: they identify cheating only after the test has concluded. This retrospective approach leaves test administrators with limited options, such as invalidating scores or requiring additional assessments.

Consequently, there is a pressing need for proactive cheating detection methods that can intervene during the test administration itself.

This paper introduces an innovative method called CHeater identification using Interim Person fit Statistic (CHIPS). This method uses information derived from RTs to define an interim person fit statistic (IPS) to identify potential cheaters during test administration. Subsequently, leveraging the adaptive nature of CAT, potential cheaters are administered items from a more secure database. In this way, the possibility of item pre-knowledge is reduced. The performance of the proposed method is tested via a simulation study. The simulation was designed to replicate a scenario in which multiple students participate in a CAT.

However, a portion of them is simulated as cheaters with pre-knowledge, meaning individuals who answer certain test items correctly and very quickly, regardless of their actual ability. The CHIPS method is incorporated into the simulation, calculating the IPS in real-time, thereby adjusting item selection and responses within the simulated test.

After analyzing the results, a modification to overcome situations with constant speed is proposed and illustrated: the Modified CHIPS (M-CHIPS). Also the M-CHIPS is tested via a simulation study, by evaluating different testing conditions, such as different significance levels, different ability distribution, fixed or variable test length and correlation between speed and ability.

The paper is organized as follows. First, the framework of the CAT and the main methods to identify cheaters using RT are introduced. Then, the CHIPS method is proposed. Afterward, the results of a simulation study are presented in order to investigate the main properties and limitations of the proposed method. Based on the limitations identified in this simulation, the M-CHIPS is introduced and tested. Some concluding remarks and suggestions for future developments end the paper.

2 Methods

This section will provide the fundamental elements for the development of the proposed new methodologies. The first part will introduce the Item Response Theory (IRT) and the CAT, delving into the various processes that compose it, as it represents the reference framework within which the new methodologies will be applied and integrated. Subsequently, there will be an examination of the concept of response time, its most commonly hypothesized distribution (the log-normal), and how it is used to identify anomalous behaviors (particularly cheating). Only in the final part of

this section will the two new methods be introduced: CHIPS and M-CHIPS, describing their methodology and operation in detail.

2.1 Item response theory

Item Response Theory (IRT) is a psychometric framework that models the probability of a correct response as a function of a latent ability and item parameters, enabling the measurement of individuals' proficiency and item properties (Lord and Novick 1968).

Under the assumption that a single trait is latent in the item response process (unidimensional model), that the response variable is dichotomous ($Y = 0\, or\, 1$), and without considering the guessing parameter, the probability of a correct response to a test item, called the Item Response Function (IRF), takes the following form:

$$P\left(Y_{nk} = 1|\theta_n\right) = \frac{\exp[a_k(\theta_n - b_k)]}{1+\exp[a_k(\theta_n - b_k)]}, \tag{1}$$

where *n* refers to the test taker, while *k* refers to the item. The IRF in Eq. (1) is called two-parameter logistic (2PL) model, and it is just one of the possible IRFs that can be hypothesized for a response process according to IRT. However, this IRF, as will also be specified in the section dedicated to the simulation study, will be the one used in this study. Not only can the IRF take on different forms (for different parameters considered or the type of equation, which can be either logistic or normal ogive), but the methods for estimating the θ_n parameter (assuming the item parameters are known) can also vary. The two most commonly used methods are maximum likelihood estimation (MLE) and the Bayesian approach (Martinková and Hladká, 2023; van der Linden and Glas 2010), described in Appendix A.

It is worth noting that, although parameter estimation in IRT is often centered on calibrating item parameters a_k and b_k, typically by conditioning or marginalizing out person ability, this is not the focus here. In a CAT environment, items already have been calibrated before operational administration, since real-time item parameter estimation is not feasible during adaptive testing. Therefore, in this work the item parameters are assumed to be known, and estimation efforts are devoted exclusively to the examinee's ability θ_n.

2.2 Computerized adaptive testing

With the advent of computer-based testing (CBT), assessments can be delivered more flexibly and efficiently, while also enabling advanced approaches such as Computerized Adaptive Testing (CAT), which improves the precision of ability estimation (van der Linden and Glas 2010).

In Computerized Adaptive Testing (CAT), the test adapts to the examinee's ability, updating the estimate after each response and selecting items to maximize precision.

Although the idea of adapting questions dates back to early practices (Binet and Simon 1905), the development of Item Response Theory provided its modern psychometric foundation. With advances in computing, CAT became feasible for large-

scale testing, offering advantages such as flexible scheduling, faster scoring, and more diverse item formats.

Compared to linear tests, where item selection, calibration, and scoring are separate and time-consuming processes, CAT performs item selection and ability estimation in real time. This requires algorithms to replace many roles traditionally held by test experts and psychometricians. Unlike linear testing, where these steps are distinct, CAT integrates them into a continuous, interdependent process: item choice depends on the current ability estimate, which in turn depends on previously selected items. In practice, CAT operates through iterative steps (from $m = 1$to M): administering an item, updating ability, checking stopping rules, and selecting the next item or ending the test, as summarized in Figure 1 and elaborated in appendix B.

2.3 Cheating detection using response time

To use response time (RT) for identifying cheaters, we must first hypothesize a distribution model. Several models have been proposed in the literature. Among them, one of the most popular models assumes a log-normal distribution of RT (van der Linden 2006). This model is based on the idea that, given a subject $n = 1, ..., N$ and an item $k = 1, ..., K$, the response time RT_{nk} is a realization of a random variable RT_{nk}, which follows a log-normal distribution.

Therefore, its probability density function is given by:

$$f\left(rt_{nk}, \zeta_n, \varphi_k, \lambda_k\right) = \frac{\varphi_k}{rt_{nk}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\varphi_k\left(\ln rt_{nk} - \left(\lambda_k - \zeta_n\right)\right)\right]^2\right\}, \tag{2}$$

where: λ_k is the time-intensity parameter of item k and represents the population-average time (on a logarithmic scale) needed to complete that item; ζ_n is the speed parameter of test taker n, which represents the latent trait underlying the response time; φ_k is the time-discrimination parameter of item k, representing the item sensitivity to different speed levels of test takers and is defined as the reciprocal of the standard deviation of the measurement error

$$\varphi_k = \frac{1}{\sigma_{\epsilon_k}}. \tag{3}$$

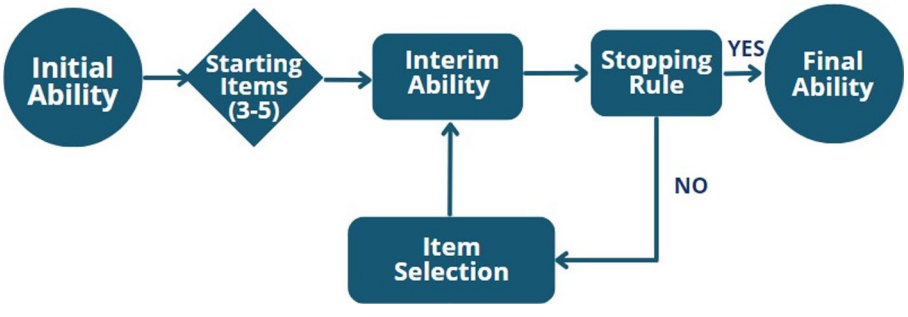

Fig. 1 Flow chart of a CAT process

From Eq. (2) follows that the logarithm of RT can be expressed using the following probabilistic function:

$$\ln RT_{nk} = \lambda_k - \zeta_n + \epsilon_{nk},$$
$$\varepsilon_{nk} \sim \mathrm{N}\left(0, \sigma^2_{\epsilon_k}\right), \tag{4}$$

where ϵ_{nk} is the error component and can capture fluctuations in RTs resulting from the random actions of a test taker, such as insert brief pauses during the test or alter their time management.

The error component specific to each item can account for these distinctions and prevents any bias in the parameter estimates.

Equation (4) was later modified by Fox and Marianti (2016), who proposed a log-normal model where the time-discrimination parameter ϕ^* is the slope of the speed,

$$\ln RT_{nk} = \lambda_k - \varphi^*_k \zeta_n + \epsilon_{nk},$$
$$\epsilon_{nk} \sim N\left(0, \sigma_{\epsilon^2_k}\right). \tag{5}$$

The time-discrimination parameter φ^* also models the covariances between RTs, which is expected to enhance the model flexibility (Fox and Marianti 2016). This follows from the fact that the covariance between the RT to item k and l of person n includes the time discriminations, which are given by:

$$\begin{aligned} cov\left(T_{nk}, T_{nl}\right) &= cov\left(\lambda_k - \varphi^*_k \zeta_n + \epsilon_{nk},\ \lambda_l - \varphi^*_l \zeta_n + \epsilon_{nl}\right) \\ &= cov\left(\varphi^*_k \zeta_n + \epsilon_{nk},\ \varphi^*_l \zeta_n + \epsilon_{nl}\right) \\ &= cov\left(\varphi^*_k \zeta_n,\ \varphi^*_l \zeta_n\right) \\ &= \varphi^*_k\, var\left(\zeta_n\right)\ \varphi^*_l \\ &= \varphi^*_k\, \sigma^2_\zeta\, \varphi^*_l. \end{aligned} \tag{6}$$

The RTs can be utilized to develop methods and statistics capable of identifying presumed cheaters and compromised items. Our study focuses on detecting cheaters and, in particular, on the possibility of intervening during the test administration. To achieve this, the paper focused on an existing statistic, later appropriately modified to be used in real-time.

The starting statistic is the person fit statistic l^t_n proposed by Mariant et al. (2014) and Fox and Marianti (2017). It is a statistic that enables the identification of potential cheaters based on the anomalous behaviors in their RTs.

The fundamental idea, supported by the literature (Cizek and Wollack 2016), is that cheaters, especially those who have pre-knowledge, tend to respond to those items much more quickly than expected. In this paper, test takers that use pre-knowledge will be referred as cheaters and all the others will be called non-cheaters. Furthermore, in the paper we focus on multiple-choice items (where the answer variable Y can be made dichotomous, with $Y=1$ indicating the correct answer and $Y=0$ indicating incorrect ones).

The idea behind the person fit statistic l_n^t is to start with the log-likelihood function of the RTs.

to eventually obtain a standardized statistic with a known distribution. The distribution underlying RTs is the one defined by Fox and Marianti (2016) and reported in Eq. (5). The non-standardized statistic is defined as

$$-2\ln L\left(\zeta_n|\overrightarrow{RT}_n\right)=\sum_{k=1}^{K}\left[\left(\frac{\ln \mathrm{RT}_{nk}-\mu_{nk}}{\sigma_k}\right)^2+\ln\left(2\pi\sigma_k^2\right)\right]=\sum_{k=1}^{K}\left[Z_{nk}^2+\ln\left(2\pi\sigma_k^2\right)\right], \quad (7)$$

where $Z_{nk}$is standard normally distributed, since it represents the standardized error of the normally distributed logarithm of RTs. For that reason, the sum of the squares of these standardized errors is, by definition, distributed as X_k^2, with K degrees of freedom. So, the likelihood-based person-fit statistic for RTs, $l_n^t$is defined as:

$$l_n^t=\sum_{k=1}^{K}Z_{nk}=\sum_{k=1}^{K}\left(\frac{\ln \mathrm{RT}_{nk}-\mu_{nk}}{\sigma_{\epsilon_k}}\right)^2, \quad (8)$$

where

$$\mu_{nk}=\lambda_k-\varphi_k^*\zeta_n. \quad (9)$$

It is possible to use significance testing, where α is the significance level and C is the.

associated threshold value for a X_k^2distribution.

If l_n^tin Eq. (8) is greater than C, then the test taker is classified in the group of cheaters and otherwise in the group of non-cheaters. This classification is expressed by a dichotomous variable F_n^t, which takes the value of 1 in the first case and 0 in the second, as follow:

$$F_n^t=\begin{cases}1\ if\ P\left(l_n^t\left(\zeta_n,\boldsymbol{\lambda},\varphi^*,\boldsymbol{\sigma}^2\right)>C\right)\\ 0\ if\ P\left(l_n^t\left(\zeta_n,\boldsymbol{\lambda},\varphi^*,\boldsymbol{\sigma}^2\right)\le C\right)\end{cases}. \quad (10)$$

Finally, using Markov chain Monte Carlo (MCMC) estimation, the dichotomous variable F_n^t is created for each test taker $n=1,\ldots,N$. For each MCMC iteration after the burn-in phase, the proportion of posterior samples in which test taker n is assigned to the cheaters class is denoted as $F_n^{t^*}$, and is used as an estimate of the posterior probability of being a cheater. This statistic can be used either independently or jointly with the person fit statistic for response accuracy (Fox and Marianti, 2017; Marianti et al. 2014).

2.4 A new proposal: the cheater identification using interim person fit statistics (CHIPS) method

The CHeater identification using Interim Person fit Statistic (CHIPS) is a method for the identification of test takers who, during a CAT, are presumed to have pre-knowledge of some or all the items in the item database.

The innovativeness of this method is that it aims at identifying such cheaters while the test is still ongoing, so immediate intervention is possible. The idea is to be able to neutralize the harmful effect that cheating causes, which is to compromise the validity of the test scores of those individuals. The reason this method was designed to primarily address pre-knowledge rather than other types of cheating lies in the fact that it is specifically intended for use in CAT. Due to its adaptive nature, CAT is less susceptible to other cheating techniques, such as copying from a nearby test taker.

The starting point is the l_n^tperson fit statistic in Eq. (8) slightly modified in order to be calculated in real time. The idea is that the new statistic must be calculable in real time during the test administration, and that should be updated on the fly. As already mentioned, the statistic $l_n^t$is defined as a sum of differences that includes all the K test items. These differences are calculated with respect to the estimate of the response speed ζ_n (Eq. 9), that is derived at the end of the test using the Gibbs sampling algorithm (Fox et al. 2021). However, this methodology not only could be time-consuming (depending on the number of test-takers), but in order to be applied it requires having the response patterns and response times of all test takers taking the test, and that they are at the same point in the test as every other test taker; otherwise, it is not possible to apply the MCMC. Clearly, such an assumption is not desirable during a CAT, whose main strengths lie in the immediacy of selecting the next item and in the fact that each test can virtually be different for each respondent (both in terms of specific items and test length/duration).

The proposal of real-time computable statistics provides for the replacement of μ_{nk} and σ_k with parameters that can be calculated at each step m of the test administration (where each step m consists of both the phase of interim ability estimation and that of item selection). The process of replacing μ_{nk} with something that can be individually calculated for each test taker shares some similarities with the approach taken by Sinharay (2018).

Under the hypothesis of log-normal distribution of RTs, such parameters could be the expected response time (Fan et al. 2012; Veldkamp 2016) and the reciprocal time-discrimination(van der Linden 2006), respectively

$$E\left[RT_{nk}|\widehat{\zeta}_{n_m}\right] = \exp\left(\lambda_k - \widehat{\zeta}_{n_m} + \frac{1}{2\varphi_k^2}\right),$$

$$\sigma_{\epsilon_k} = \frac{1}{\phi_k}, \tag{11}$$

$$\widehat{\zeta}_{n_m} = \max_k L\left(\zeta_n\right) = \frac{\sum_{k\in R_m}\left[\varphi_k^2(\lambda_k - lnrt_{nk})\right]}{\sum_{k\in R_m}\left[\varphi_k^2\right]}, \tag{12}$$

where $\widehat{\zeta}_{n_m}$ is the maximum-likelihood estimator (MLE) for the person speed, and R_m is the set of items administered since step m.

As can be seen from Eq. (12), $\widehat{\zeta}_{n_m}$ can be calculated at each step m using the known parameters of the items and the RTs. This means that the adaptive algorithm can easily compute the $\widehat{\zeta}_{n_m}$ statistic in real-time during the test. Once $\widehat{\zeta}_{n_m}$ is calculated, obtaining the expected value $E\left[RT_{nk}|\widehat{\zeta}_{n_m}\right]$ is straightforward (Eq. 11).

Therefore, by replacing μ_{nk} and σ_k with $E\left[RT_{nk}|\widehat{\zeta}_{n_m}\right]$ and σ_{ϵ_k} respectively, a new statistic was defined, called the interim person fit statistic (IPS):

$$l^t_{n_m} = \sum_{k=1}^{m}\left(\frac{lnRT_{nk} - lnE\left[RT_{nk}|\widehat{\zeta}_{n_m}\right]}{\frac{1}{\varphi_k}}\right)^2 \sim \chi^2_m. \tag{13}$$

This statistic in Eq. (13) was demonstrated to follow a X^2 distribution with m degrees of freedom, even in the presence of a plausible number of cheaters (5%). This hypothesis has been tested with a one-sample (N=1000) Kolmogorov-Smirnov (K-S) test, for $m = 35$ ($D = 0.101$; $p - value = 0.27$), and also through graphical representations such as the one in Fig. 2, which shows how the data curve closely matches the density of the corresponding chi-square distribution, except for the tail end on the right, where the simulated cheaters accumulate.

This means that if a significance level α is chosen, the threshold value C_m can be easily found from the χ^2_m distribution.

Individuals with an IPS lower than C_m, falling within the non-rejection region of the null hypothesis, will be referred to as non-cheater, while the remaining individuals will be classified as cheaters.

Once the statistic to be used during the test is defined, a method has also been developed to leverage $l^t_{n_m}$ to interrupt the malicious behavior of cheaters. The idea is not to penalize the cheaters for exploiting pre-knowledge (especially because they might have been misclassified as such), but rather to seek a method that can simultaneously restore both the validity of the test, which cheating may have undermined, and its fairness.

Following the idea of Veldkamp (2016), after a test taker has been flagged as a cheater, the item selection algorithm, through the shadow test approach, will start to administer the next items from a more secure item database (an item bank that has a very low exposure rate and will be more frequently updated), in order to reduce the probability that the cheater has pre-knowledge on those items.

In fact, the use of STA allows for the introduction of constraints for the item selection criterion (ISC) at each step m. The STA operates by generating, whenever a new item must be selected, a complete test, the shadow test, that includes all previously administered items and satisfies all imposed constraints. From this provisional test, only the item to be administered at step $m$is then chosen. In this way, the entire adaptive test adheres to the constraints throughout the administration.

In our case, the constraint concerns the choice of the item database from which all subsequent items for the shadow test at step $m$are selected. If $l^t_{n_m}$ is lower than C_m, the item selection proceeds as usual from the main database. Conversely, if

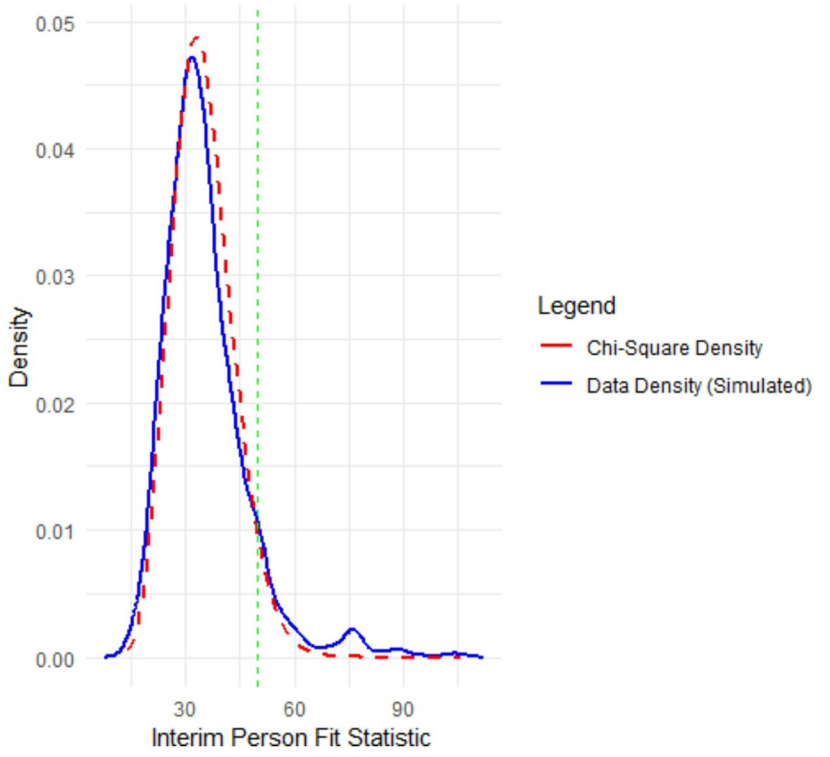

Fig. 2 Simulated $l^{t}_{n_m}$ density distribution compared with the χ^2_{35} density distribution. m = 35. $N = 1000$ of which 5% are cheaters

$l^t_{n_m}$ exceeds C_m, indicating sufficient statistical evidence to flag the test taker as a cheater, then all subsequent items for the shadow test at step $m$are drawn from the more secure database.

Since the statistic $l^t_{n_m}$ is recalculated and updated at each step, shadow test $m+1$ may differ from the previous one, depending on the responses given and the updated value of $l^t_{n_{m+1}}$.

Therefore, it may occur that a test taker who was *not* flagged as a cheater at step $m$becomes flagged at step $m+1$. Symmetrically, a test taker flagged at a certain step may no longer exceed the threshold at a later point.

In this study, we deliberately allow such transitions: the classification into cheater or non-cheater may change dynamically during the test. We are aware that an alternative approach would be to make the flag permanent, that is, once a test taker is identified as a cheater, they would continue to be treated as such for the rest of the test. The comparison between these two operational choices represents an interesting direction for future research.

Finally, the computation of $l^t_{n_m}$ begins only after the administration of an initial set of items.

2.5 M-CHIPS

The Modified-CHIPS (M-CHIPS) is essentially the same as CHIPS; however, in the phase immediately preceding the calculation of the IPS, an additional modification was introduced with the aim of more accurately identifying those cheaters who have complete pre-knowledge of all the items in the original database and therefore will always respond both correctly and very quickly.

This time, after the algorithm administers the initial items (e.g. 5), only the MLE of speed (Eq. 12) is computed. At this point, for those with a $\widehat{\zeta}_{n_m}$ greater than a certain value U, the subsequent 4 items (items 6–9) are selected from the more secret database. Only after the ninth item is answered, the M-CHIPS returns to work as CHIPS. In this way, cheaters will tend to slow down and the IPS will be able to better identify them. The reason why the algorithm does not stop at calculating only $\widehat{\zeta}_{n_m}$ but still proceeds to compute the IPS lies in the fact that comparing $\widehat{\zeta}_{n_m}$ with an arbitrarily chosen threshold could incorrectly identify as cheaters those who are not, but who simply respond very quickly (either because they are skilled or because they are answering randomly). It is then the IPS (through the control of its asymptotic distribution) that is able to separate those who are actual cheaters from those who are merely fast responders.

For this reason, the best expected outcome is that the modification can simultaneously improve the detection of actual cheaters (particularly when they have pre-knowledge of the entire main database) and avoid wrongly classifying as cheaters those who are not.

To verify the properties and limitations of CHIPS and then test whether the modification in M-CHIPS actually achieves the intended goal, several simulations were carried out.

The next section presents the key features of the simulations, while the following one reports the results obtained.

3 Simulation study

The performance and advantages of the CHIPS are investigated through a simulation study. All the simulations were performed with R studio (R Core Team 2013) using the packages: LNIRT (Fox et al. 2021), ShadowCAT (Kroeze 2017) and catIrt (Nydick 2014). These packages were combined and appropriately modified to correctly implement the CHIPS.

We simulated a scenario that could best reflect real-life testing situations where some individuals engage in cheating behaviors. This scenario was used for comparing a CAT using the CHIPS approach to a CAT with the traditional IRT approach. The comparison was based on performance indices, such as the BIAS and the root mean square error (RMSE) for the ability.

Additionally, the statistical test within the CHIPS method was assessed by analyzing errors (Type I and Type II errors) and evaluating the power of the test. Based on

the results of this analysis, several research questions were formulated. To address the questions, one or more characteristics of the simulation were modified.

Finally, the obtained results were discussed also in terms of limitations and potential future developments.

3.1 Simulation design

The steps addressed in the simulation study are described below:

1. The ability and speed of $N = 100$ test takers were simulated from a bivariate normal distribution (Fox and Marianti 2016; Fox et al. 2021; van der Linden 2007) with mean equal to zero and negative correlation (-0.5), to adhere to the speed-accuracy trade- off (van der Linden 2006). The correlation value was derived from an analysis of the 2018 math grade 10 INVALSI data (INVALSI, 2018) using the Hierarchical distribution model (Fox and Marianti 2016; Fox et al. 2021; van der Linden 2007). This approach ensures that the estimated abilities are like those of real students.
2. Out of the 100 students, 20 were simulated as cheaters ($N_C = 20$). Their abilities and speeds were generated from the same distribution as the non-cheaters ($N_H = 80$). However, a cheater is assumed to always respond correctly to items on which they have pre-knowledge, regardless of their true ability and the item difficulty. Moreover, cheaters respond faster than average to items on which they have pre-knowledge

3. The main database was not simulated but consists of 170 items taken from the Credential Form database (available in the R package LNIRT; Fox et al. 2021), whose psychometric characteristics (a_k, b_k, φ_k and λ_k) have been estimated using the R package LNIRT. The more secure database is a perfect copy of the main one.
4. The setups for the CAT were fixed as follows:

   (a) For each test taker, the first 5 starting items were randomly selected.
   (b) As item selection rule, the MIC (Equation B1) was chosen.
   (c) To estimate the interim abilities, the MAP (Equation A2) method was employed. For the estimation of the final ability, the MLE (Eq. 2) method was used.
   (d) As stopping rule, was imposed a fixed-lenght CAT with $K = 35$ items.

   The characteristics (a_k, b_k, φ_k and λ_k) have been estimated using the R package *LNIRT*. The more secure database is a perfect copy of the main one.

5. To each cheater was randomly assigned items from the main database for which they have pre-knowledge. The levels considered were 50%, 75%, and 100% of the items from the main database. We did not consider lower levels (e.g. 25%) because at these levels the BIAS is very low and therefore the contribution of the

method would have been less effective. Pre-knowledge regarding items from the secret database was always assumed to be 0%.

6. For each $i = 1, ..., 100$ replication of the simulation, the answers and the RTs to each item of the test were simulated for all the 100 test takers. The estimation process followed these steps:

   (a) The first 5 items were randomly administered to the test takers. For each of them, a correct response ($Y_k = 1$) or an incorrect response ($Y_k = 0$) was simulated following a two-parameter logistic (2PL) model (Eq. 1).

   For the cheaters, the responses to the items in which they have pre-knowledge were always recorded as correct. A similar procedure was applied to the response times (RTs), which are simulated based on Eq. (4). If a cheater had pre-knowledge of that item, then the response time is divided by 4. The choice of this value was made following preliminary analysis, which highlighted how this value ensures that only the most extreme values in the right tail of the distribution of speed for non-cheaters overlap with the distribution of speed for cheaters.

   (b) Subsequently, the algorithm calculate the IPS (Eq. 13). At this point, the algorithm compares the IPS with the threshold value C_m corresponding to the chosen significance level α (set at 0.05). If it is higher, then the algorithm is constrained (because of the STA) to choose the next item from the more secure database. After the new item is selected, the responses are simulated as described in point (a).
   (c) Step (b) is repeated for $m = 6, \ldots, M$ times until the test reaches the predefined maximum length ($M = 35$). At this point, for each test taker, the final ability is estimated. It is important to emphasize that this entire phase goes beyond the standard settings of a CAT. This represents the additional computational component that is integrated into the classic CAT algorithm. This integration is made possible by framing the operation as a constraint, which is then incorporated into the CAT computational process through the STA.

   Therefore, this phase is not present in the 'classic' IRT-based CAT, which in the results section will serve as a benchmark for evaluating the performance of the proposed methods. However, this is the only substantial difference. Thus, both procedures being compared (CHIPS and classic IRT) share all the same characteristics and processes described at the simulation level, except for this additional step (present in CHIPS and M-CHIPS but absent in classic IRT).

7. Lastly, the average BIAS and RMSE were computed across all 100 replications. The reported results refer to the average values.

The next section presents the results of the simulation. First, the comparison between CHIPS and classic IRT is reported. Subsequently, the results concerning M-CHIPS are also presented.

Table 1 BIAS and RMSE of ability for cheaters and non-cheaters, IRT and CHIPS

BIAS	Noncheaters		Cheaters	
P-K	IRT	CHIPS	IRT	CHIPS
50%	0.030	0.030	0.861	0.207
75%	0.030	0.030	1.869	0.452
100%	0.030	0.030	5.939	5.44
RMSE	Noncheaters		Cheaters	
P-K	IRT	CHIPS	IRT	CHIPS
50%	0.088	0.088	1.088	0.149
75%	0.088	0.088	4.488	0.575
100%	0.088	0.088	35.838	32.421

Table 2 Percentage variation of cheaters' BIAS and RMSE using CHIPS instead of IRT

P-K	$\Delta BIAS_C$	$\Delta RMSE_C$
50%	-76%	-86%
75%	-76%	-87%
100%	-8%	-10%

4 Results

Table 1 presents BIAS and RMSE of the ability estimator for both classic IRT and CHIPS methods for each level of pre-knowledge and distinguished between cheaters and non-cheaters.

When considering only the cheaters, CHIPS performs better than IRT, but both BIAS and RMSE will increase with growing pre-knowledge. Specifically, for a 100% pre-knowledge, the values between the two models are almost identical. On the other hand, when considering only the non-cheaters, these indices are the same for both methods. This findings demonstrate how CHIPS is able to improve the ability estimation for cheaters while it does not affect the one for non-cheaters.

To better assess the impact of using CHIPS on cheaters' ability estimation performance, Table 2 displays the percentage reduction in both BIAS and RMSE, using CHIPS instead of IRT:

$$\Delta BIAS_C = 100\frac{BIAS_{CHIPS_C} - BIAS_{IRT_C}}{BIAS_{IRT_C}},$$

$$\Delta RMSE_C = 100\frac{RMSE_{CHIPS_C} - RMSE_{IRT_C}}{RMSE_{IRT_C}}. \tag{14}$$

Once again, for the first two levels of pre-knowledge, the results are quite promising, displaying an high reduction for both indices. However, the reduction sharply decreases for a 100% pre-knowledge.

Subsequently, a test analysis was conducted. In this context the Type I error rate is the rate of misclassifying non-cheaters as cheaters, while the true negative rate is the rate of correctly classifying non-cheaters. Similarly, the Type II error rate is the rate

of misclassifying cheaters as non-cheaters, with the power of the test representing the rate of correctly identifying cheaters.

Misclassifying a cheater (Type II error) risks overestimating their ability, while misclassifying an non-cheater (Type I error) slightly exposes the database without penalizing the test taker. For that reason, minimizing the Type II error rate is preferable to enhance the power of the test, as long as the Type I error rate remains within acceptable limits. Table 3 present test outcomes, for the three levels of pre-knowledge.

Once again, the results are consistent with what has been discussed so far. In fact, the Type I error rate is very low and very close (even slightly lower) to the significance level α. The Type II error rate remains relatively stable, but only for the first two pre-knowledge levels. In fact, it increases substantially (0.894) for 100% pre-knowledge, mirroring the behavior of BIAS and RMSE for cheaters.

The increase in the Type II error rate can be linked to how the IPS is formulated (Eq. 25). In fact, $l^t_{n_m}$ relies heavily on the difference between expected and actual response times,

rather than solely on the estimated speed. Even individuals with an extreme estimated speed won't have their H_0 rejected if they maintain a consistent speed throughout the test. Thus, in cases of full pre-knowledge, cheaters consistently respond at an extreme speed.

The 10% of accurately classified cheaters are attributed to the random component ϵ_{nk} outlined in Eq. (4).

Since in this simulation the value of the divider for RT (set at 4) was kept constant, the observed limit could be attributed to the simulation itself. Therefore, it is not guaranteed that CHIPS would encounter the same issues in a real scenario. Nonetheless, to overcome the limits identified in this simulation, we decided to test a modification to CHIPS: the Modified-CHIPS (M-CHIPS).

4.1 M-CHIPS results

As already described in the section dedicated to M-CHIPS, where the new method is presented, the only difference between CHIPS and M-CHIPS lies in the fact that

Table 3 Decision table

P-K = 50%	Decision about H0	
	Fail to reject	Reject
True	0.957	0.043
False	0.043	0.957
P-K = 75%	Decision about H0	
	Fail to reject	Reject
True	0.957	0.043
False	0.040	0.960
P-K = 100%	Decision about H0	
	Fail to reject	Reject
True	0.957	0.043
False	0.894	0.107

Pre-knowledge 50%, 75% and 100%

in this second method, before the IPS calculation phase, there is a preliminary step in which 3 items from the most secure database are administered based solely on the MLE of speed ($\widehat{\zeta}_{n_m}$). To implement this 'modification' within the simulation, during the CHIPS phase (i.e., phase 6 described in the Simulation section), after the administration of the initial 5 items, the system does not immediately begin calculating the IPS but instead limits itself to calculating $\widehat{\zeta}_{n_m}$. At this point, for those with a $\widehat{\zeta}_{n_m}$ greater than a certain value U, the subsequent 4 items (items 6–9) are selected from the more secret database. Only after the ninth item is answered, the M-CHIPS returns to work as CHIPS. In this way, cheaters will tend to slow down and the IPS will be able to better identify them.

Regarding the choice of U, choosing a value that is too large would risk not identifying many cheaters. Conversely, selecting a value that is too small would risk administering secret items to many non-cheaters, potentially overexposing the more secret database. To strike a balance, assuming no information about the actual speed distributions of test takers is available, a value of $U = 0.693$ was chosen. This value corresponds to the speed of those who answer questions in half the average time needed.

In order to test the choice of U, Table 4 shows the classification indices at the ninth item, to observe how many cheaters are correctly classified with the modification and how many non-cheaters are mistakenly identified, even before the statistic is calculated.

The results show that the chosen value U allows the items in the more secret database to avoid being overexposed, as the Type I error rate remains below the 0.05 threshold, while correctly identifying the cheaters, in fact the Type II error rate is very close to zero when pre-knowledge is 100%. Considering that, at this stage, only the expected speed and the recorded speed are compared, it is expected that for lower pre-knowledge levels (50%), the percentage of cheaters correctly classified is lower than in the case of maximum pre-knowledge (100%). In fact, in the latter case, a cheater is assumed to always answer correctly and very quickly to all nine items, whereas in the former case, a cheater only increases their response speed on certain

Table 4 Decision table

P-K = 50%	Decision about H0	
	Fail to reject	Reject
True	0.965	0.035
False	0.560	0.440
P-K = 75%	Decision about H0	
	Fail to reject	Reject
True	0.965	0.035
False	0.279	0.721
P-K = 100%	Decision about H0	
	Fail to reject	Reject
True	0.965	0.035
False	0.061	0.939

M-CHIPS. 9-th item. Pre-knowledge 50%, 75% and 100%

Table 5 BIAS and RMSE of ability for cheaters and non-cheaters, IRT, CHIPS and M-CHIPS

BIAS	Noncheaters			Cheaters		
P-K	IRT	CHIPS	M-CHIPS	IRT	CHIPS	M-CHIPS
50%	0.030	0.030	0.031	0.861	0.207	0.197
75%	0.030	0.030	0.031	1.869	0.452	0.249
100%	0.030	0.030	0.031	5.939	5.44	0.653
RMSE	Noncheaters			Cheaters		
P-K	IRT	CHIPS	M-CHIPS	IRT	CHIPS	M-CHIPS
50%	0.088	0.088	0.087	1.088	0.149	0.135
75%	0.088	0.088	0.087	4.488	0.575	0.195
100%	0.088	0.088	0.087	35.838	32.421	1.964

Table 6 Percentage variation of cheaters' BIAS and RMSE using M-CHIPS instead of IRT

P-K	$\Delta BIAS_C$	$\Delta RMSE_C$
50%	77%	88%
75%	87%	96%
100%	89%	95%

questions. However, these are only partial results. To see the true effect of the change, we need to focus on the full test.

Table 5 illustrates how the M-CHIPS not only outperforms the IRT but also effectively overcomes the CHIPS limitation for 100% pre-knowledge. In fact, the M-CHIPS manages to enhance the estimation of cheaters' abilities compared to CHIPS, without negatively impacting non-cheaters' estimation. The BIAS and RMSE values for the latter remain nearly identical to those of the IRT and CHIPS methods, across all three pre-knowledge levels.

Moreover, by comparing Table 6 with Table 2, it can be observed that both the $\Delta BIAS_C$ and $\Delta RMSE_C$ of the M-CHIPS method are not only slightly higher than those of CHIPS for the first two pre-knowledge levels but exhibit a considerable difference when the last level is reached.

By analyzing the BIAS and RMSE for cheaters when they are correctly or incorrectly classified, for 100% pre-knowledge, it is revealed that the BIAS and RMSE for correctly classified cheaters are lower (BIAS = 0.331, RMSE = 0.361) than ones displayed in Table 5. In fact, they are strongly influenced by those for the incorrectly identified cheaters (BIAS = 2.266, RMSE = 8.109), even though these instances are only few.

The results are confirmed by the test analysis. In fact, Table 7 not only shows that the Type I error rate has consistently remained below 0.05 but also demonstrates that, for 100% pre-knowledge, the test power has considerably increased compared to CHIPS, rising from 0.107 (Table 3) to 0.885 (Table 7).

Therefore, M-CHIPS, compared to CHIPS, manages to enhance the identification of cheaters, especially for high pre-knowledge levels, without deteriorating the identification of non-cheaters.

Lastly, to understand if M-CHIPS is sensitive to certain characteristics, additional simulations were conducted with some modifications to the simulation setup. The main results are reported in the Supplementary material.

Table 7 Decision table

P-K = 50%	Decision about H0	
	Reject	Fail to reject
True	0.957	0.043
False	0.039	0.961
P-K = 75%	Decision about H0	
	Reject	Fail to reject
True	0.957	0.043
False	0.030	0.970
P-K = 100%	Decision about H0	
	Reject	Fail to reject
True	0.957	0.043
False	0.115	0.885

M-CHIPS. Pre-knowledge 50%, 75% and 100%

5 Concluding remarks

In this work, the CHeater identification using Interim Person fit Statistic (CHIPS) method for item pre-knowledge on CAT has been proposed.

From a methodological and computational perspective, the main contribution of this work lies in transforming a likelihood-based person-fit statistic for response times, traditionally evaluated only after the full test, into an online statistic that can be recomputed at every step of a computerized adaptive test. By replacing the parameters that normally require MCMC estimation with interim maximum-likelihood estimators of person speed and expected log-response times, we derive an Interim Person-fit Statistic (IPS) whose asymptotic χ^2distribution is analytically tractable at each adaptive step. This reformulation turns a global diagnostic tool into a sequential, computationally lightweight statistic that can be embedded directly into CAT's iterative architecture.

The computational novelty extends to how the IPS is incorporated into the Shadow Test Approach (STA). At each step of the CAT, the algorithm generates a full shadow test that satisfies all content and exposure constraints, included an additional constraint determined by the IPS. When the IPS exceeds the predefined threshold (which depends on the distribution χ^2, on the step $m$that determines its degrees of freedom, and on the chosen significance level), the optimization routine that constructs the shadow test is forced to draw all remaining items from a more secure item bank. In this way, item selection is no longer guided solely by the classical item-selection rule (here, the maximization of Fisher information), but by a joint criterion combining statistical evidence of cheating and psychometric optimality.

The STA therefore becomes a dual-objective constrained optimization problem: keeping information high while respecting a dynamically activated security constraint. This integration represents an explicit computational strategy for steering an adaptive algorithm across two item pools in real time, something that is not available in traditional CAT implementations.

The methodological extension introduced in M-CHIPS further refines this mechanism. By incorporating an early-stage intervention based strictly on estimated speed, M-CHIPS forces cheaters with extreme and consistent pre-knowledge to receive sev-

eral items from the secure database before the IPS is computed. This preliminary perturbation breaks the stability of response-time patterns typical of full pre-knowledge cases, patterns that would otherwise make the IPS less sensitive, thereby improving the statistic's discriminatory power without increasing computational burden. The design remains compatible with the STA: the constraint is activated temporarily based on a deterministic rule involving the speed estimate, and then, after the first nine items have been administered, the algorithm move to the IPS-based constraint. This two-step mechanism shows how real-time statistical monitoring can be combined with adaptive, constraint-based item selection to make the procedure more robust when test takers behave strategically or attempt to cheat.

Taken together, these elements position CHIPS and M-CHIPS as contributions not only to cheating detection, but to the computational statistics of adaptive testing more broadly. They demonstrate how fit statistics with known asymptotic behaviours can be reformulated for sequential use, how they can be embedded as constraints within shadow-test optimization, and how dynamic item-bank switching can be managed algorithmically in a way that preserves psychometric efficiency while increasing test security.

The results of the simulation study confirm the effectiveness of CHIPS in improving the accuracy of ability estimates for cheaters while maintaining the validity of assessments for. non-cheaters. The method demonstrated significant reductions in bias and root mean square error (RMSE) in the ability estimations of cheaters across various levels of item pre-knowledge.

However, for cases where cheaters had complete pre-knowledge of all test items, CHIPS exhibited limitations in its detection capability due to the stability of response patterns. To address this issue, the Modified version, M-CHIPS, was introduced. M-CHIPS enhances detection accuracy by incorporating an early-stage intervention based on response speed, effectively counteracting cases of extreme and consistent cheating behavior. M-CHIPS results in a better performance of ability estimation, especially for 100% item pre-knowledge, without overexposing the more secure database. The simulation study also revealed the flexibility of the methods to various factors, including the test length, the significance level, the correlation between speed and ability, and the ability distribution of cheaters.

Specifically, regarding test length, the method maintained stable estimation performances in fixed-length CAT scenarios, while also proving effective in variable-length CAT when stopping criteria were adjusted. The significance level α played a crucial role in determining the trade-off between Type I and Type II errors, with higher α values increasing the power of the test to detect cheaters while slightly increasing the exposure of the more secure item database. The correlation between speed and ability influenced the accuracy of cheating detection: the method demonstrated higher robustness when ability and speed were negatively correlated, as cheaters with lower ability but higher speed were more effectively identified. Even when the correlation is positive, the method appears to have very good performances. Furthermore, when the ability distribution of cheaters was lower than that of non-cheaters, the methods continued to function reliably, with M-CHIPS slightly improving its detection rate due to the stronger discrepancy between expected and observed response behaviors. These findings confirm that the proposed methods can be effectively adapted to vari-

ous testing conditions while maintaining a strong capacity to detect and mitigate cheating in CAT environments.

Despite these positive results, the method is not without limitations, which are more related to general considerations rather than to specific evidence found in the simulation. Specifically, the method is conceived for binary items and when the assumption of log normality for the RTs is fulfilled.

The simulations were based on specific assumptions, such as cheaters having no pre-knowledge of items in the more secure database and no knowledge of how the identification method works. Otherwise, cheaters might intentionally slow down to try to deceive the method. Additionally, setting a fixed value to divide the response time of cheaters is a limitation of the simulation, which is due to the fact that in the literature there are no studies that actually analyze the response speed of a cheater when they have pre-knowledge versus when they do not. Instead, there are only studies that use data involving cases of cheating, which vary from study to study. A possible development could involve generating different values from a random variable, or hypothesizing the existence of a secondary speed, such as a cheating speed (similar to a question-reading speed).

Moreover, the two methods presented in this work rely solely on RT. As with the person-fit statistics that the IPS draws upon (Fox and Marianti 2016), it is possible to modify the statistic to account jointly for both response time and response accuracy (RA). Such an addition would provide greater stability to the IPS, making it more robust against type II errors, particularly when non-cheaters resort to fast-guessing. However, this modification would require a complete extension of the IPS, with the introduction of a counterpart incorporating RA, which would need to be updated step by step and whose asymptotic distribution would have to be studied. For this reason, we consider it a future step that requires significant research to further improve the IPS and the method for detecting cheaters.

In the context of the simulation, it would be interesting to explore what happens to CHIPS and M-CHIPS when the RTs of test takers do not follow a log-normal distribution. Furthermore, it would be methodologically of interest to study whether it is possible to modify the fit statistic to make it more general and adaptable to different RT models.

Additionally, investigating scenarios where cheaters have some level of pre-knowledge about items in the more secure database or intentionally slow down to deceive the method, could provide valuable insights. Furthermore, the simulations did not make assumptions about unusual behaviors of non-cheaters, such as fast-guessing behaviors.

Beyond the simulation, other developments could involve the implementation of methods capable of generating the more secure items on-the-fly, starting from items in the main database (*item cloning*). This would drastically decrease the likelihood that cheaters have pre-knowledge of those items and reduce the burden of constantly updating the more secure database.

Lastly, a crucial future development would be to implement the method in an operational CAT.

Appendix A

IRT ability estimation methods

The MLE method is based on maximizing the likelihood function (Equation A1),

$$\widehat{\theta}_{n_{MLE}} = arg\max_{\theta_n} \left\{ L\left(\theta|\overrightarrow{Y}_n\right) : \theta_n \in (-\infty, \infty) \right\},$$

which in the case of the 2PL model has the following form:

$$L\left(\theta|\overrightarrow{Y}_n\right) = \prod_{k=1}^{K} \left\{ \frac{\exp[a_k(\theta_n - b_k)]}{1+\exp[a_k(\theta_n - b_k)]} \right\}^{Y_{nk}} \left\{ 1 - \frac{\exp[a_k(\theta_n - b_k)]}{1+\exp[a_k(\theta_n - b_k)]} \right\}^{1-Y_{nk}}. \tag{A1}$$

Since the first derivative of the likelihood function (Equation A1) does not have a closed-form solution, $\widehat{\theta}_{n_{MLE}}$ is generally estimated using nonlinear minimization methods, for example employing a Newton-type algorithm (Dennis Jr and Schnabel 1996; Schnabel et al. 1985). It is an iterative procedure that aims to minimize a given function $f(x)$. In this case, the function to be minimized is the negative log-likelihood:

$$f(x) = -\ln\left(\theta|\overrightarrow{Y}_n\right).$$

The procedure begins with an initial guess for the solution, $\theta_n 0$. This value can be chosen randomly or based on prior knowledge. After that, at each iteration $m = 0, ..., M$ the gradient vector $f'(\theta_{n_m})$ and the Hessian matrix $f''(\theta_{n_m})$ are computed with respect to the parameter θ_n evaluated at the current estimate θ_{n_m}. Then, the parameter estimate is updated using the Newton-Raphson formula:

$$\theta_{n_{m+1}} = \theta_{n_m} - \frac{f'(\theta_{n_m})}{f''(\theta_{n_m})}.$$

The process is repeated until the termination criterion is met, for example reaching a maximum number of iterations (M) or achieving a desired level of accuracy.

Regarding alternative estimation methods to ML, among the most commonly used in the field of IRT models are Bayesian methods. These estimation methods are based on Bayes' theorem.

In these methods, θ_n is treated as a random variable, and the goal is to find its *posterior distribution* $g\left(\theta|\overrightarrow{Y}_n\right)$ starting from a *prior distribution* $g(\theta_n)$ which is hypothesized based on known characteristics of θ:

$$g\left(\theta|\overrightarrow{Y}_n\right) = \frac{L\left(\theta|\overrightarrow{Y}_n\right) g(\theta_n)}{\int L\left(\theta|\overrightarrow{Y}_n\right) g(\theta_n) d\theta_n},$$

where $L\left(\theta|\overrightarrow{Y}_n\right) g(\theta_n)\, d\theta_n$ is the marginal likelihood, representing the overall probability of observing the data under the model. Just like the likelihood, the posterior distribution in Equation (A2) can be maximized to define the estimate of θ_n. In this case, the estimator is called the Maximum A Posteriori (MAP) estimator (Lord 1986; Mislevy 1986).

$$\widehat{\theta}_{n_{MAP}} = \underset{\theta_n}{arg\max} \left\{ g\left(\theta|\overrightarrow{Y}_n\right) : \theta_n \in (-\infty, \infty) \right\}, \tag{A2}$$

The small-sample properties of the MAP estimator depend on the likelihood and also on the shape of the prior distribution. In fact, for uniform prior, the posterior distribution in Equation (A2) becomes proportional to the likelihood function over the support of the prior, and the maximizers in Eqs. (A1) and (A3) are equal.

Regarding the estimation procedure, as with ML estimators, since there is no closed-form solution, the same nonlinear minimization method previously described can be employed, as well as the Expectation-Maximization (EM) methods (McLachlan and Krishnan 2007). For this method as well, an initial value θ_{n_0} is chosen.

Subsequently, for each step $m = 0, ..., M$, there is first an Expectation step (E-step), in which the expected value is computed of individuals' responses to each item based on the current estimates of θ_{n_m} and the item parameters $I_K = (a_k, b_k, c_k)$. These expected values represent the probability of a correct response for each item.

$$E[Y_{nk}] = P(Y_{nk} = 1|\theta_n \mathrm{m}, I_K),$$

$$E\left[\overrightarrow{Y_n}\right] = (E[Y_{n1}], \ldots, E[Y_{nk}]). \tag{A3}$$

Depending on the IRT model selected, $E\left[\overrightarrow{Y_n}\right]$ is used to calculate the corresponding expected complete-data log-likelihood:

$$Q(\theta_n|\theta_n \mathrm{m}) = E\left[\ln L\left(\theta_n \mathrm{m}|E\left[\overrightarrow{Y_n}\right]\right)\right] + \ln g(\theta_n).$$

This phase is followed by a Maximization step (M-step), where the expected complete-data log-likelihood is maximized to find the new estimate of θ_n.

$$\widehat{\theta}_{n_{m+1}} = \underset{\theta_n}{arg\max} \{Q(\theta_n|\theta_n \mathrm{m})\}.$$

The E-step and the M-step are iteratively performed until the termination criterion is met.

Appendix B

Further analyses of M-CHIPS performance

The adaptive testing process involves several key steps:

- Ability Initialization: The adaptive process begins with the algorithm estimating the test-taker presumed ability. Common choices for the initial ability (θ_{0_n}) include setting it to 0, choosing a random value (*random initialization*), or using available information from prior tests. In this study, ability is initialized at zero.
- Selection of Initial Items: Some initial items are selected to collect essential information for the subsequent ability estimation process. Typically, 3–5 items are chosen, either based on the presumed ability or selected randomly. In the simulation reported later, the first five items are selected completely at random.
- Estimation of the First Ability: The first ability is estimated after the administration of initial items. Challenges arise in cases where all items are answered correctly or incorrectly. Various approaches, such as fixing the estimate temporarily or using Bayesian estimation methods, are employed to address these issues. Precisely to avoid such issues, the chosen estimation method is MAP.
- Estimation of Interim Abilities θ_{m_n} : Interim abilities are estimated, at each step $m = 1, ..., M$of the CAT, after each answer to a new item, forming the core of the iterative process. Both maximum likelihood (ML) and Bayesian estimation methods can be used. The method used to estimate abilities after the initial one does not necessarily have to be the same as the one employed in the previous step. This alternation provides great flexibility to the CAT. However, since the simulation assumes the presence of cheaters, who may answer all posed items correctly at a given point, and given that ML models present certain limitations in such cases (van der Linden and Glas 2010), the MAP method was chosen for estimating intermediate abilities as well.
- Item Selection Criterion (ISC): It is the rule used to select the next item in a test, with the goal of matching item difficulty to the examinee's estimated ability for greater precision. In the 2PL model, item information is highest when ability θ is close to item difficulty b_k, meaning well-targeted items minimize estimation error.
- For this reason, item selection relies on the updated ability estimate at each step. Common criteria include maximum information and Bayesian approaches, depending on the context. In this study, the Maximum Information Criterion (MIC) is used, selecting each item based on the current estimate θ_n.

When the ML is chosen as estimator for θ_n, under smoothness conditions on the IRF, $\widehat{\theta}_{n_{MLE}}$ is asymptotically distributed as $N\left(\theta_n^*, I^{-1}\left(\theta_n^*\right)\right)$ where θ_n^* is the true value of the latent ability, while $I\left(\theta_n^*\right)$ is the Fisher information related to θ_n^*. Therefore, the inverse of $I\left(\theta_n^*\right)$ is the asymptotic variance of $\widehat{\theta}_{n_{MLE}}$. So, the larger is the Fisher information, the smaller is the asymptotic variance of $\widehat{\theta}_{n_{MLE}}$. $I\left(\theta_n^*\right)$ is also the expected value of the second derivative of $\ln L\left(\theta_n\right)$, with inverted sign.

$$I\left(\theta_n^*\right) = -E\left[\frac{\partial^2 \ln L(\theta_n)}{\partial^2 \theta_n^2}\right].$$

In addition, this expected value is equivalent to the sum of the individual Item Information Functions (IIF), indicated with $I_k\left(\theta_n\right)$, that indicates how much information can be derived from each single item, with reference to a generic θ_n.

$$\left(\theta_n^*\right) = \sum_{k=1}^{K} I_k\left(\theta_n\right),$$

$$I_k\left(\theta_n\right) = \frac{[P(Y_{nk}=1|\theta_n)]^2}{P(Y_{nk}=1|\theta_n)(1-P(Y_{nk}=1|\theta_n))}.$$

The MIC involves choosing, at each selection step $m = 1, ..., M$, the item that maximizes $I\left(\theta_n^*\right)$.

$$k_{m+1} = \underset{l}{argmax}\left\{I_l\left(\widehat{\theta}_{m_n}\right) : l \in R_m\right\}, \tag{B1}$$

where:

- $m = 1, ..., M$ indicates the number of items that have already been administered.
- R_m is the set of all items that have not yet been administered.
- l indicates the generic element of the set R_m.
- $\widehat{\theta}_{m_n}$ is the generic estimate of θ_n at step m. As this is a generic formulation, $\widehat{\theta}_{m_n}$ can be obtained using any estimation method, not necessarily ML.
- $I_l\left(\widehat{\theta}_{m_n}\right)$ indicates the information provided by item l at the ability estimate at step m.

Generally, the larger $I_l\left(\widehat{\theta}_{m_n}\right)$ is, the smaller is the asymptotic variance of $\widehat{\theta}_{m_n}$. Specifically, when using an ML estimator, $\widehat{\theta}_{n_{MLE}}$ will have the smallest possible asymptotic variance. For this reason, when $\widehat{\theta}_{n_{MLE}}$ is chosen as the estimator for interim abilities, it is common to select the MIC.

However, there are no methodological constraints, and it is also quite popular to use the MAP in combination with the MIC (van der Linden and Glas 2010).

- Stopping Rule: The iterative process is terminated based on a chosen stopping rule, commonly either after a fixed number of items (*fixed length test*) or when interim ability estimation stabilizes (*variable length test*). In this study, we present both a simulation with a fixed-length test (35 items) and one with variable length.
- Estimation of the Final Ability θ_n: Final ability estimates are computed to ensure optimal statistical properties and are often converted into an equated number-correct score using methods such as the test characteristic function or equipercentile transformation. The choice of estimation method depends on several CAT com-

ponents, including item selection, item pool, and constraints. In this study, which uses dichotomous multiple-choice items, the final ability θ_n is not transformed.

Moreover, the method for the final estimate may differ from earlier ones: here, MAP is used for initial and interim estimates, while the final ability is estimated using MLE.

Finally, regarding the ISC, there are also approaches that integrate the item selection phase. One of these is the Shadow Test Approach (STA)(van der Linden and Reese 1998).

STA is often combined with information-based item selection methods because it compensates for their lack of control over item exposure and helps enforce test specifications, such as content constraints. A shadow test is built using algorithms similar to those for linear tests, but it remains latent and is regenerated at each step of the CAT. For a fixed-length test *M*, each shadow test also has length *M*.

Each shadow test satisfies all constraints and includes previously administered items. Thus, the first shadow test resembles a newly assembled linear test, while the last corresponds to the final adaptive test and fully respects all constraints (van der Linden and Glas 2010). This feature is essential when constraints are integrated into the process. In the proposed method, a key step is adding a constraint that guides item selection from two databases. STA is therefore required to ensure this constraint is maintained throughout the test, as will be explained in detail later.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00180-026-01739-1.

Funding Open access funding provided by Alma Mater Studiorum - Università di Bologna within the CRUI-CARE Agreement. The publication was partially produced with funding from the Italian Ministry of University and Research under the Call for Proposals related to the scrolling of the final rankings of the PRIN 2022 call. Latent variable models and dimensionality reduction methods for complex data (PI Prof. Paolo Giordani) Project No. 20224CRB9E, CUP C53C24000730006 Grant Assignment Decree No. 1401 adopted on 18.9.2024 by MUR.

References

Binet A, Simon T (1905) Méthodes nouvelles pour le diagnostique du niveau intellectuel des anormaux. L' Annee Psychologique 2:245

Cizek GJ, Wollack JA (2016) Handbook of quantitative methods for detecting cheating on tests. Taylor & Francis Group, New York

Dennis Jr JE, Schnabel RB (1996) Numerical methods for unconstrained optimization and nonlinear equations. SIAM, Philadelphia

Fan Z, Wang C, Chang H-H, Douglas J (2012) Utilizing response time distributions for item selection in CAT. J Educational Behav Stat 37(5):655–670. https://doi.org/10.3102/1076998611422912

Fox J-P, Marianti S (2016) Joint modeling of ability and differential speed using responses and response times. Multivar Behav Res 51(4):540–553. https://doi.org/10.1080/00273171.2016.1171128

Fox J-P, Klotzke K, Simsek AS (2021) LNIRT: An R package for joint modeling of response accuracy and times. *arXiv preprint arXiv:2106.10144*. https://doi.org/10.48550/arXiv.2106.10144

Fox J-P, Marianti S (2017) Person-fit statistics for joint models for accuracy and speed. J Educ Meas 54(2):243–262. https://doi.org/10.1111/jedm.12143

Iannario M, D'Enza AI, Romano R (2024) A hybrid approach for the analysis of complex categorical data structures assessment of latent distance learning perception in higher education. Comput Stat 39(1):161–179. https://doi.org/10.1007/s00180-022-01272-x

INVALSI (2018) Rapporto prove invalsi 2018. INVALSI. Rome, Italy. https://www.invalsi.it/invalsi/doc_evidenza/2018/Rapporto_prove_INVALSI_2018.pdf

Kroeze K (2017) Multidimensional computer adaptive testing with the shadow testing routine. https://github.com/Karel-Kroeze/ShadowCAT.git

Lord FM (1986) Maximum likelihood and Bayesian parameter estimation in item response theory. J Educ Meas 23:157–162

Lord FM, Novick MR (1968) Statistical Theories of Mental Test Scores. Addison-Wesley, Reading

Marianti S, Fox J-P, Avetisyan M, Veldkamp BP, Tijmstra J (2014) Testing for aberrant behavior in response time modeling. J Educational Behav Stat 39(6):426–451. https://doi.org/10.3102/1076998614559412

Martinková P, Hladká A (2023) Computational aspects of psychometric methods: With r. Chapman. https://doi.org/10.1201/9781003054313. Hall/CRC

McLachlan GJ, Krishnan T (2007) The em Algorithm and Extensions. Wiley, New Jersey

Mislevy RJ (1986) Bayes modal estimation in item response models. Psychometrika 51:177–195. https://doi.org/10.1002/j.2330-8516.1985.tb00118.x

Nydick S (2014) *Simulate IRT-based computerized adaptive tests*. https://github.com/swnydick/catIrt

R Core Team (2013) R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/

Ranger J, Schmidt N, Wolgast A (2023) Detecting cheating in large-scale assessment: The transfer of detectors to new tests. Educ Psychol Meas 83(5):1033–1058. https://doi.org/10.1177/00131644221132723

Schnabel RB, Koonatz JE, Weiss BE (1985) A modular system of algorithms for unconstrained minimization. ACM Trans Math Softw (TOMS) 11(4):419–440

Sinharay S (2017) Detection of item preknowledge using likelihood ratio test and score test. J Educational Behav Stat 42(1):46–68. https://doi.org/10.3102/1076998616673872

Sinharay S (2018) A new person-fit statistic for the lognormal model for response times. J Educ Meas 55(4):457–476

Stocking ML, Lewis C (1998) Controlling item exposure conditional on ability in computerized adaptive testing. J Educational Behav Stat 23(1):57–75. https://doi.org/10.3102/10769986023001057

Sympson J, Hetter R (1985) Controlling item-exposure rates in computerized adaptive testing. *Proceedings of the 27th Annual Meeting of the Military Testing Association*, 973–977

van der Linden WJ (2006) A lognormal model for response times on test items. J Educational Behav Stat 31(2):181–204. https://doi.org/10.3102/10769986031002181

van der Linden WJ (2007) A hierarchical framework for modeling speed and accuracy on test items. Psychometrika 72(3):287. https://doi.org/10.1007/s11336-006-1478-z

van der Linden WJ, Glas CAW (2010) Elements of adaptive testing, vol 10. Springer, New York. https://doi.org/10.1007/978-0-387-85461-8

van der Linden WJ, Reese LM (1998) A model for optimal constrained adaptive testing. Appl Psychol Meas 22(3):259–270

van Krimpen-Stoop EM, Meijer RR (2000) Detecting person misfit in adaptive testing using statistical process control techniques. In *Computerized adaptive testing: Theory and practice* (pp. 201–219). Springer. https://doi.org/10.1007/0-306-47531-6_11

Veldkamp BP (2016) On the issue of item selection in computerized adaptive testing with response times. J Educ Meas 53(2):212–228. https://doi.org/10.1111/jedm.12110

Wainer H, Dorans NJ, Flaugher R, Green BF, Mislevy RJ (2000) Computerized adaptive testing: A primer. Routledge. https://doi.org/10.4324/9781410605931

Wollack JA, Maynes DD (2016) Detection of test collusion using cluster analysis. Handbook of quantitative methods for detecting cheating on tests. Routledge, pp 124–150

Zhou T, Jiao H (2023) Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. Educ Psychol Meas 83(4):831–854. https://doi.org/10.1177/00131644221117193

Authors and Affiliations

Luca Bungaro[1,3] · Mariagiulia Matteucci[1] · Stefania Mignani[1] · Bernard P. Veldkamp[2]

✉ Luca Bungaro
luca.bungaro2@unibo.it

[1] University of Bologna, Bologna, Italy

[2] University of Twente, Enschede, The Netherlands

[3] INVALSI, Rome, Italy