Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN
SCIENZE STATISTICHE

Ciclo XXXVI

Settore concorsuale: 13/D1 - STATISTICA
Settore Scientifico Disciplinare: SECS-S/01 - STATISTICA

Response times in computerized adaptive testing: A method for cheating detection

---

Presentata da: Luca Bungaro

Coordinatore Dottorato
Prof.ssa Monica Chiogna

Supervisore
Prof.ssa Mariagiulia Matteucci
Co-supervisori
Prof.ssa Stefania Mignani
Prof. Bernard P. Veldkamp

Esame finale anno 2024

Abstract

In the field of educational and psychological measurement, the shift from paper-based to computerized tests has become a prominent trend in recent years. Computerized tests allow for more complex and personalized test administration procedures, like Computerized Adaptive Testing (CAT).

CAT, following the Item Response Theory (IRT) models, dynamically generates tests based on test-taker responses, driven by complex statistical algorithms. Even if CAT structures are complex, they are flexible and convenient, but concerns about test security should be addressed. Frequent item administration can lead to item exposure and cheating, necessitating preventive and diagnostic measures.

In this thesis a method called "CHeater identification using Interim Person fit Statistic" (CHIPS) is developed, designed to identify and limit cheaters in real-time during test administration. CHIPS utilizes response times (RTs) to calculate an Interim Person fit Statistic (IPS), allowing for on-the-fly intervention using a more secret item bank. Also, a slight modification is proposed to overcome situations with constant speed, called Modified-CHIPS (M-CHIPS).

A simulation study assesses CHIPS, highlighting its effectiveness in identifying and controlling cheaters. However, it reveals limitations when cheaters possess all correct answers. The M-CHIPS overcame this limitation. Furthermore, the method has shown not to be influenced by the cheaters' ability distribution or the level of correlation between ability and speed of test-takers.

Finally, the method has demonstrated flexibility for the choice of significance level and the transition from fixed-length tests to variable-length ones.

The thesis discusses potential applications, including the suitability of the method for multiple-choice tests, assumptions about RT distribution and level of item pre-knowledge. Also limitations are discussed to explore future developments such as different RT distributions, unusual honest respondent behaviors, and field testing in real-world scenarios. In summary, CHIPS and M-CHIPS offer real-time cheating detection in CAT, enhancing test security and ability estimation while not penalizing test respondents.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Overview

In recent years, computerized tests for educational and psychological measurement, have been replacing paper-based ones, a trend set to continue, likely making computer-based assessments the norm. Statistical models for these tests, particularly Item Response Theory (IRT) dating back to the 1950s, offer a strong foundation for assessing and comparing individual skills (Lazarsfeld, 1949, Lord, 1952, Rasch, 1960).

These models remain relevant despite complex developments in test methodologies, such as Computerized Adaptive Testing (CAT; van der Linden and Glas, 2010a). CATs dynamically generate the test based on the test-taker's responses, driven by a complex statistical algorithm.

CAT structures are intricate, with various components. Some methodologies aim to enhance CAT by incorporating additional data sources, like response times (RT). Several models have been proposed to model RT distributions (van der Linden, 2009, De Boeck and Jeon, 2019), useful not only for improving CAT but also for detecting those who implement incorrect behavior in order to answer correctly to items, namely cheaters.

Continuous testing, while flexible and convenient, raises security concerns. Frequently administered items may become compromised, undermining test

integrity. To counter this, researchers have developed preventive measures like item exposure control methods (Sympson and Hetter, 1985). However, even successful controls can't fully prevent cheating, necessitating diagnostic measures to identify anomalous behaviors. These involve detecting aberrant response patterns or RTs across administered items (Marianti et al., 2014, Fox and Marianti, 2017). While much literature covers methods for cheating detection, the majority of the proposed one are only viable post-testing, making interventions less effective and sometimes impractical.

1.2 Main contributions of the thesis

This work starts with a deep analysis of the main IRT models, RT distributions, and CAT. Then, it proceeds to the combined use of all these tools to develop a method capable of identifying and limiting cheaters during the test administration.

RTs are used to define a statistic capable of identifying a suspected cheater. The peculiarity of this statistic is that it is calculated while the test is being taken, allowing for real-time intervention. The proposed method is called "CHeater identification using Interim Person fit Statistic" (CHIPS). It acts by administering to suspected cheaters items taken from a different and more secure item bank, which is less exposed. The items in this bank are changed more frequently and share the psychometric characteristics with those in the main bank.

CHIPS has been tested in a simulation study. It has been highlighted how CHIPS effectively improves the estimates of the true abilities of cheaters without influencing those of non-cheaters (honest respondents). However, it seems to have a limitation when cheaters possess all the correct answers to the test. This limitation is overcome following a slight modification of the method. The modified method (M-CHIPS) has been tested, and through multiple

simulations, some of its potentialities have been highlighted. These include the possibility to be used for variable-length tests and to be flexible regarding the choice of precision in correctly identifying cheaters at the expense of exposing items from the more secure item bank. It is also independent from the distribution of cheaters' abilities and the correlation between ability and response speed.

Finally, some limitations of the method and the analysis are discussed, such as its applicability to multiple-choice tests only and the assumption of a complete absence of item pre-knowledge in the more secure item bank. The study lends itself to future developments to overcome these limitations, ideally through application in a non-simulated context.

Chapter 2

Literature review

2.1 Item Response Theory (IRT)

Item Response Theory (IRT) is a theory aimed at studying and developing a structure capable of measuring and comparing the level of ability of different subjects through tests or questionnaires. It originated around 1950 as a result of independent research conducted in parallel by the Australian sociologist Paul Lazarsfeld (1949), the Danish mathematician George Rasch (1960), and the American psychometrician Frederic M. Lord (1952). Due to the limited power of computers at the time, it was only since the 1980s that this theory has found increasingly concrete applications, thanks also to the work of two psychometricians: the American Benjamin Drake Wright (1981) and the Australian David Andrich (1978).

IRT was born in the field of psychometrics, which is the branch of psychology that deals with techniques and theories for the objective measurement of abilities, knowledge, attitudes, personality traits, and academic outcomes. IRT is generally considered as the continuation and evolution of Classical Test Theory (CTT) (Lord and Novick, 1968), also a psychometric theory of the early 1900s based on the assumption that a person's score (X_n) in a test is nothing but the sum of a *true score* (T_n) and an *error score* (ε_n).

$$X_n = T_n + \varepsilon_n, \tag{2.1}$$

where ($n \in 1, \ldots, N$) refers to the nth subject.

The advancement brought by IRT, in the field of educational measurement, was to hypothesize the existence of latent constructs, such as abilities, which directly underlie and influence the probability of a correct or incorrect answer to the questions of a test. In fact, educational measurement involves utilizing educational assessments and analyzing data, such as scores derived from these assessments, to make inferences about the abilities and proficiencies of students. IRT models are indeed defined as *latent trait models* for continuous latent traits (abilities) and categorical observed variables (item responses), precisely because they assume the existence of traits that are not directly observable and determine observable manifestations, such as answers to questions. In essence, IRT expresses the probability of responding correctly to each *item* of the test, conditioned on a given level of ability (the latent trait). Such models do not only estimate the level of ability of the examined subjects, but also the item parameters, like the difficulty of the item, the discriminant power of each item and sometimes also the probability of randomly responding correctly to a question (*guessing*), considering both the respondents and the items. This is therefore the most significant innovation compared to CTT, where the probability of correct response is modeled only from the total score obtained in the test, resulting in sample-dependent estimates and making it difficult to compare different tests.

In fact, one of the characteristic of the IRT models is the *parameter invariance*. That means that the characteristics of test items, like difficulty and discrimination, remain constant when the test is applied to different groups or populations. There are two key types of invariance:

- Scalar Invariance. The measurement properties of items are equivalent across different groups, allowing for fair group comparisons.
- Strict Invariance. A more stringent requirement that includes scalar invariance and demands that item response variability is also equal across groups.

IRT is thus based on models for which the probability of correct response is conditioned both on the characteristics of the items (*difficulty*, *item discrimination power*, and *guessing*) and on the characteristics of the responding subjects (abilities). We will discuss these characteristics, called *parameters*, and their meaning more thoroughly shortly.

2.1.1 Types of IRT models

Several IRT models can be specified, and these models vary depending on internal model specifications.

- **The type of input data can be either dichotomous or polytomous.** Observed variables are usually the responses given to a questionnaire or a test. If a subject is asked to indicate a preference on a scale, such as a Likert scale, the input variables would be polytomous, as they could vary within a range that allows for more than two responses. The polytomous variable can then be *ordinal* (if the answers have an internal order) or *nominal* (if the answers all have the same importance). In the case of the Likert scale, the variable is of the ordinal type.

  The second type of variable is the *dichotomous*. In this case there are only two possible answers, generally coded with 0 and 1, which in educational assessment usually represent incorrect and correct responses, respectively. Sometimes, as for example in multiple choice tests, even if there are more than two possible answers, it is possible to consider the

response variable as dichotomous, indicating with 1 the correct answer and with 0 all the wrong one.

In this thesis, only datasets containing dichotomous response variables will be used, and models that work with such data will be shown and discussed.

- **The item response function (IRF).**
  The IRF is a mathematical function that expresses the probability of correctly answering an item depending on item parameters and person ability.
  Depending on the chosen mathematical function, different models can be defined. The two most common choices are the *logistic model*, that is based on a logistic function, and the *normal ogive model* (or *probit model*) that is based on the normal probability distribution (Hambleton and Swaminathan, 1985).

- **The number of latent traits.**
  If the test is assumed to measure only one latent trait (meaning that only one type of ability is necessary to model the answering process correctly), then the considered model is defined as *unidimensional*. However, if there are multiple abilities required to model the answering process correctly, then the model is considered *multidimensional* (see, e.g. Johnson et al., 2006, Reckase et al., 2009, Toland et al., 2017, Mair and Gruber, 2022).

- **The number of estimated parameters.** As previously mentioned, IRT models rely on parameters of both the items and the responding subjects. The number of parameters varies depending on the model selected.
  As for the subjects, in the unidimensional case, there is only one type of parameter, called incidental (or *person parameter*), and it refers to

the ability, indicated by the Greek letter Theta (θ). It has a general form of θ_n with ($n \in 1, 2, \ldots, N$), where n refers to the nth subject. The most common assumption is that ability has a normal distribution with mean zero. This means that those with a value of θ_n close to zero have average ability, those with a high positive value are considered very skilled in that particular area, and those with a high negative value have ability well below average. This parameter has a positive (nonlinear) relationship with the probability of responding correctly to a test item. This means that the higher the value of θ_n, the higher the probability that the subject will respond correctly to a test question. This nonlinear relationship is explained by the IRF.

Parameters that refer to items, on the other hand, are defined as structural and come in three types: item discrimination (a), item difficulty (b), and the so-called item guessing parameter (c). As parameters that refer to individual test items, items in the database are denoted by $k \in 1, \ldots, K$.

- a_k: This parameter is called item *discrimination* and indicates how much the item is able to distinguish between a subject with high ability and one with low ability. Geometrically, it indicates the slope of the curve associated with that item. In the common logistic parametrization, an item with a value of this parameter equal to 1 moderately discriminates the ability to which it refers. A value between 0 and 1 will indicate an item that can hardly distinguish between those with high ability and those with low ability, while a value greater than 1 will be associated with an item that easily distinguishes between different levels of ability. This parameter is free to vary between zero and infinity.
- b_k: This parameter indicates the *difficulty* of the item. Geometrically, it represents the location of an item with respect to the ability scale. It corresponds to the ability level at which we would expect

examinees to have a probability of 0.5 (assuming no guessing) of answering the item correctly. Usually θ and b are on the same scale. This means that when the mean value of θ is set to 0, $b = 0$ indicates an item of average difficulty, higher values indicate more difficult items, and values lower than zero indicate easier items. It has a normal distribution.

- c_k: This parameter indicates the *guessing* of the item, which is the probability of answering the item correctly for subjects with very low ability. It may take values from 0 to 1, where 0 indicates a question that cannot be answered correctly by chance, and 1 corresponds to a question that can always be answered correctly by chance. Geometrically, it corresponds to the lower asymptote of the curve.

Depending on the number of item parameters and on the choice of the IRF, different unidimensional IRT models are defined.

2.1.2 Unidimensional IRT

The first IRT models were unidimensional (Rasch, 1960, Lord, 1980). It is assumed, therefore, that the results in a test are determined by a single ability. This is justified by the hypothesis that, even if there could be multiple latent traits, only one of these would be the dominant one that alone would suffice to explain the given responses. For example, to correctly answer a math test item, one can hypothesize that the only ability that would contribute to this purpose would be the one related to math (without considering complementary abilities such as geometric or text comprehension). Obviously, this is a limiting assumption, but in the course of this study, only cases (real or simulated) in which this assumption of unidimensionality is respected will be analyzed. Another characteristic of IRT models is the so-called *local independence*

assumption, meaning it is assumed that the probability that a subject has to answer a single test item correctly, conditionally on their ability, is not influenced by the other item responses. In different terms, the probability of correctly answering the entire test correctly is equal to the product of the probabilities of correctly answering each item. Therefore, given the n-th subject's K-dimensional answers vector:

$$\overrightarrow{\mathbf{Y}_n} \;=\; (Y_{n1},\; \ldots\, Y_{nk},\; \ldots\, Y_{nK}), \tag{2.2}$$

it holds the equation:

$$P(\overrightarrow{\mathbf{Y}_n} = \mathbf{1} \mid \theta_n) \;=\; \prod_{k=1}^{K} P(Y_{nk} = 1 \mid \theta_n). \tag{2.3}$$

As mentioned earlier, this probability depends on the mathematical formulation of the IRF (logistic or normal ogive) and the number of item parameters (1, 2, or 3). For simplicity, we will start from the three-parameter model for both types of IRF, as the other two are specific cases of it.

For the 3-parameter logistic (3PL) (Birnbaum, 1968) model, the probability of a correct answer is:

$$P(Y_{nk} = 1 \mid \theta_n) \;=\; c_k + (1 - c_k)\frac{\exp[a_k(\theta_n - b_k)]}{1 + \exp[a_k(\theta_n - b_k)]}. \tag{2.4}$$

From this more general model, we can derive the 2-parameter logistic (2PL) model and the 1-parameter logistic (1PL or Rasch) model by setting $c_k = 0$ and then $a_k = 1$.

2PL model (Lord, 1952):

$$P(Y_{nk} = 1 \mid \theta_n) = \frac{\exp[a_k(\theta_n - b_k)]}{1 + \exp[a_k(\theta_n - b_k)]}. \tag{2.5}$$

1PL model (Rasch, 1960):

$$P(Y_{nk} = 1 \mid \theta_n) = \frac{\exp[(\theta_n - b_k)]}{1 + \exp[(\theta_n - b_k)]}. \tag{2.6}$$

Regarding the 3-parameter normal ogive (3PNO) (Lord, 1952, Lord and Novick, 1968) model , the IRF is defined as follows

$$P(Y_{nk} = 1 \mid \theta_n) = c_k + (1 - c_k)\Phi[a_k(\theta_n - b_k)], \tag{2.7}$$

where Φ is the normal cumulative distribution function (CDF).

Similarly to logistic models, one can switch from the 3-parameter model to the other two models (1PNOM and 2PNOM) in the same way.

2 PNO model (Lord, 1952, Lord and Novick, 1968):

$$P(Y_{nk} = 1 \mid \theta_n) = \Phi[a_k(\theta_n - b_k)], \tag{2.8}$$

1PNO model (Lord, 1952, Lord and Novick, 1968):

$$P(Y_{nk} = 1 \mid \theta_n) = \Phi(\theta_n - b_k). \tag{2.9}$$

Finally, by defining the IRF, it becomes possible to identify the likelihood function of a response pattern using unidimensional IRT models. This is due to the local independence assumption, which allows the responses to each item to be considered as independent and identically distributed random variables. As a result, the likelihood function will be the product of the univariate density functions.

In the case of the 2PL model, the likelihood function will be equal to:

$$L(\theta_n|\overrightarrow{Y_n}) = \prod_{k=1}^{K} \left\{ \frac{\exp[a_k(\theta_n - b_k)]}{1+\exp[a_k(\theta_n - b_k)]} \right\}^{Y_{nk}} \left\{ 1 - \frac{\exp[a_k(\theta_n - b_k)]}{1+\exp[a_k(\theta_n - b_k)]} \right\}^{1-Y_{nk}}. \quad (2.10)$$

While for the 2PNO model, it will be equal to:

$$L(\theta_n|\overrightarrow{Y_n}) = \prod_{k=1}^{K} \{\Phi[a_k(\theta_n - b_k)]\}^{Y_{nk}} \{1 - \Phi[a_k(\theta_n - b_k)]\}^{1-Y_{nk}}, \quad (2.11)$$

Defining the likelihood function in this way, allows for estimating the ability using one of the most commonly used methods in the literature, namely maximum likelihood (ML).

Other estimation methods consist of evaluating θ as random variable and estimating it using the *Bayesian approach*.

In this thesis, the focus will be more on estimating the ability given the item parameters, rather than estimating the values of these parameters. For that reason, the ML estimation methods and the Bayesian estimation methods for ability, will be described in the next sections of this chapter.

2.2 ML estimation methods

The ML estimator is the maximizer of the likelihood in Equations (2.10, 2.11) over the range of possible θ values:

$$\hat{\theta}_{n_{MLE}} = \arg\max_{\theta_n} \left\{ L(\theta|\overrightarrow{Y_n}) : \theta_n \in (-\infty, \infty) \right\}. \quad (2.12)$$

This is one of the earliest estimation methods used for IRT models, and it continues to be widely used despite the increase in alternative methods. In fact, it has the properties of being consistent and asymptotically efficient. However, regarding its limitations, for the 3PL model, a unique maximum for

the likelihood function does not always exist and for response patterns with all items correct or all incorrect, no finite ML estimates exist.

Since the first derivative of the likelihood function (Equation 2.12) does not have a closed-form solution, the value of $\hat{\theta}_{n_{MLE}}$ is generally estimated using nonlinear minimization methods, for example employing a Newton-type algorithm (Schnabel et al., 1985, Dennis Jr and Schnabel, 1996). 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:

$$f(x) = -\ln L(\theta_n | \overrightarrow{Y_n}). \tag{2.13}$$

The procedure begins with an initial guess for the solution, θ_{n_0}. This value can be chosen randomly or based on prior knowledge. After that, at each iteration $m = 0, \ldots, M$ 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:

$$\theta_{n_{m+1}} = \theta_{n_m} - \frac{f'(\theta_{n_m})}{f''(\theta_{n_m})}. \tag{2.14}$$

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.

2.3 Bayesian estimation methods

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* $g(\theta_n|\overrightarrow{Y_n})$ starting from a *prior distribution* $g(\theta_n)$ which is hypothesized based on known characteristics of θ:

$$g(\theta_n|\overrightarrow{Y_n}) = \frac{L(\theta_n|\overrightarrow{Y_n})g(\theta_n)}{\int L(\theta_n|\overrightarrow{Y_n})g(\theta_n)\,d\theta_n}, \tag{2.15}$$

where $\int L(\theta_n|\overrightarrow{Y_n})g(\theta_n)\,d\theta_n$ is the *marginal likelihood*, representing the overall probability of observing the data under the model. Just like the likelihood, the posterior distribution in Equation (2.15) can be maximized to define the estimator of θ_n. In this case, the estimator is called the Maximum A Posteriori (MAP) estimator (Lord, 1986, Mislevy, 1986).

$$\hat{\theta}_{n_{MAP}} = \arg\max_{\theta_n} \left\{ g(\theta_n|\overrightarrow{Y_n}) : \theta_n \in (-\infty, \infty) \right\}. \tag{2.16}$$

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 (2.15) becomes proportional to the likelihood function over the support of the prior, and the maximizers in Equation (2.12) and Equation (2.16) are equal. Hence, the MAP estimator shares all the above properties of the ML estimator (i.e. being consistent and asymptotically efficient, but for the 3PL model, a unique maximum for the posterior distribution does not always exist and for response patterns with all items correct or all incorrect, no finite estimates exist). Instead, for nonuniform prior distributions, depending on the choice of prior distribution, the posterior distribution may be multimodal. If so, unless precaution is taken, MAP estimation may result in a local maximum.

Regarding the estimation procedure, as with ML estimators, since there is no closed-form solution, the same nonlinear minimization method (Sec-

tion 2.2) 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 $m = 0, \ldots, M$, there is first an Expectation step (E-step), in which is computed the expected values of individuals' responses to each item based on the current estimates of θ_{n_m} and the item parameters $\boldsymbol{I_K} = (a_k, b_k, c_k)$. These expected values represent the probability of a correct response for each item.

$$\begin{aligned} E[Y_{nk}] &= P(Y_{nk} = 1 | \theta_{n_m}, \boldsymbol{I_K}), \\ E[\overrightarrow{Y_n}] &= (E[Y_{n1}], \ldots, E[Y_{nK}]) . \end{aligned} \tag{2.17}$$

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

$$Q(\theta_n | \theta_{n_m}) = E\left[\ln L(\theta_{n_m} | E[\overrightarrow{Y_n}])\right] + \ln g(\theta_n). \tag{2.18}$$

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.

$$\theta_{n_{m+1}} = \arg\max_{\theta_n} \{Q(\theta_n | \theta_{n_m})\} . \tag{2.19}$$

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

As an alternative to MAP, in Equation (2.16), the Expected A Posteriori (EAP), in Equation (2.20), estimator is typically suggested (Bock and Mislevy, 1982). This estimation method is based on calculating the expected value of the posterior distribution, obtained by integrating it with respect to θ_n.

$$\int \theta_n g(\theta_n)\, d\theta_n. \tag{2.20}$$

In the case of a suitable prior distribution, the EAP estimator always exists, and it offers the advantage of straightforward computation. It doesn't necessitate iterative processes; typically, a single round of numerical integration is sufficient. This attribute was once significant but has diminished in importance due to the increased capabilities of modern calculators.

2.4 Response time

In the earlier sections, some of the most commonly used IRT models were described. These statistical models included IRFs (Equations, 2.4 - 2.9) that establish a mathematical connection between a hidden trait and a student's measurable test response. However, it's important to note that these models do not comprehensively represent all the cognitive processes involved from reading a question to choosing an answer.

These cognitive *processes* (De Boeck and Jeon, 2019) encompass the behaviors that lead a student to choose what they consider to be the most appropriate response from the potential answers. By "appropriate," we do not mean merely correct. For example, one of the most common attitudes when the answer is unknown is to select a random one. We have already shown models that consider this situation (Equations, 2.4, 2.7); however, these model fails to provide the motivations behind a student's decision to rely on a random answer rather than invest more time in finding the correct one.

In summary, relying solely on the pattern of responses provided to define a latent model does not guarantee a comprehensive analysis of the underlying processes in the test-taker's reasoning.

Clearly, having more information about these processes would lead to an improvement in the proposed models and, consequently, the estimation of ability. Fortunately, technological advancements that have occurred since the inception of IRT have allowed progress in this direction. Nowadays, the increasingly widespread use of computerized tests enables obtaining much more information beyond the mere response patterns. One of these valuable and easily acquired pieces of information during a computerized test is undoubtedly the response time (RT).

In fact, processes inherently require time. Returning to the previous example, there clearly is a difference between someone who takes a long time to respond to a specific item and then provides an answer, that could also be wrong, and someone who does the same but spends much less time. In the former case, we are likely witnessing a failed attempt to rely on one's knowledge, whereas in the latter case, we probably have what is known as *rapid guessing*. This is just one of the numerous cases in which RT demonstrates its fundamental importance in clearly distinguishing the functioning of a process.

Naturally, there are other methods to delve into the functioning of a mind at work on a test in detail. For example, eye-tracking systems or the analysis of the physical and mental conditions of students before and during the exam. However, in this work, we will focus solely on RT, both because it is an auxiliary source of information consistently present in computerized tests and also because it is a known and frequently discussed and utilized investigative system that is very easy to monitor.

2.4.1 Response time models

We have just introduced RT as an auxiliary source of information for IRT models. It becomes crucial, for this purpose, to be able to model RT. The

literature, in fact, since the early 1980s, has been exploring which models are most appropriate for describing RT in a test.

Following the classification proposed by De Boeck and Jeon (2019), partially inspired by the work of van der Linden (2009), response time models can now be grouped into four categories.

1. Models in which RT is the sole dependent variable. These models can be classified into three subgroups:

    - Distribution Models for Response Times. These models hypothesize that, given a subject $n = 1, ..., N$, and given an item $k = 1, ..., K$, the response time rt_{nk} is a realization of a random variable RT_{nk}, which follows a known distribution with a variance that increases with the mean, such as the gamma distribution (Maris, 1993), the log-normal distribution (van der Linden, 2006), the ex-Gaussian distribution (Matzke and Wagenmakers, 2009), the Weibull and Gumbel distribution (Loeys et al., 2011), the inverse Gaussian distribution (Lo and Andrews, 2015), and the shifted Wald distribution (Anders et al., 2016). Among these known distributions, one of the most popular is the log-normal distribution proposed by van der Linden (2006) to model RT data due to its easy implementation and good fit to the data. Moreover, it serves as the foundation for developing other models, which will be presented in the following categories.
    - Explanatory Response Models. These models assume that RT is the sum of the times required for all the processes involved in selecting the appropriate response. In fact, it is often assumed that multiple processes contribute to the response selection and that each process requires a different amount of time (Sternberg, 1977, 1985). The sum of these times would be the observed response time. A possible

example, as proposed by De Boeck and Jeon (2019), could be a question of the association type, such as "Son is to aunt as daughter is to?"

In this case, the processes involved would be:

- Encoding: The process required for encoding the terms used, such as "son," "aunt," and "daughter."
- Inference: The process that involves comparing terms A and B (i.e., "son" and "aunt") and leads to the identification of two differences (gender and generation).
- Mapping: The process that involves comparing terms A and C (i.e., "son" and "daughter") and leads to the identification of a single difference (gender).
- Application: The process through which the relationship between terms A and B is applied to term C in order to find the missing term D.

In this example, the RT would be the sum of the time required for performing encoding (X_a), inference (X_b), mapping (X_c), and application (X_d), plus the time needed for reading the question and making a decision (the intercept). Finally, the statistical model is defined by the presence of an error component (ε) that can have different distributions depending on the underlying assumptions.

$$RT = intercept + aX_a + bX_b + cX_c + dX_d + \varepsilon, \tag{2.21}$$

where a, b, c, d are the temporal parameters associated with each of the four processes.

- Response Times as a Function of Response Accuracy. These models reverse the basic assumption in the literature, namely that RT can explain Response Accuracy (RA). In fact, in the IRT context, RA

refers to the alignment between a person's actual ability and their responses to test items. RA can be quantified by comparing a person's actual responses to the predicted probabilities based on their ability level. Higher RA indicates a more precise assessment of a person's skills or traits, while lower accuracy suggests a mismatch between the test and the individual's abilities. For this reason, typically, RT is considered as a factor that can help explain RA. But, in that case, the assumption is that it is instead RA that can explain RT (Novikov et al., 2017).

2. Joint Models. These are models in which both RT and RA are dependent variables. Depending on different characteristics, these models can be divided into three subgroups:

   - Hierarchical Models. These are models that share the same framework, called the Bivariate Generalized Linear Item Response Theory modeling (B-GLIRT) framework (Molenaar et al., 2015). These are bivariate models because they assume the existence of one dimension for RA, interpreted as ability (θ), and another dimension for RT, interpreted as speed (ζ). They are called hierarchical models because the distributions of θ and ζ are evaluated at two levels: first, the marginal distributions of the two latent traits are assumed, and then they are jointly evaluated with a certain degree of correlation. The item parameters are also linked to both RA and RT, and they are correlated with each other. The distinguishing factor among the different models is the choice of the marginal distribution for RT. For the marginal distribution of RA, a 1-, 2-, or 3-parameter IRT model (either logistic or normal ogive) is chosen depending on the case. For the RT distribution, one of the aforementioned known distributions is chosen. One of the early hierarchical models is van der Linden's (2007), which assumes that RT follows a log-normal

distribution, just like the log-normal distribution model himself proposed in 2006. It is also one of the first models to take into account a negative correlation between θ and ζ, which can be explained by the fact that if the respondent wants to prioritize accuracy in the test, they tend to take more time and thus go slower (*speed-accuracy trade-off*). This model has then inspired models that assume variable time during the test (Fox and Marianti, 2016), models with a different distribution of RT, such as the shifted Weibull (Loeys et al., 2011), and models that allow for accommodating most types of distributions, such as the semi-parametric proportional hazards model proposed by Wang et al. (2013) and Kang (2017).

- Diffusion Model. These are models explicitly proposed as an alternative to hierarchical models (van der Maas et al., (van der Maas et al., 2011)). The underlying idea of these models is that there exists a single type of primary process called *information accumulation*, which occurs in response to a *stimulus* caused by the administration of an item. This process of information accumulation is influenced by three fundamental parameters, which are:
  - The diffusion drift parameter, typically represented by its mean v, indicates the propensity (in this case, the average propensity) to choose what is considered to be the correct answer (the possible answer, in these models, is referred to as the *bound* and is denoted by the letter X). This parameter depends on the information accumulated following the item administration and somehow reflects the ability of the test-taker as utilized in a classical IRT model (Equation 2.23).
  - Boundary separation (a), indicating the response caution of the subject, which may be influenced by instructions and rewards. If boundary separation is decreased, both RT and the

probability of terminating at the correct boundary (that is, at the correct answer) are reduced. In this way, the inverse relation between speed and accuracy is naturally accommodated in the model. These parameters influence both the RA (Equations 2.22, 2.23) and the RT (Equation 2.24), as well as their relationship (Equation 2.25). In fact, considering the simplified case of a dichotomous response, the probability of choosing the correct response, i.e., terminating at the upper boundary ($X = 1$), is given by:

$$P_+ = P(X = 1) = \frac{\exp(-2zv) - 1}{\exp(-2av) - 1}, \tag{2.22}$$

and in the even more specific case where the process is *unbiased* (i.e., $z = \frac{a}{2}$), this simplifies and becomes:

$$P_+ = P(X = 1) = \frac{\exp(-av) - 1}{\exp(-2av) - 1} = \frac{\exp(av)}{1 + \exp(av)}, \tag{2.23}$$

which closely resembles the IRFs of logit IRT models (Equations, 2.4 - 2.6).

– Non-Decision Time (T_{er}). Indeed, these models assume that RT is the sum of a Decision Time (DT), which is the time required for the information accumulation process, and a Non-Decision Time (T_{er}) that may include stimulus perception and the time needed to execute a motor response. Therefore, it is the DT that is influenced by the aforementioned parameters (a and v), and its expected value (in the dichotomous case) is equal to:

$$E(DT) = \frac{a}{2v} \frac{1 - \exp(-av)}{1 + \exp(-av)}. \tag{2.24}$$

As for the joint distribution of RA and RT, the joint density function is given by:

$$f_{X,RT}(x,rt) = \frac{\pi\sigma^2}{a^2}\exp\left(\frac{(ax-z)v}{\sigma^2} - \frac{v^2}{2a^2}(rt-T_{er})\right) \times \sum_{m=1}^{\infty}\sin\left(\frac{\pi m(ax-2zx+x)}{a}\right) \times \exp\left(-\frac{1}{2}\frac{\pi^2\sigma^2m^2}{a^2}(rt-T_{er})\right). \quad (2.25)$$

In sum, it is a model primarily based on the concept of a process and directly considers the speed-accuracy trade-off (through parameter a). However, it has its limitations in being based on a one-process assumption (i.e., the existence of information accumulation as the only primary process), making it, ultimately, a kind of rotation of the hierarchical models.

- Race Models: These models hypothesize a *competition* among the so-called *accumulators* (Audley and Pike, 1965, Smith, 2000), which represent the accumulation of evidence supporting one answer over another. For each question, there are as many accumulators as there are available responses. The idea behind these models is that each accumulator requires a different amount of time to reach its upper limit. When an accumulator reaches its upper limit first, the answer associated with that specific accumulator is chosen. According to the model proposed by Rouder et al. (2015), the actual RT for an item is determined by the sum of the time required to reach the upper limit of the "winning" accumulator (*finishing time*) plus a time component required for non-decision processes such as stimulus encoding and response execution (*shift parameter*: ψ). This specific model assumes that each accumulator depends on the latent trait θ, unlike the model proposed by Ranger et al.

(2015). In this latter case, the accumulators can be divided into an *information accumulator*, which refers to the actual correct answer, and *misinformation accumulators*, which refer to all incorrect responses. Thus, two latent traits are hypothesized: θ, which explains the increase in information, and ω, which explains the increase in misinformation. Similarly to diffusion models, each stimulus has an upper limit, and once reached, the corresponding response is selected. However, in this case, the total response time is a function of the processing capacity k, which is the sum of θ and ω, without the addition of any shift parameter. Like diffusion models, race models also have the limitation of relying on a single process (in this case, the "race" to which accumulator reaches its upper limit first) and having a parametrization in two dimensions. The difference from hierarchical models is primarily interpretive.

3. Local Dependency Models. These are models that, like joint models, hypothesize both RT and RA as dependent variables, but they also postulate a deeper connection between these two latent traits, where one can explain the other and vice versa. They can be divided into two main subgroups:

   - Latent variable models with remaining dependencies: these are models in which the dependency between the two latent traits is explained by introducing a *local dependency parameter* (van der Linden and Glas, 2010b, Bolsinova et al., 2017, De Boeck et al., 2017).
   - Class models: these are models in which the existence of multiple response classes is hypothesized, specifically two classes corresponding to two distinct ways of responding, namely the *fast mode* and the *slow mode*. Each class has its specific model and thus a

specific process to arrive at the response. These response classes can be *manifest* (Partchev and De Boeck, 2012) or *latent* (Wang and Xu, 2015, Molenaar et al., 2016, 2018). It is important to emphasize that these classes refer to the type of response (not specific items or individuals). In fact, there are also class models that refer to individuals. In this case as well, we find two classes, namely the *rapid guessers* and the *regular problem solvers* (Meyer, 2010, Jeon and De Boeck, 2019).

The application of these models opens up new and interesting developments regarding the speed-accuracy trade-off (although speed is not explicitly included in these models). In fact, it is confirmed that higher RA values correspond to larger RTs (which in hierarchical models translated to slower speed), but this tends to be true especially for difficult items (Bolsinova et al., 2017), while the opposite relationship holds for easy items (Bolsinova et al., 2017, De Boeck et al., 2017), precisely because easier items, on average, require less time (Partchev and De Boeck, 2012, DiTrapani et al., 2016, Molenaar et al., 2016, 2018).

In summary, class models are especially advantageous when investigating two specific types of processes (fast and slow responses), but their limitation lies in their inability to provide additional information about other types of processes.

4. Response Times as Covariate Models. These are models in which RA is the dependent variable and RT is one of its covariates. Even for this case, the models can be divided into subgroups:

   - Speed-Accuracy tradeoff (SAT) based models: These are models that directly incorporate the SAT (van der Linden, 2007) within the function explaining the probability of a correct response to an item. The *success rate* becomes a (exponential) function of time (Roskam,

1987, Verhelst et al., 1997). Additionally, according to the models proposed by Lohman (1989) and Wang and Hanson (2005), the growth rate of the function can be equated to the latent trait of speed in hierarchical models.

- Generalized Linear Mixed Model (GLMM) based covariate models: These are mixed models that assume that the SAT holds (or is more prominent) depending on the different tasks required by the item. Furthermore, several studies (Goldhammer et al., 2014, 2015, 2017, Naumann and Goldhammer, 2017) have highlighted how the SAT also depends on the respondent's ability itself. In fact, the higher the ability, the stronger the SAT.

Obviously, the classification presented here is not the only possible one, and the models shown, albeit in a fairly general manner, are just some of the theorized ones. However, they are still among the most discussed and used in the literature.

In the following subsections, more specifically, the log-normal distribution model proposed by van der Linden (2006) and the hierarchical model proposed by van der Linden (2007) will be presented. Additionally, a modification of the latter model, proposed by Fox and Marianti (2016) and Fox et al. (2021) will be introduced. The reason why these models will be explored is that they will be used for analyses on real data reported later (Section 2.4.4) and because they form the basis on which the new method for identifying cheaters, proposed in Chapter 3, was developed.

2.4.2 Log-normal distribution model

As previously mentioned, 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:

$$f(rt_{nk}, \zeta_n, \phi_k, \lambda_k) = \frac{\phi_k}{\sqrt{2\pi rt_{nk}}} \exp\left\{-\frac{1}{2}\left[\phi_k\left(\ln rt_{nk} - (\lambda_k - \zeta_n)\right)\right]^2\right\}, \quad (2.26)$$

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. It serves as an equivalent of item difficulty in IRT models and can be positive or negative, with a mean of 0.

- ζ_n is the speed parameter of test-taker n, which represents the latent trait underlying the response time (similar to how ability θ relates to the response pattern). It reflects the constant working speed of the test-taker, accounting for systematic differences in response times given λ_k. A speed value of zero indicates a test-taker who, on average, works at the same speed as the population average. For instance, for an item with a time-intensity $\lambda_k = 4.61$, the average time for a test-taker with average speed ($\zeta_n = 0$) on that item is around 100 seconds on the regular time scale ($e^{4.61} \cong 100$). Conversely, a negative (positive) speed indicates a test-taker who, on average, works slower (faster) than the population average.

- ϕ_k is the time-discrimination parameter of item k, representing the item sensitivity to different speed levels of test-takers. It serves as an analogue to the item discrimination parameter in IRT models and is strictly greater than zero. It is defined as the reciprocal of the standard deviation of the normal distribution[1]:

$$\phi_k = \frac{1}{\sigma_k}. \quad (2.27)$$

[1] Since RT_{nk} follows a log-normal distribution, its natural logarithm, denoted as $\ln RT_{nk}$, will have a normal distribution with variance σ^2_{nk}.

Therefore, a higher value for ϕ_k indicates less variability in the log response time discrimination of the item among individuals with different levels of ζ.

From Equation (2.26), since it represents the probability density function of a standard normal distribution, we can confirm that the standard deviation is the reciprocal of the time-discrimination parameter and we can also identify the mean,

$$\mu_{nk} = \lambda_k - \zeta_n, \tag{2.28}$$

from which,

$$\ln RT_{nk} \sim N\left(\lambda_k - \zeta_n, \frac{1}{\phi_k^2}\right). \tag{2.29}$$

However, for any value of ε, the distribution in Equation (2.26) remains the same under the transformations,

$$\begin{aligned} &\lambda_k - \varepsilon, \\ &\zeta_k - \varepsilon. \end{aligned} \tag{2.30}$$

Therefore, the model is not identified. To establish identifiability, the following constraint is imposed on the parameter ζ_n:

$$\sum_{n=1}^{N} \zeta_n = 0. \tag{2.31}$$

This constraint also brings an advantage in terms of interpretation for both λ_k and ζ_n. In fact, Equation (2.28) implies that:

$$\frac{\sum_{n=1}^{N} \lambda_k}{N} - \frac{\sum_{k=1}^{K} \zeta_n}{K} = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} \mu_{nk}}{NK}. \tag{2.32}$$

By imposing the constraint in Equation (2.31), it is reduced to:

$$\frac{\sum_{n=1}^{N} \lambda_k}{N} = \frac{\sum_{n=1}^{N} \sum_{k=1}^{K} \mu_{nk}}{NK}. \tag{2.33}$$

That is, the average item parameter λ_k is equal to the average expected log-time over the persons and items. As a consequence, ζ_n is a deviation from this average.

Finally, given the probability density function in Equation (2.26), it follows that the logarithm of RT can be expressed using the following probabilistic function:

$$\begin{gathered} \ln RT_{nk} = \lambda_k - \zeta_n + \varepsilon_{nk}, \\ \varepsilon_{nk} \sim N(0, \sigma^2_{\varepsilon_k}) \end{gathered} \tag{2.34}$$

As mentioned, this model is the basis for another widely used model in the literature, namely the hierarchical log-normal model (van der Linden, 2007).

2.4.3 Hierarchical distribution model

The underlying idea of this model, as previously mentioned, is that the RA and RT are jointly and, with a certain degree of correlation, the dependent variables. It has also been mentioned that it is called a hierarchical model because it is specified at multiple levels.

At the first level, two distinct models are presented for the RA and RT. At the second level, the prior distributions of the parameters of the first-level models are defined, and the parameters of the two models are allowed to be dependent. This improves the estimation of the parameters for both models because the estimation of the RA model parameters utilizes the additional information provided by RTs, and vice versa.

At the third level, hyperparameters are defined, which correspond to the parameters of the prior distributions defined at the second level.

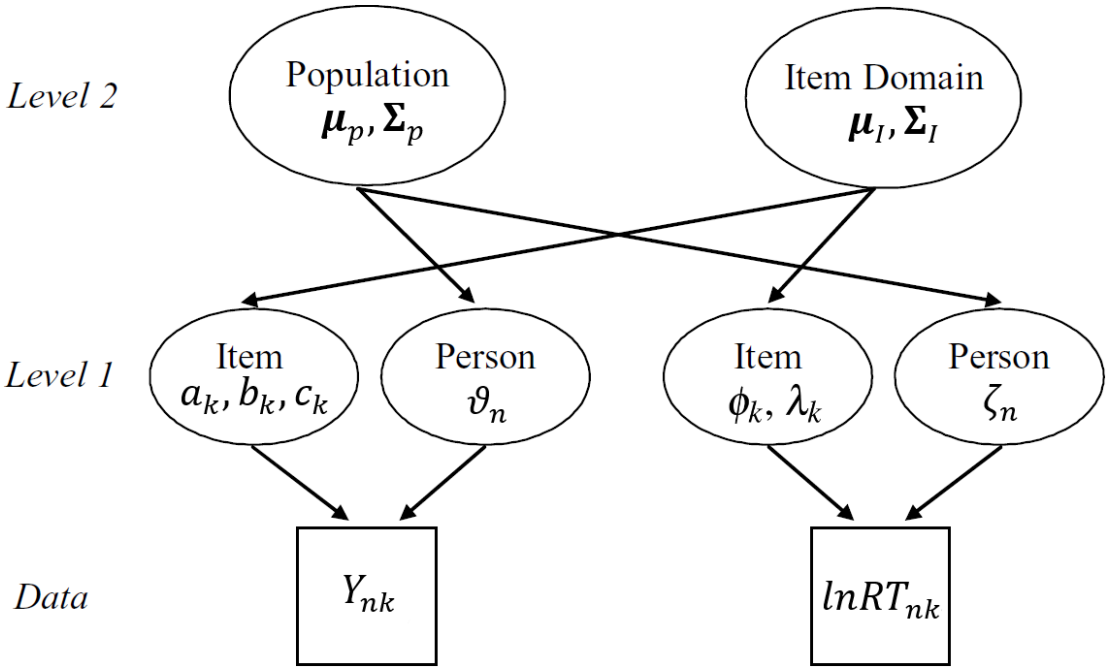

Fig. 2.1 Hierarchical structure of the jointly distribution of RT and RA.

First-Level Models: For the RA, a classic IRT model is hypothesized. Like in the case of van der Linden (2007), a 3PL model is chosen (Equation 2.4). On the other hand, for the RT, one of the distribution models described in the previous paragraph is selected. In the specific case examined in this study, it is the log-normal model described by van der Linden (2006) (Equation 2.26).

Second-Level Models: At this level, the RA and RT are modeled jointly. Multivariate distributions are chosen for both the item parameters $\boldsymbol{\psi}_k = (a_k, b_k, c_k, \phi_k, \lambda_k)$ and the individual parameters $\boldsymbol{\xi}_n = (\theta_n, \zeta_n)$.

Concerning the item parameters, a multivariate normal distribution with

parameters $\boldsymbol{\mu}_I$ and $\boldsymbol{\Sigma}_I$ is chosen as the prior distribution:

$$\begin{aligned}\left(a_k, b_k, c_k, \phi_k, \lambda_k\right)^\top &\sim N(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I),\\ \boldsymbol{\mu}_I &= \left(\mu_a, \mu_b, \mu_c, \mu_\phi, \mu_k\right)^\top,\\ \boldsymbol{\Sigma}_I &= \begin{pmatrix} \sigma_a^2 & \sigma_{a,b} & \sigma_{a,c} & \sigma_{a,\phi} & \sigma_{a,\lambda} \\ \sigma_{a,b} & \sigma_b^2 & \sigma_{b,c} & \sigma_{b,\phi} & \sigma_{b,\lambda} \\ \sigma_{a,c} & \sigma_{b,c} & \sigma_c^2 & \sigma_{c,\phi} & \sigma_{c,\lambda} \\ \sigma_{a,\phi} & \sigma_{b,\phi} & \sigma_{c,\phi} & \sigma_\phi^2 & \sigma_{\phi,\lambda} \\ \sigma_{a,\lambda} & \sigma_{b,\lambda} & \sigma_{c,\lambda} & \sigma_{\phi,\lambda} & \sigma_\lambda^2 \end{pmatrix},\end{aligned} \tag{2.35}$$

with density function:

$$f(\boldsymbol{\psi}_k; \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I) = \frac{1}{\sqrt{(2\pi)^5 |\boldsymbol{\Sigma}_I|}} \exp\left(-\frac{1}{2}(\boldsymbol{\psi}_k - \boldsymbol{\mu}_I)^\top \boldsymbol{\Sigma}_I^{-1} (\boldsymbol{\psi}_k - \boldsymbol{\mu}_I)\right). \tag{2.36}$$

For the two discrimination parameters a_k and ϕ_k, the constraint of being positive is imposed.

As for the individual parameters, a bivariate normal distribution with parameters $\boldsymbol{\mu}_P$ and $\boldsymbol{\Sigma}_P$ is chosen as the prior distribution:

$$\begin{aligned}\begin{pmatrix}\theta_i \\ \zeta_i\end{pmatrix} &\sim N(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P),\\ \boldsymbol{\mu}_P &= \begin{pmatrix}\mu_\theta \\ \mu_\zeta\end{pmatrix},\\ \boldsymbol{\Sigma}_P &= \begin{pmatrix}\sigma_\theta^2 & \sigma_{\theta,\zeta} \\ \sigma_{\theta,\zeta} & \sigma_\zeta^2\end{pmatrix},\end{aligned} \tag{2.37}$$

with density function:

$$f(\boldsymbol{\xi}_n; \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P) = \frac{1}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}_P|}} \exp\left(-\frac{1}{2}(\boldsymbol{\xi}_n - \boldsymbol{\mu}_P)^\top \boldsymbol{\Sigma}_P^{-1} (\boldsymbol{\xi}_n - \boldsymbol{\mu}_P)\right). \tag{2.38}$$

Third -Level Models: In this third and final level, the *hyperprior distributions* for the two bivariate models described in the second level are defined. A hyperprior distribution is a distribution that is used to model uncertainty or variation in the parameters of the prior distribution. In hierarchical models, parameters themselves can be treated as random variables, and their distribution is referred to as a hyperprior. Regarding the item parameters, van der Linden (2007) suggests simplifying the model by separating the parameterization of c_k from the other item parameters. For this reason, we introduce a new vector of item parameters that does not include c_k:

$$\begin{aligned} \boldsymbol{\psi}_k^* &= \left(a_k, b_k, \phi_k, \lambda_k\right)^\top \sim N_4(\boldsymbol{\mu}_I^*, \boldsymbol{\Sigma}_I^*), \\ \boldsymbol{\mu}_I^* &= \left(\mu_a, \mu_b, \mu_\phi, \mu_k\right)^\top, \\ \boldsymbol{\Sigma}_I^* &= \begin{pmatrix} \sigma_a^2 & \sigma_{a,b} & \sigma_{a,\phi} & \sigma_{a,\lambda} \\ \sigma_{a,b} & \sigma_b^2 & \sigma_{b,\phi} & \sigma_{b,\lambda} \\ \sigma_{a,\phi} & \sigma_{b,\phi} & \sigma_\phi^2 & \sigma_{\phi,\lambda} \\ \sigma_{a,\lambda} & \sigma_{b,\lambda} & \sigma_{\phi,\lambda} & \sigma_\lambda^2 \end{pmatrix}. \end{aligned} \tag{2.39}$$

As hyperpriors for $\boldsymbol{\psi}_k^*$, independent normal/inverse-Wishart prior distributions are chosen; that is:

$$\begin{aligned} \boldsymbol{\mu}_I^* | \boldsymbol{\Sigma}_I^* &\sim \text{MVN}(\boldsymbol{\mu}_{I0}, \boldsymbol{\Sigma}_I / \kappa_{I0}), \\ \boldsymbol{\Sigma}_I^* &\sim \text{Inverse-Wishart}(\boldsymbol{\Sigma}_{I0}^{-1}, \nu_{I0}), \end{aligned} \tag{2.40}$$

where $\nu_{I0} \geq 4$ is a scalar degrees-of-freedom parameter, $\boldsymbol{\Sigma}_{I0}^{-1}$ is a 4×4 (positive definite symmetric) scale matrix for the hyperprior on $\boldsymbol{\Sigma}_I^*$, and $\boldsymbol{\mu}_{I0}$ and κ_{I0} are the vector with the means of the posterior distribution and the strength of prior information about these means, respectively.

Instead, for the guessing parameter a Beta hyperprior distribution is assumed,

$$c_k \sim \text{Beta}(\gamma, \delta), \quad k = 1, \ldots, K. \tag{2.41}$$

Regarding the person parameters, once again a normal/inverse-Wishart hyperprior distribution is chosen:

$$\begin{aligned} \boldsymbol{\mu}_P | \boldsymbol{\Sigma}_P &\sim \text{MVN}(\boldsymbol{\mu}_{P0}, \boldsymbol{\Sigma}_P / \kappa_{P0}), \\ \boldsymbol{\Sigma}_P &\sim \text{Inverse-Wishart}(\boldsymbol{\Sigma}_{P0}^{-1}, \nu_{P0}), \end{aligned} \tag{2.42}$$

where, the parameters for the hyperprior distributions of $\boldsymbol{\mu}_P$ and $\boldsymbol{\Sigma}_P$ are defined analogously to those of $\boldsymbol{\mu}_I^*$ and $\boldsymbol{\Sigma}_I^*$, and with a minimum number of degrees of freedom for the scale parameter ν_{P0} set to 2.

For this choice of hyperprior distributions, the joint posterior distribution of the parameters factors is:

$$\begin{aligned} f(\xi, \psi, \mathbf{c}, \mu_P, \mu_I^*, \gamma, \Sigma_P, \Sigma_I^*, \delta \mid \mathbf{y}, \mathbf{rt}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} f(y_{nk}; \theta_n, a_k, b_k, c_k) \times \\ \times f(rt_{nk}; \zeta_n, \phi_k, \lambda_k) f(\xi_n; \mu_P, \Sigma_P) f(\psi_i^*, c_k; \mu_I^*, \gamma, \Sigma_I^*, \delta) \times \\ \times f(\mu_P, \Sigma_P) f(\mu_I^*, \Sigma_I^*) f(c_k). \end{aligned} \tag{2.43}$$

Lastly, to establish identifiability, the following constraints are imposed:

$$\mu_\theta = 0, \quad \sigma_\theta^2 = 1, \quad \sum_{n=1}^{N} \zeta_n = 0. \tag{2.44}$$

The first two constraints are usual in IRT parameter estimation. The third constraint is the same used in the log-normal distribution model for RT and gives the model the same advantages previously described (Section 2.4.2).

As previously mentioned, this hierarchical model was subsequently modified by Fox and Marianti (2016). The substantial modification was made to the

first level of the hierarchical model, specifically concerning the distribution of RTs. A new time-discrimination parameter, denoted as ϕ_k^*, was introduced. Unlike its previous representation as the reciprocal of standard deviation (Equation 2.27), ϕ_k^* now represents the slope of the speed. This parameter also models the covariances between RTs, which is expected to enhance the model flexibility (Fox and Marianti, 2016). Therefore, the logarithm of RT can be expressed using the following probabilistic formula:

$$\begin{aligned} \ln RT_{nk} &= \lambda_k - \phi_k^* \zeta_n + \varepsilon_{nk}, \\ \varepsilon_{nk} &\sim N(0, \delta_{\varepsilon_k}^2). \end{aligned} \tag{2.45}$$

Moreover, the error component in Equation (2.45) can capture fluctuations in RTs resulting from the random actions of a test-taker. When individuals vary in their pace of responding, insert brief pauses during the test, or alter their time management, the RTs might exhibit more systematic variability than can be attributed to the underlying average performance. The error component specific to each item can account for these distinctions and prevents any bias in the parameter estimator.

Regarding the other two levels of the hierarchical model, the specifications have remained unchanged compared to the model by van der Linden (2007) described earlier. However, unlike that model, the identification avoids restricting $\sigma_\theta^2 = 1$. For that reason, the variance of the latent factors is identified by restricting the product of discriminations and time discriminations to one, $\prod_{k=1}^K a_k$ and $\prod_{k=1}^K \phi_k$, respectively. Additionally, a different approach is suggested for determining the mean of the latent factors. This involves either setting the sum of the difficulty and time-intensity parameters to zero, $\sum_{k=1}^K b_k = 0$ and $\sum_{k=1}^K \lambda_k = 0$ respectively, or fixing the mean of the ability parameter to zero, $\mu_\theta = 0$, and the mean of the speed parameter to zero, $\mu_\zeta = 0$.

Finally, with regard to the model estimation methods, these are the same for both the model by van der Linden (2007) and the modified model by Fox and Marianti (2016). These are Markov Chain Monte Carlo (MCMC) methods. These methods are a class of statistical techniques used to approximate complex probability distributions, especially when direct sampling is difficult or infeasible. MCMC methods are widely applied in Bayesian statistics. The fundamental idea behind MCMC is to construct a Markov chain that explores the target distribution of interest, eventually producing samples that closely resemble draws from that distribution. The Markov chain is a sequence of random states where each state depends only on the previous one. It is designed to have a stationary distribution, and in MCMC, this stationary distribution represents the desired probability distribution for sampling.

Within this class of methods, the Gibbs sampling (Geman and Geman, 1984) is commonly used for the hierarchical models. It is a specific MCMC algorithm, particularly useful for multivariate problems. It is used for sampling from the conditional distributions of each variable one at a time, given the current values of all the other variables. First of all, initial values for all the variables in the model are chosen. Then, for each variable in the model, it is sampled a new value from its conditional distribution given the current values of all other variables. This step is done iteratively for each variable in the model. The conditional distribution for each variable is derived from the joint probability distribution of all variables in the model, with all other variables held fixed. This process is repeated for a chosen large number of iterations. After that, the convergence of the Markov chain is assessed by examining the samples obtained after a *burn-in* period (early iterations where the chain may not be in equilibrium). The samples obtained after convergence are drawn from the joint distribution of all variables in the model.

Regarding the estimation of the parameters of the hierarchical model, the main steps of the procedure used by (Fox and Marianti, 2016) are summarized below (for a scenario without the c_k guessing parameter).

First, to simplify the sampling process, two auxiliary variables are defined using a technique called *data augmentation*. This technique involves introducing a latent (or auxiliary) variable Z, connected to the observed data through a one-to-many relationship. The variable Z is generally constructed in such a way that:

$$P(y|\theta) = \int_{f(z)=y} p(z|\theta)dz. \tag{2.46}$$

In this way, it is equivalent to performing inference on the parameter θ when using the model for the observed data $p(y|\theta)$ or the augmented data model $p(z|\theta)$. This technique is advantageous when it is easier to sample from $p(z|\theta,y)$ and $p(\theta|z)$ than to sample from $p(y|\theta)$. In the case of the hierarchical model under examination, the auxiliary variable z for the given responses y_{nk} is defined as follows:

$$\begin{aligned} z_{nk} &= a_k\theta_n - b_k + e_{nk}, \\ e_{nk} &\sim N(0,1). \end{aligned} \tag{2.47}$$

In this way:

$$Y_{nk} = \begin{cases} 1 & \text{if } Z_{nk} > 0 \\ 0 & \text{if } Z_{nk} \leq 0 \end{cases} \tag{2.48}$$

Using the auxiliary variable in Equation (2.47), the conditional distribution becomes a normal distribution, making it straightforward to sample from:

$$Z_{nk}|Y_{nk},\theta_i,\mathbf{\Psi}_k^* \sim N(a_k\theta_n - b_k, 1). \tag{2.49}$$

After the definition of the auxilary variable in Equation (2.47), the initial parameter values are determined by separately estimating the model for responses and the model for RT. Subsequently, the algorithm follows the following steps for each iteration $m = 1, \ldots, M$:

- *Step 1*: Sample the augmented data from $p(z_{nk}|a_k, b_k, \theta_n)$, given the previous values of the item parameters and ability.
- *Step 2*: Sample the item parameter values from:

$$p(\boldsymbol{\Psi}_k^*|\boldsymbol{z}_k^*, \boldsymbol{\xi}, \boldsymbol{\mu}_I^*, \boldsymbol{\Sigma}_I^*), \tag{2.50}$$

  where:

$$\begin{gathered} \boldsymbol{z}_k^* = (\boldsymbol{H}_\theta \oplus -\boldsymbol{H}_\zeta)\boldsymbol{\Psi}_k^* + e_k, \\ e_k \sim N(0, I_{2N}), \\ \boldsymbol{H}_\theta = (\theta, -1_N), \\ \boldsymbol{H}_\zeta = (-\zeta, 1_N). \end{gathered} \tag{2.51}$$

- *Step 3*: Sample the person parameter values from:

$$p(\boldsymbol{\xi}_n|\boldsymbol{z}_n^*, \boldsymbol{\Psi}^*, \boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P), \tag{2.52}$$

  where:

$$\begin{gathered} \boldsymbol{z}_n^* = (\boldsymbol{a} \oplus -\boldsymbol{\phi})\boldsymbol{\xi}_n + e_n, \\ e_n \sim N(0, I_{2K}). \end{gathered} \tag{2.53}$$

- *Step 4*: Sample the hyperparameter values from:

$$p(\boldsymbol{\mu}_I^*|\boldsymbol{\Sigma}_I^*, \boldsymbol{\mu}_{I0}, \boldsymbol{\Psi}^*, \boldsymbol{V}_{I0}), \tag{2.54}$$

and from:

$$\begin{gathered} p(\sigma_{\theta,\zeta}|\boldsymbol{\theta},\boldsymbol{\zeta},\tilde{\sigma}^2_\zeta,\bar{\sigma}_{\theta,\zeta},\sigma^2_\rho), \\ \tilde{\sigma}^2_\zeta = \sigma^2_\zeta - (\sigma_{\theta,\zeta})^2. \end{gathered} \tag{2.55}$$

These 4 steps are iteratively repeated until the completion of the M-th iteration.

In the following subsection, a study on real data is presented, highlighting the capabilities and operation of this joint estimation approach.

2.4.4 Example with RT real data

To show the features and the potential of the hierarchical model described above, a study was conducted using real data from Italian students, which included both their responses and the respective RTs to a standardized school test. The hierarchical model used for this purpose followed the approach of Fox and Marianti (2016), and it was entirely implemented within an *R* package *LNIRT* by Fox et al. (2021). The data used for this joint analysis was provided by the Italian National Institute for the Evaluation of the Education and Training System (INVALSI).

In Italy, INVALSI every year administers standardized tests via Computer Based Testing (CBT) to students attending grades 8, 10, and 13. In this study, the 2018 mathematics data for grade 10 were used to jointly estimate the ability and speed of students.

The tests are administered to the whole student population, around 500,000 students. INVALSI also builds a random sample of around 41,000 units. The sampling procedure is a two-stage with Italian geographical region and school track stratification at the first stage. The units of the first stage are the schools and the units of the second stage are the classes. In this study were analyzed the results of the sample after a cleaning procedure from missing and implausible values with respect to RTs ($N = 35,970$). INVALSI imposes

a time limit of 90 minutes on grade 10 tests, which is considered enough for students to read and answer all the questions.

Since the INVALSI tests were generated from a Rasch item bank, and the *LNIRT* package use the normal ogive models, the difficulty parameters were reparameterized for the logit model (multiplying the values obtained from the estimates by the conversion variable equal to 1.7).

The main results for item parameters are summarized in Table 2.1, which shows mean, minimum, and maximum of the EAP estimates.

	Item Difficulty (Rasch Model)	**Time Intensity**	**Time Discrimination**
Mean	-0.070	4.229	1.175
Minimum	-2.574	3.114	0.011
Maximum	2.726	5.151	2.288

Table 2.1 Item parameters. Hierarchical model. LNIRT.

For person parameters, the estimates of ability and speed are given in Table 2.2.

	Person Ability	**Person Speed**
Mean	0.000	0.000
Minimum	-2.311	0.611
Maximum	1.946	2.283

Table 2.2 Person parameters. Hierarchical model. LNIRT.

The ability follows a normal distribution, while the speed distribution curve is slightly skewed. From the residual analysis, it turns out that the residuals of the response times violate the assumption of log-normal distribution for most items. Following several analyses, it was possible to note that this violation is due to the large number of test-takers (35,970).

The correlation matrices for item parameters are given in Table 2.3. The analysis of these results shows that there is, on average, a positive relationship

between the difficulty of the items and their intensity and discriminating power, in terms of time. This means that the most difficult (easy) items are also the ones that discriminate better (worse) and require more (less) time to perform. The negative correlation between time-discrimination and time-intensity, on the other hand, indicates that on average the items that require more (less) time are the ones that discriminate worse (better), but with a very low and not significant magnitude.

	Item Difficulty	**Time Intensity**	**Time Discrimination**
Item Difficulty	1.000	0.370 (0.000)	0.234 (0.004)
Time Intensity	0.370 (0.000)	1.000	- 0.014 (0.436)
Time Discrimination	0.234 (0.004)	- 0.014 (0.436)	1.000

Table 2.3 Item correlation matrix. Value (p-value). Hierarchical model. LNIRT.

Table 2.4 shows the correlation matrix for person parameters. It provides important information about the correlation between the speed and ability of the test-takers (-0.574), which is negative and significant (p-value < 0.001). So, test-takers with a higher (lower) ability tends to be slower (faster).

	Person Ability	**Person Speed**
Person Ability	1.000	-0.574 (0.000)
Person Speed	-0.574 (0.000)	1.000

Table 2.4 Person correlation matrix. Value (p-value). Hierarchical model. LNIRT.

This result goes to consolidate the speed-accuracy trade-off hypothesis (van der Linden, 2007), for which those who are prepared want to engage and show their skills, even during a test that does not directly affect their school average, while those who are less prepared tend to be less interested and more hasty.

In conclusion, RTs can indeed be effectively utilized to implement the ability estimation process and study its relationships with speed. Additionally, the analysis of item parameters helps in better understanding the intrinsic

characteristics of the questions. This information can be used to develop tests (both linear and adaptive) that are more effective and can better meet all the constraints that can be imposed, such as those related to the test completion time.

While this section has focused on RT and its applications to IRT models, the next one will cover, in its essential points, another key topic for this study, namely CAT.

2.5 Computerized Adaptive Testing (CAT)

For a long time, educational testing primarily centered around paper-and-pencil exams and performance assessments. Starting in the late 1980s, with the widespread adoption of personal computers in education, these testing formats expanded to become suitable for computer-based delivery. The utilization of computer-based testing (CBT) offers several advantages. It enables on-demand testing, allowing examinees to take tests whenever and wherever they are ready. Moreover, modern PCs' computational power and multimedia capabilities can be harnessed to create innovative question formats and more realistic testing environments. Additionally, computers can enhance the statistical accuracy of test scores through Computerized Adaptive Testing (CAT) (Wainer et al., 2000, van der Linden and Glas, 2010a). In CAT, instead of administering the same fixed test to every examinee, the test is tailored on the examinee's ability estimate, updating the estimate after each new answer and selecting subsequent questions to optimize the measurement precision. The concept of adaptive item selection has historical roots, dating back to practices like the Binet and Simon (1905) intelligence test and oral examinations, where questions were tailored to an examinee's perceived knowledge level.

The development of IRT (Section 2.1) in the mid-20th century provided a solid psychometric foundation for adaptive testing. The initial research into implementing adaptive testing focused on finding approximate estimation methods and alternative adaptive formats suitable for traditional paper-and-pencil testing, owing to the limitations of early computers. As computer technology advanced, adaptive testing became feasible for large-scale, high-stakes testing programs. The transition from paper-and-pencil to CBT gained momentum with the diffusion of the first personal computers and subsequently expanded into various fields, including psychology, marketing, and health-outcome research.

The shift to computerized testing administration offered several benefits, such as flexible test scheduling for examinees, more comfortable test-taking environments, faster electronic data processing and score reporting, and a broader range of question types and content.

The greater advantages of CAT over linear tests can be better understood by considering how the latter are typically constructed. Generally, newly crafted items are evaluated for difficulty and placed in pretest sections by test experts. Items that pass statistical scrutiny during the pretest phase (or *calibration phase*) become eligible for the final test form assembly. A preliminary test form is created using automated algorithms for test assembly, which is then reviewed and potentially adjusted by test experts. This form is then pre-equated before operational administration to a sample of examinees. Subsequently, number-right scores are transformed onto a common scale by psychometricians employing IRT scaling and true-score equating. The time between operational administrations and score reporting may also take several weeks.

In contrast, within a CAT environment, item selection and ability estimation happen in real-time. Consequently, computer algorithms must take on the roles of both test experts and psychometricians. Since the test adapts

to the examinee, the task of item selection and ability estimation becomes significantly more challenging, requiring robust procedures for solving this complex measurement problem with minimal or no human intervention.

Another subtle distinction between linear and CAT formats is that, as mentioned earlier with the linear example, item selection and ability estimation in linear tests usually occur separately, albeit sometimes utilizing similar methods such as IRT. In CAT, however, item selection and ability estimation occur in tandem. The efficiency of ability estimation is closely tied to the selection of appropriate items for an individual, creating a circular relationship between item appropriateness and interim ability estimates.

The structure of CAT is, in essence, quite complex. To summarize it for descriptive purposes, it can be said to consist of various phases (Wainer et al., 2000, van der Linden and Glas, 2010a):

- Ability initialization.
- Selection of initial items.
- Estimation of the first ability.
- Estimation of interim abilities.
- Selection of the next item.
- Estimation of the final ability.

As previously mentioned, the item selection process and the interim ability estimation process are directly linked to each other, and this cycle continues until the process reaches a so-called *stopping rule* (van der Linden and Glas, 2010a).

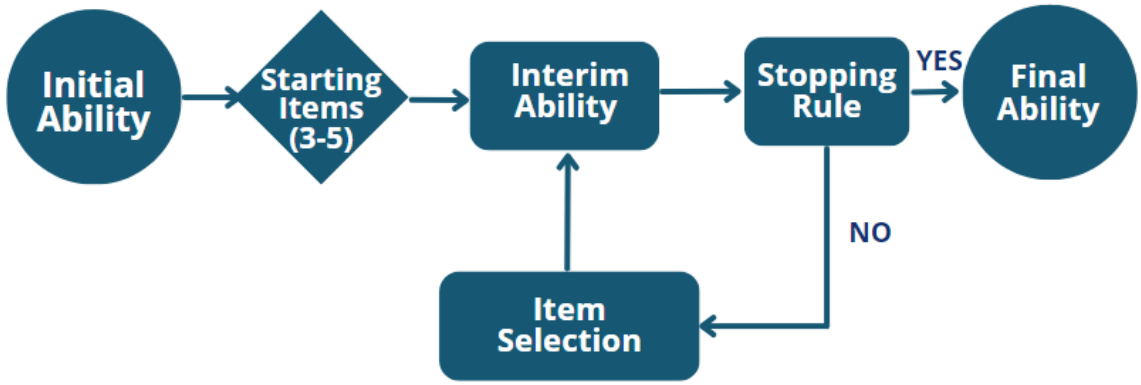

Fig. 2.2 Flow chart of a CAT process.

To delve into more detail, the single steps are examined more specifically below.

- Ability initialization. In order to start the adaptive process, the computer algorithm that manages the CAT needs to know the examinee's presumed ability. This is primarily to kickstart the subsequent ability estimation processes, which, as mentioned in Sections (2.2, 2.3), typically employs iterative methods that require the selection of an initial ability. There are some common choices for θ_{0_n}. In cases where there is no prior information about the specific examinee, θ_{0_n} is set to 0 or randomly chosen (*random initialization*). However, if there is available information about the individual ability (for example, if one or more tests have been previously taken), that information can be used to choose the initial value for the ability.

- Selection of initial items. Similarly, the selection of initial items is crucial in starting the automated process. These are the items that will be administered before the first ability estimate is made. They are essential for collecting as much information as possible to kickstart the estimation process. For this reason, there must be more than one initial item, but they also should not be too many, as this could make the test overly

long and negate the actual advantage of CATs, which is to adapt to the respondent's true ability. Typically, 3-5 items are used for this purpose. There are various possibilities for selecting these initial items. One approach is to assume θ_{0_n} as the true ability of the examinee and choose the items accordingly. However, this approach tends to be problematic when the same value is chosen for all θ_{0_n}, for example $\theta_{0_n} = 0$ for each $n = 1, \ldots, N$, as it risks giving the same initial items to all examinees. The alternative and commonly used approach is to select them randomly.

- Estimation of the first ability. This involves estimating the first ability following the administration of the initial items.

  As mentioned previously in Section 2.2, the ML estimation method does not yield finite estimates for response patterns where all items are answered correctly or incorrectly. This limitation poses challenges for ability estimation, particularly at the beginning of a test when such response patterns are more likely to occur. Several approaches have been proposed to address this issue.

  First, one suggestion is to temporarily fix the ability estimate at a small value for incorrect responses or a large value for correct responses until finite estimates can be obtained. Second, in some cases, ability estimation is delayed until the examinee has answered a larger set of items. Third, this problem has prompted the use of Bayesian methods like EAP estimator. Fourth, when relevant empirical information about examinees, such as scores on earlier related tests, is available, initial ability estimates can be derived from this supplementary data.

  However, none of these solutions is entirely flawless. The first two approaches involve arbitrary choices regarding ability values or the selection of items. The third approach requires choosing a prior distribution, which, in the absence of response data, heavily influences the

choice of the first item (Section 2.3). If the prior distribution is far from the true ability of the examinee, it can be counterproductive and may lead to a longer initial string of correct or incorrect responses than necessary. Regarding the fourth solution, while there are no technical obstacles to using empirical priors, their selection should be done with caution. For instance, relying on general background variables can introduce bias and should be avoided.

Fortunately, the challenge of inferring an initial ability estimate is primarily an issue for short tests, such as 10-item tests within a battery. For longer tests, typically consisting of more than 20 to 30 items, the ability estimator usually has sufficient opportunities to recover from a suboptimal initial estimate.

In any case, regardless of the chosen estimation method, this initial estimated ability serves to initiate the item selection process, which will be specifically discussed in the next section. The key point is that once this first ability is estimated, the adaptive algorithm will select the first item outside of the initial items, and the answer to that item will be used to update the ability estimate. From this point forward, the new ability estimates will be referred to as interim ability estimates.

- Estimation of interim abilities θ_{m_n}. This is the estimation of ability that gets updated after each answer to a new item. It is the heart of the iterative process because it is from the value of this estimate that the next item will be selected.

  As with the estimation of the initial ability, both ML estimation methods (especially if a certain number of answers have already been collected, to avoid the convergence issues mentioned earlier) and Bayesian estimation methods can be used. Furthermore, the chosen method does not necessarily have to be the same as that used for the initial ability

estimate, nor does it have to coincide with the one chosen for the final ability estimate. This, combined with the fact that the item selection process also changes depending on the chosen method (Section 2.5.1), makes CAT highly versatile and modifiable in one or more of its parts, depending on the needs.

This iterative process of interim ability estimation and item selection continues for a number of iterations $m = 1, \ldots, M$, until the selected stopping rule is reached.

- Stopping rule. This is the rule chosen to terminate the iterative process of item selection and interim ability estimation. In this case too, there are various options, but usually, two are the most commonly used.

  The first option is to stop the process once a fixed M number of items have been administered. In this case, the test is generally referred as *fixed length test*.

  The other procedure is to stop the test when the estimation of interim ability stabilizes, meaning when the standard deviation of the estimated ability falls below a predetermined value known as the *target value*. In this case, the test is called *variable length test*.

  A common practice is to consider both procedures and end the test when the first one is met. Sometimes, for variable length tests, a minimum test length value is also chosen to ensure that there is a minimum amount of data available for a plausible estimate of the final ability.

- Estimation of the final ability θ_n. While final ability estimates should ideally possess optimal statistical properties, their primary purpose shifts away from guiding item selection. Instead, they serve to offer the examinee a meaningful summary of their performance in the form of the best possible score. To achieve this, final estimates are sometimes converted into an equated number-correct score on a reference test, which is es-

sentially a released linear version of the test being administered. Two common methods for performing this conversion are the test characteristic function (Lord, 1980) and the equipercentile transformation (Segall, 1997). The former becomes known once the test items are calibrated, while the latter requires estimation through a separate empirical study. To simplify the scoring process for examinees and avoid the need to explain complex ML scoring methods, Stocking (1996) suggested a modification to the likelihood equation, ensuring that its solution forms a monotonic relationship with the number-correct score. However, the need for score adjustments can be entirely eliminated by implementing appropriate constraints on item selection during the test, automatically equating the number-correct scores on an adaptive test to those on a reference test (van der Linden and Glas, 2010a).

Determining the best method for ability estimation is intricately connected to various other aspects of CAT. Firstly, the choice of the criterion used for item selection plays a critical role. Additionally, factors like the composition of the item pool, whether the estimation procedure incorporates collateral information about the examinees, and the presence of content constraints on item selection can all impact ability estimates.

For simplicity in this study, which is focused on dichotomized multiple-choice items (correct answer = 1 and wrong answer = 0), the value of the final ability θ_n will always be left unchanged. Therefore, no conversion will be applied to make it a score, since this is not the focus of the study.

Just as was previously done in Sections (2.2, 2.3) regarding the main IRT estimation models, next Section 2.5.1 will present some of the main Item Selection Criteria (ISC) prevalent in the literature. Again, in this case, RT can be used as an auxiliary source of information that enhances the methodologies. Therefore, after an initial analysis of the main ISC, a series of methods that also use informations about RT will be proposed.

It's worth noting here that while RT can be used in CAT both in the phases of estimating the different abilities (first ability, interim ability, and final ability) and in item selection processes, it tends to be mostly used in the latter case. This is because, as seen in Section 2.4.3, joint estimation methods tend to be complex and time-consuming (using MCMC techniques) and are, therefore, hardly applicable during a CAT. This point will also be discussed in Chapter 3.

2.5.1 Item Selection Criterion

The term Item Selection Criterion (ISC) refers to the method used to choose the next item to administer in a CAT. There are many such criteria, but all of them are based on the principle of exploiting the information that is collected during the test in order to choose the best item to administer (Wainer et al., 2000, van der Linden and Glas, 2010a). Some of the ISC examined in this work make use also of the information about the RT for each item. These are methods that go to modify some of the more classic ones, for this reason it is necessary to first make a brief presentation of such criteria (only the one that can be modified) and then move on to see in more detail those employing RTs.

Classic ISC:

- Maximum information criterion (MIC). As stated in the introduction, the ISC uses the estimate of θ_n calculated in the previous steps to determine which item to administer. When the ML is chosen as estimation method, under smoothness conditions on the IRF, $\hat{\theta}_{MLE_n}$ is asymptotically distributed as $N(\theta_n^*, I^{-1}(\theta_n^*))$ where θ_n^* is the true value of the latent ability, while $I(\theta_n^*)$ is the Fisher information related to θ_n^*. Therefore, the inverse of $I(\theta_n^*)$ is the asymptotic variance of $\hat{\theta}_{MLE_n}$. So, the larger is the Fisher information, the smaller is the asymptotic variance of $\hat{\theta}_{MLE_n}$.

The function $I(\theta_n)$ is the expected value of the second derivative of $\ln L(\theta_n)$, with inverted sign.

$$I(\theta_n) = -E\left[\frac{\partial^2 \ln L(\theta_n)}{\partial \theta_n^2}\right]. \tag{2.56}$$

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

$$\begin{aligned} I(\theta_n) &= \textstyle\sum_{k=1}^{K} I_k(\theta_n), \\ I_k(\theta_n) &= \frac{[P(Y_{nk}=1 \mid \theta_n)]^2}{P(Y_{nk}=1 \mid \theta_n)(1-P(Y_{nk}=1 \mid \theta_n))}. \end{aligned} \tag{2.57}$$

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

$$k_{m+1} = \arg\max_{l} \{I_l(\hat{\theta}_{m_n}) : l \in R_m\}, \tag{2.58}$$

where:

- $m = 1, \ldots, M$ 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.
- $\hat{\theta}_{m_n}$ is the generic estimator of θ_n at step m. As this is a generic formulation, $\hat{\theta}_{m_n}$ may refer to any estimation method, not necessarily ML.
- $I_l(\hat{\theta}_{m_n})$ indicates the information provided by item l at the ability estimate at step m.

Generally, the larger $I_l(\hat{\theta}_{m_n})$ is, the smaller is the asymptotic variance of $\hat{\theta}_{m_n}$. Specifically, when using an ML estimation method, $\hat{\theta}_{MLE_n}$ will

have the smallest possible asymptotic variance. For this reason, when $\hat{\theta}_{MLE_n}$ 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 EAP in combination with the MIC (van der Linden and Glas, 2010a).

Despite its advantages, this criterion is not free from critical issues, in particular because Fisher's information is a measure of local information, the further away $\hat{\theta}_{m_n}$ is from the real value of θ_n, the lower the effectiveness of this criterion. For this reason, it tends to be less efficient with low-length tests, because the estimator has more difficulty in stabilizing the estimate. Furthermore, this criterion always tends to choose the most discriminating items. For this reason some items are much more exposed than others and MIC don't have features for satisfying test constraints.

For these reasons, and especially for the latter, another ISC that is usually proposed is the *a*-stratified with *b*-blocking item selection method (Chang et al., 2001).

- The *a*-stratified with *b*-blocking item selection method (ASB). The a-stratified (AST) with b-blocking item selection method (ASB) tries to solve the problem of information-based item selection methods that, no matter how large the item pool size is, only a small fraction of the items tend to be used. In Chang and Ying (1999) and Chang et al. (2001), has been shown that ASB can help balance the item exposure distribution, and hence yields the better test security while maintaining acceptable estimation efficiency. This method makes possible to select the most discriminating items, namely those with a higher *a*-parameter, in the most advanced stages of the test, due to the fact that high discrimination parameters are more useful in later stages of an exam than in the early stages, when there is considerable uncertainty about the ability parameter θ_n. In order to do that, the ASB performs the following steps:

- Arrange items in ascending order of difficulty and then divide the item bank into W equal-length blocks. So the first block contains the easiest items and the W-th block contains the most difficult ones.
- For each $w = 1, \ldots, W$, arrange items in ascending order of discrimination and then divide the block into J equal-length strata. In this way, for each of the W blocks, the first stratum contains the less discriminating items, and the J-th stratum contains the most discriminating ones. Thus, an item selected form the w-th block and his j_w-th stratum is easier but as discriminating than an item selected from the $(w+1)$-th block and his j_{w+1}-th stratum.
- Now, for each $j = 1, \ldots, J$, recombine the j_w-th stratum items across the W blocks into a single stratum. In this way, now each stratum contains items with very similar a parameters but with different b parameters, sorted in ascending order.
- Divide the test into J stages.
- In the j-th stage, the items considered are only those belonging to the j-th stratum and they are chosen based on the closeness of b values to the current estimate of θ_{m_n}:

$$k_{m+1} = \arg\max_{l} \left\{ \frac{1}{|\hat{\theta}_{m_n} - b_l|} : l \in R_{m_j} \right\}, \tag{2.59}$$

  where R_{m_j} contains the remaining items in the j-th stratum that have not been administered.

This method is able to balance the item exposure distribution, and hence yields the better test security while maintaining acceptable estimation efficiency. However requires, on average, much larger item banks with items that have very similar distributions of a-parameters per difficulty class.

- Shadow test approach (STA) for item selection. It's not really an ISC, but rather an approach that is used to resolve the critical issues of the selected ISC (van der Linden and Reese, 1998), that is how test specifications like content restrictions have to be taken into account during item selection. STA is very often used together with information-based item selection methods, because it is able to counterbalance the lack of control on the exposure of the items that these ISC presented, and also allows to deal with test specifications, like content specifications. A shadow test is a test that is generated by a test-assembly algorithm similar to those used for the linear pen-and-paper tests, but that is not disclosed, but rather remains latent. It is generated every time the algorithm has to choose the next item to be administered and, for CATs of fixed length M, are generated M shadow tests of length M. These tests are generated in order to respect all constraints and to contain the items already administered. For this reason, the first shadow test is equivalent to a linear pen-and-paper test just assembled, while the last shadow test is the actual adaptive test and always meets all constraints (van der Linden and Glas, 2010a).

  The procedure follows these steps:

  - *Step 1*: Initialize the ability estimator. That means that the initial ability θ_{0_n} is chosen following the initial choice rule.
  - *Step 2*: A shadow test of length M is assembled by the test-assembly algorithm following all the established constraints, and by ensuring that all items already administered are present.
  - *Step 3*: Among the unused items in the shadow test is administered the item that best satisfies the chosen ISC. For instance, in the case of opting for information-based item selection methods, the administered item would be the one with the highest information content among the unused items in the shadow test.

- *Step 4*: After the test-taker has provided his answer to the administered item, the interim ability estimate is updated following the chosen estimation method.
- *Step 5*: The test-assembly algorithm is updated to include, in the next shadow test, the administered item.
- *Step 6*: All unused items in the shadow test are returned to the pool.
- *Step 7*: Steps 2-6 are repeated until M items have been administered.

It is important to note that, because each of the shadow tests meets all the constraints, it means that also the actual adaptive test meets them all. For that reason, with this approach it is easily possible to take advantage of the information deriving from the RT, since it is enough to add to the test a constraint (van der Linden and Veldkamp, 2004).

ISC with RT:

- Maximum information per time unit criterion (MIT) (Fan et al., 2012). This is an improvement of the MIC, which in addition to maximizing the Fisher's information at the current ability estimate, also takes advantage of the auxiliary information provided by the RT. This is an improvement because often it happens that a highly informative item can be quite time-consuming, so it has less practical value compared to an equally or somewhat less informative item that requires less time to complete. Specifically, instead of maximizing raw item information $I_l(\hat{\theta}_{m_n})$, the next item is chosen based on:

$$k_{m+1} = \arg\max_{l} \left\{ \frac{I_l(\hat{\theta}_{m_n})}{E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}]} : l \in R_m \right\}, \tag{2.60}$$

where:

- RT_{nl} is the average time required to complete the item l by test-taker n.

– $\hat{\zeta}_{n_{MLE_m}}$ is the maximum likelihood estimator of the speed parameter ζ_n at the current step m for test-taker n,

$$\hat{\zeta}_{n_{MLE_m}} = \max_k L(\zeta_n) = \max_k \prod_{k=1}^{m} \frac{\phi_k}{rt_{nk}\sqrt{2\pi}} e^{-\frac{1}{2}[\phi_k(\ln rt_{nk} - (\lambda_k - \zeta_n))]^2} = \\ = \frac{\sum_{k \in R_m}\left[\phi_k^2(\lambda_k - \ln rt_{nk})\right]}{\sum_{k \in R_m} \phi_k^2}. \tag{2.61}$$

– $E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}]$ is the expected time that a test-taker n takes to complete item l, given their current MLE of working speed in Equation (2.61). Under the log-normal model for RT in Equation (2.26), treating ϕ_k and λ_k as known parameters, it can be computed as:

$$E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}] = \int_{-\infty}^{\infty} \frac{\phi_l}{\sqrt{2\pi}} e^{-\frac{1}{2}\left[\phi_l\left(\ln rt_{nl} - (\lambda_l - \hat{\zeta}_{n_{MLE_m}})\right)\right]^2} drt_{nl} = \\ = e^{\left(\lambda_k - \hat{\zeta}_{n_{MLE_m}} + \frac{1}{2\phi_k^2}\right)}. \tag{2.62}$$

Following Fan et al. (2012), MIT has the same advantages as MIC in terms of capacity to control variance, particularly when $\hat{\theta}_{m_n}$ and $\hat{\zeta}_{n_{MLE_m}}$ are close to the true θ_n and ζ_n, respectively. Compared to the MIC, it saves substantial testing time, with only a small loss of measurement precision. But this criterion also has some disadvantages, like it tends to be less efficient with small-length tests, but above all is not able to balance item exposure rates well. Furthermore, it requires fitting an RT model to explain the data. For this reason, a model miss-fitting could lead to an incorrect estimation of the parameters ϕ_k, λ_k and ζ_n, which affect the correct process of information maximization. In order to solve the first problem, improvements for ASB and STA were proposed in

the literature and are discussed later in this section. For the second disadvantage, Cheng et al. (2017) proposed a simplified version of MIT.

- A simplified version of the maximum information per time unit criterion (MIT-S). As already mentioned, good model fit is a prerequisite to using MIT criterion, and when using real data, this could lead to serious errors in identifying the best items to administer. In fact, Equation (2.62) is based on the log-normal model for response time proposed by van der Linden (2006) (Equation 2.26), but there are many other models for RT, as widely reported in Section 2.4.1. In addition, it has been found that the shapes of empirical response time distributions for items within a test and of similar types of tests can vary (Klein Entink et al., 2009, Ranger and Kuhn, 2012) and no single model may universally fit well all items in an item bank (Patton, 2015). For this reason Cheng et al. (2017) have tried to simplify the MIT criterion so as not to require fitting a response time model to the individual-level response time data, and they call this criterion MIT semplified (MIT-S). In order to do so, they modified Equation (2.60) by completely removing the information regarding ζ_n, a choice motivated by the fact that such an individualized measure does not make any difference in rank-ordering the items for item selection. In fact, they note that, in Equation (2.61), the speed estimate does not truly depend on the specific items l considered, this is because, even with a different set of items, if the test taker responds according to their real speed, then the estimate should remain almost the same. What they did is then replace the denominator in Equation (2.60) with a factor that does not include the speed estimate and that is independent form the model chosen for RT:

$$k_{m+1} = \arg\max_{l} \left\{ \frac{I_l(\hat{\theta}_{m_n})}{\overline{\ln RT_{nl}}} : l \in R_m \right\}, \tag{2.63}$$

where $\overline{\ln RT_{nl}}$ is the average of the log-transformed RT to item l.

As an average, although this factor has as its subscript the reference to the single test-taker n, in reality it is a measure that does not vary between the test-takers, but that depends only on the item l (Cheng et al., 2017). Moreover, it is effectively independent of the model chosen for the response times, even if it has different properties depending on the real RT distribution. For example, if van der Linden (2006) model holds, then this measure would be able to enclose inside itself the information about ζ_n. In fact, from Equation (2.26) follows that,

$$E(\ln RT_{nl} \mid \phi_k, \lambda_k, \zeta_n) = \lambda_k - \zeta_n, \tag{2.64}$$

and

$$E(\lambda_k - \zeta_n) = \lambda_k - \mu_\zeta, \tag{2.65}$$

where μ_ζ is the mean working speed of the considered population.

For that reason, $\overline{\ln RT_{nl}}$ is the MLE of the mean of the log-normal distribution over the entire population, and it serves as an estimator of the difference between the time-intensity parameter of an item and the group-level speed.

Cheng et al. (2017), doing a comparison study on real data, found that MIT-S saves substantial testing time, with only a small loss of measurement precision, and it seems to perform better in saving time than MIT for Rasch models. Then, MIT-S is less demanding in terms of data, pre-processing and computational resources, since real time updating of the working speed estimate is not needed. From a theoretical point of view, it is a more robust criterion to model misfit compared to MIT (however, the Authors did not explore this aspect in the comparison

study). However, Since MIT-S is a simplification that does not in any way solve the problem of exposure control, it is still not able to balance item exposure rates well, particularly for Rasch models, because it always favors highly time-saving items since all the items share the same a discrimination parameter ($a = 1$). The next criteria presented will focus on how best balance the item exposure rate, by exploiting the information from RTs.

- a-Stratified with b-blocking and time weighting criterion (ASB-TW). The first of these criteria is an improvement of the ASB criterion (Fan et al., 2012) that introduce a simple adjustment for time to the fifth step of the ASB algorithm presented previously (maintaining the others unchanged). They modified Equation (2.59) considering as denominator not only the difference in absolute value between θ_n and b, but also the expected value $E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}]$ defined in Equation (2.62),

$$k_{m+1} = \arg\max_{l} \left\{ \frac{1}{|\hat{\theta}_{m_n} - b_l||E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}]|} : l \in R_{m_j} \right\}. \tag{2.66}$$

Fan et al. (2012) conducted a simulation study that showed how, also in this case, using the information from RTs has led to saving substantial testing time, with only a small loss of measurement precision. The process of creating several strata allows ASB-TW criterion to have a much higher exposure rate control of the items compared to maximum-information criteria (MIC, MIT, MIT-S), similar to what happens with the ASB criterion.

However, The introduction of $E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}]$, as weighting factor, lead to a problem of correct item selection when we have a model misfit iden-

tification, and the need to real time update the working speed estimate, as already discussed for MIT.

- Constraints with respect to response times. The simplest way to use the information about RTs needed to answer the already submitted items, is to put that information into a constraint for total test time (van der Linden, 2008, Veldkamp, 2016),

$$t_{\text{tot}} \geq \sum_{k \in A_m} rt_{nk} + \sum_{l \in R_m} E[RT_{nl} | \hat{\zeta}_{n_{MLE_m}}] x_l, \tag{2.67}$$

where:

  - t_{tot} is the total time that test-takers have before the test ends.
  - x_l is the decision variable, denoting if an item will be in the test ($x_l = 1$) or not ($x_l = 0$).
  - A_m is the set of all items that have already been administered.

In other terms, the sum of the times rt_{nk} spent on answering the previous items and the expected times $E[RT_{nl} | \hat{\zeta}_{n_{MLE_m}}]$ on the remaining items, has to be lower than the total amount of time t_{tot} available for the test. For that reason, if an item has an expected response time greater that $(t_{\text{tot}} - \sum_{k \in A_m} rt_{nk})$ will not be taken into account by the selected criterion.

Since it is a constraint and not an ISC, it can be combined with the different criteria showed so far, and can also be used with item selection approaches, such as STA.

Regarding its advantages, the more accurate $E[RT_{nl} | \hat{\zeta}_{n_{MLE_m}}]$ is, the more sure is that the test-taker is able to finish the test. Furthermore, Since this is a constraint, it may be bundled with other constraints to compensate any lack that the selected ISC has, such as those relating to the control of the items exposure rate.

However, as already pointed out for MIT and ASB-TW, having to estimate the expected value $E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}]$, a model miss-fitting could lead to a problem of correct item selection, particularly on the early stages of CAT, when hardly any information about ζ_n is available. As a result of that, the test-taker might need more time than expected and might run into time trouble toward the end of the test.

In order to overcome this limit, Veldkamp (2016) proposed the introduction of an ISC called Robust CAT. First, however, it is worth mentioning another ISC, also proposed by Veldkamp (2016), which integrates the constraint about the maximum RT in Equation (2.67), directly within the objective function of the ISC.

- Penalized violations of maximum response time criterion. This criterion is based on the assumption that it may be acceptable to allow a small percentage of test-takers to exceed the maximum RT. This is because, although it can bring great advantages to make sure that all the examined are able to complete the whole test, it is also true that a constraint such as that on Equation (2.67) could strongly limit the choice of some items, especially in the more advanced phases of the test, going to eliminate, for some test-takers, those items that, if effectively answered, would be able to supply more information about the true value of θ_n. In order to do this, Veldkamp (2016) proposes a strategy that is based on the *goal programming* or *penalty strategy* for dealing with test specifications (Veldkamp, 1999). In this type of strategy, what is called *goal* or *target* is fixed and then a certain weight or penalty, P, is established, which is applied during the item selection process to all those items that do not respect the predetermined goal. This approach ensures that items failing to meet the objective are not automatically eliminated; however, their selection becomes considerably more challenging. In this criterion,

the goal is not to surpass the maximum response time, as defined in Equation (2.67), and and the objective function of the ISC is defined as:

$$k_{m+1} = \arg\max_l \sum_{l \in R_m} [I_l(\hat{\theta}_{m_n}) x_l - P^* \max_l \{(\sum_{k \in A_m} rt_{nk} + \\ + \sum_{l \in R_m} E[RT_{nl} | \hat{\zeta}_{n_{MLE_m}}] x_l) - rt_{\text{tot}}, 0\}]. \tag{2.68}$$

Note that if the goal in Equation (2.67) is respected, then the second member of the sum will be zero and the objective function becomes that of a classic information maximization criterion. If instead the goal is violated, that is $t_{\text{tot}} < \sum_{k \in A_m} rt_{nk} + \sum_{l \in R_m} E[RT_{nl} | \hat{\zeta}_{n_{MLE_m}}] x_l$, then the value of the information provided by that specific item l, will be penalized the more $\sum_{k \in A_m} rt_{nk} + \sum_{l \in R_m} E[RT_{nl} | \hat{\zeta}_{n_{MLE_m}}] x_l$ is greater than t_{tot} and the more P has been chosen high. Note also that, the higher P is chosen, the closer this method is to the constraints with respect to the response times method, that was previously illustrated. So, in that case, the items violating Equation (2.67) are never selected for any test-taker. Conversely, a P too small, tending to zero, would reduce Equation (2.68) to that of a MIC (Equation 2.58), losing completely the contribution that the information about RTs can give to the ISC. For these reasons, it is important to carefully choose the value to be attributed to the weight P. In general, it can be advised to determine the most appropriate value of P empirically, in a simulation study.

Finally, as also pointed out by Veldkamp (2016), this ISC can also be applied within the STA, so that, during the construction of each shadow test, each item within them was first weighed.

In addition to the advantages already mentioned for the other criteria of information maximization, this criterion is also able to manage those items that would require a higher RT than others, but without automatically eliminating them from the selection.

However, like all the other criteria already mentioned that make use of the expected value $E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}]$, this method is very vulnerable to model-misfitting, and is also important to select with great care the value of P, to avoid undermining the benefits that this criterion brings with it, thus making it very important to carry out preliminary analyses, such as a simulation study.

- Robust CAT criterion. The latest method analyzed is what Veldkamp (2016) defines as Robust CAT criterion. This is a criterion that tries to solve the problem about the uncertainty that there may be in the estimate of the expected value $E[RT_{nl}|\hat{\zeta}_{n_{MLE_m}}]$, especially in the early stages of the test, when hardly any information about speed is available. It is called *robust* because it takes the uncertainty into account, by selecting the items based on a conservative estimate of the parameters involved. This method is based on the assumption, argued by Bertsimas and Sim (2003), that uncertainty is normally distributed, and for this reason has a large impact only on the final solution for a limited number Γ of items. Starting from this assumption, Veldkamp (2013) has developed a pseudo-algorithm that allows the application of the Bertsimas and Sim (2003) within the ISC, but that does not consider the information deriving from the RTs, and subsequently modified its own pseudo-algorithm so that it could also consider such information (Veldkamp, 2016).

  After establishing a value for $\Gamma = 1,\ldots,M$, the modified pseudo-algorithm follows the following steps to decide which $(m+1)$-th item to administer:

  - Rank the item such that $I_1(\hat{\theta}_{m_n}) \geq I_2(\hat{\theta}_{m_n}) \geq \ldots \geq I_M(\hat{\theta}_{m_n})$.
  - Calculate the difference $d_k = E[RT_{nk}|\hat{\zeta}_{n_{MLE_m}}] - E[RT_{nk}|\hat{\zeta}^{robust}_{n_{MLE_m}}]$, for every $k = 1,\ldots,K$, where $E[RT_{nk}|\hat{\zeta}^{robust}_{n_{MLE_m}}]$ is a robust estimation

of the expected RT calculated taking into account also an error component due to uncertainty.

– For $w = 1, \ldots, (M-m)+1$, find the item that solves.

$$G^w = \max_l \sum_{l \in R_m} I_l(\hat{\theta}_{m_n}) x_l. \tag{2.69}$$

– At this point, among the $(M-m)+1$ items that have the largest G^w, the $(m+1)$-th item administered will be the one that maximizes G^w respecting the following constraint:

$$\begin{aligned} t_{\text{tot}} \geq \textstyle\sum_{k \in A_m} rt_{nk} + \sum_{l \in R_m} E[RT_{nl} \mid \hat{\zeta}_{MLE_m n}] x_l + \\ + \left[\textstyle\sum_{k=1}^{w} (d_k - d_w^*) x_k + \min(M-(m+1), \Gamma) d_w^* \right], \end{aligned} \tag{2.70}$$

$$k_{m+1} = \arg\max_w \{G^w : w = 1, \ldots, (M-m)+1\} \tag{2.71}$$

where $d_w^* = \min_{k \leq w}\{d_k\}$.

To summarize, the RT constraint in Equation (2.67) is corrected for uncertainty in Γ of the items by adding the term $\sum_{k=1}^{w}(d_k - d_w^*)x_k + \min(M-(m+1), \Gamma)d_w^*$. In this way the risk of a test-taker having too little time to complete the test is reduced, at the cost that a series of w integer programming problems has to be solved in the item selection step instead of just one.

In conclusion, it is worth noting that, exactly as shown in the works of the various authors already mentioned, among all these methods reported here there is not one that is objectively better than the others, but it is true that some of these might be more recommendable to others, depending on the need. For example, in the case of real data that do not fully reflect the log-normal distribution of response times, can be used the MIT-S criterion because of its

property of robustness against the model-misfitting, otherwise can be used constraints with respect to response times when we want to be more confident that the subjects can finish the test entirely in time. Furthermore, regardless of the choice of ISC, this section has shown how RT can be effectively used to define valid item selection methods.

More generally, the main components of a CAT have been presented in this Section, emphasizing how this type of CBT can be complex, yet, at the same time, they are capable of adapting to the requirements depending on how they are configured. In this regard, a comparative study between CATs with different configurations is presented below to better understand their differences. This study was conducted using real data, specifically the same data from INVALSI used in the previous Section 2.4.4. Although this Section and, more broadly, this Chapter have extensively discussed RT and its potential implementation in CATs, the example study to be presented will not use this auxiliary source of information. Instead, it will be in Chapter 3 that RT and CAT will be combined to address a persistent issue in testing, both computerized and pen-and-paper, that of cheaters.

2.5.2 CATs comparison with real data

The aim of this analysis is to compare CATs with different configuration via a simulation study using the 2018 mathematics INVALSI data for grade 10, already used for the application in Section 2.4.4. For this grade, 12 parallel test forms were created by INVALSI from a Rasch item bank through Automated Test Assembly (ATA) methods. These forms are fixed, with 35 items each, linear and follow specific content constraints. In the first step of this analysis, a fixed-length CAT and a variable-length CAT are compared. Subsequently, the performance of the CAT is evaluated by applying the same constraints used by INVALSI to generate their linear forms, and to do that the STA was used. The main objective of the simulation is to explore how the

different configuration will diverge in terms of accuracy of the estimates and efficiency, using as performance indexes the BIAS and the MSE of the ability and the correlation between the true and the estimated ability. Furthermore, alterations in the test length have been investigated.

The item bank contains 143 mathematics items, calibrated according to the Rasch model.

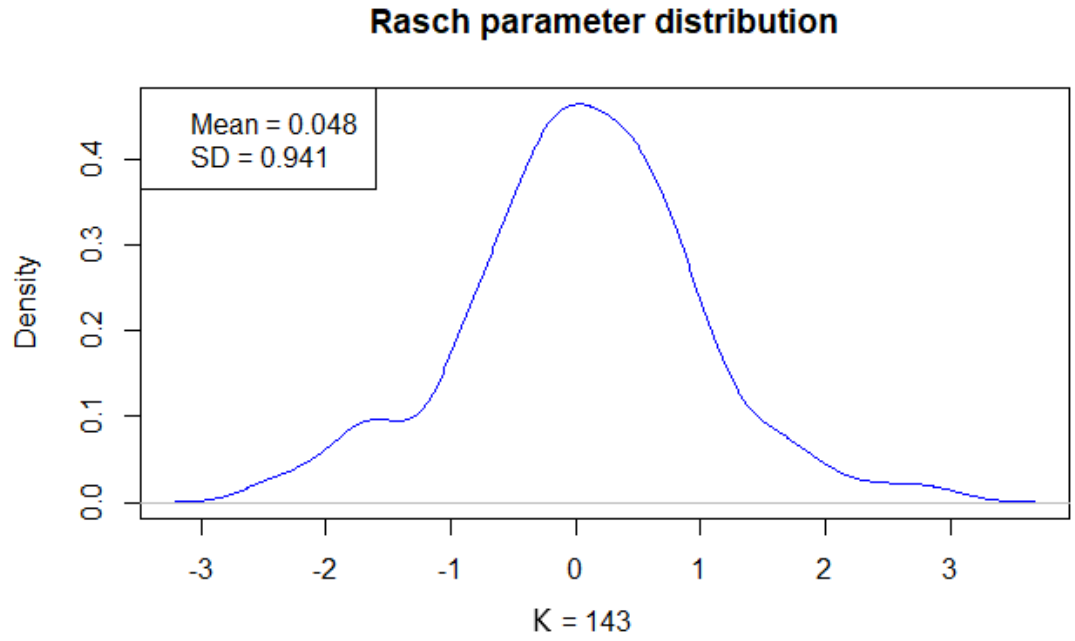

Fig. 2.3 Item difficulty parameter distribution (K=143). CATs comparison with real data. INVALSI data.

These items also have content characteristics (*item type*, *domain*, *dimension*) and INVALSI has used all this information to assemble, in an automated way, 12 tests of 35 items each, as homogeneous as possible.

The simulation was then carried out in this way:

- θ_n abilities were randomly simulated for a different number of respondents (N=100, N=500 and N=1000), from a standard normal distribution.
- Through the use of the R package *catIrt* (Nydick, 2014), the simulated θ_n were used to simulate answers to two different types of CAT. The

first was a fixed length CAT, with a number of items equal to that of the 12 linear forms ($n = 35$), while the second was a variable length CAT, which continued to administer items until the $\hat{\theta}_n$ variance was less than 0.16 (with a minimum test length of 24 items). In both cases, the first 5 items administered were randomly selected and the initial ability estimate was assumed to be 0 for each test-taker. In addition, the interim ability estimate was calculated using the MLE $\hat{\theta}_{n_{MLE_m}}$ and the items to be administered after the fifth were chosen following the MIC criterion (Equation 2.58). At the end of the simulated tests, the final abilities of each test-taker were calculated, also in this case using the MLE.

- Later, using the same simulated θ_n but a different R package called *ShadowCAT* (Kroeze, 2017) were simulated answers to a variable length CAT that met the same content constraints used by INVALSI to generate the 12 linear tests. In order to implement all these constraints at the same time, the STA was used, integrated into the *ShadowCAT* package. In order to be able to make the comparison, the simulation setup were the same of the variable test length.

- Having both, the estimated and the real value of the abilities, for each model and for each population number (100, 500 and 1000), the BIAS and the MSE of the ability and the correlation between the real and the estimated ability have been calculated. Those indexes were used as performance indicators.The average test length was also calculated. In addition, other indicators, such as the minimum, maximum and average exposure rate for each item (i.e, the percentage that indicates in how many tests that item has been selected) have also been calculated. These simulations were repeated 100 times (with 100 different seeds) and then means of the indexes obtained were made. As results obtained, these means were reported.

The results of this simulation analysis are summarised in Table 2.5. In general, it is possible to say that an increase in population does not involve an impactful change in the values of the indexes, for all the methods. This is because ability estimates are performed individually for each test-taker, so since they are not a joint estimate, but marginal, the responses of the rest of the population do not affect the ability estimation of the individual respondent. Clearly, an increase in population still has a stabilizing effect on the average.

TYPE	N = 100				N = 500				N = 1000			
OF TEST	**BIAS**	**MSE**	**COR.**	**LENGTH**	**BIAS**	**MSE**	**COR.**	**LENGTH**	**BIAS**	**MSE**	**COR.**	**LENGTH**
CAT FL	0.002	0.142	0.935	35.000	0.000	0.145	0.934	35.000	0.000	0.147	0.934	35.000
CAT VL	0.007	0.158	0.927	29.068	0.000	0.156	0.929	29.172	0.000	0.158	0.928	29.211
CAT C	0.022	0.239	0.903	28.582	0.011	0.214	0.904	28.575	0.009	0.207	0.906	28.579

CAT FL = CAT fixed-length; CAT VL = CAT variable-length; CAT C = CAT with constraints.

Table 2.5 Performance indexes of 3 different types of CAT. INVALSI data.

As is well known in the literature, it is possible to note that there is a trade-off between the BIAS and the MSE. For this reason, to make easier to compare the performance of the models, the value of the correlation between the true and the estimated ability has also been reported. In summary, it is possible to see that the variable-length CAT model, as could be expected, has slightly lower performance, but on average saves 6 items per test. The CAT model with constraints is the one that has the worst performance in estimation accuracy, but it is at the same time the one that has the smallest length tests and that always ensures that each test respects the content constraints.

In fact, unrestricted CATs are not independently able to maintain all content constraints. To understand how effectively it can influence the performance to be able to maintain all the constraints at the same time, was also carried out a simulation (N=100) in which only one constraint at a time is considered. As can be seen from the Table 2.6, relaxing the constraints involves an effective improvement in both performance and length.

TYPE OF TEST	BIAS	MSE	CORRELATION	LENGTH
CAT 1 constraint	**0.0178**	**0.207**	**0.917**	**28.428**
CAT all constraint	0.022	0.239	0.903	28,582

Table 2.6 Comparison between CAT with 1 constraint and CAT with all constraints (N=100) . INVALSI data.

Further simulations were then carried out, and the most interesting results are those concerning the variable-length CAT but with a lower maximum allowable variance (0.11 instead of 0.16) (Table 2.7). In fact, decreasing the maximum variance allowed, the variable-length CAT improves its performance, going to slightly exceed those of the fixed-lenght CAT, but at the expense of the test length, which increases by an average of 7 items.

TYPE OF TEST	BIAS	MSE	CORRELATION	LENGTH
CAT FL	**0.002**	0.142	0.935	**35.000**
CAT VL	0.005	**0.106**	**0.949**	42.630

Table 2.7 Comparison between CAT fixed-length and CAT variable-length (max variance 0.11; N = 100) . INVALSI data.

This is in line with expectations, as a decrease in the target value forces the CAT to continue administering items to further reduce the variance of the estimates, thereby increasing their precision.

The last analysis carried out concerns the monitoring of the item exposure rate. The results found show that the exposure rate distributions do not differ much between the different methods (Table 2.8), however these are not optimal distributions, since some items have an exposure rate greater than 45%. A possible development, not performed in this analysis, might be to analyze the performances of a CAT that also considers the control of the exposure rate distribution (Sympson and Hetter, 1985).

TYPE OF TEST	MEAN	MIN	MAX
CAT FL	24%	11%	40%
CAT VL	24%	8%	42%
CAT C	24%	11%	47%

Table 2.8 Exposure rate indexes (N=100). CATs comparison with real data. INVALSI data.

To summarize, it can be said that CAT tends to be adaptable to the test giver's needs. The choice of a fixed-length approach is optimal when there is the need to ensure that all test-takers are administered the same number of items, while variable-length CAT allows for improving estimation performance, albeit at the risk of having some tests longer than others. Furthermore, CAT has also been able to manage all the content constraints in real-time, demonstrating adaptability in this regard as well, with modulable performance depending on the quantity and type of constraints.

In the next Chapter the RT and the CAT will be used together to address the issue of test-takers who cheat during computerized test.

Chapter 3

Using RT to identify cheaters in CAT

3.1 Cheaters in CAT

Test cheating is a widespread issue in various settings, spanning from elementary school assessment programs to higher education and professional certification programs that grant specialized licenses or credentials in specific fields. Obtaining precise statistics on the extent of cheating is challenging, but it appears that cheating remains a significant problem. For example, a review of studies involving American college students conducted by Whitley (1998) found that, on average, 43% of college students reported cheating on exams. But cheating is a widespread and ongoing issue worldwide. For example, during the COVID-19 pandemic, as reported by a systematic review conducted by Newton and Essex (2023), there was a significant rise in the number of tests conducted online through computers, leading to an increase in cheating cases (from 29.9% pre COVID to 54.7% during COVID) because "*there was an opportunity to do so*".

Estimating the prevalence of cheating on licensure and certification exams is more complex, but given the high stakes involved in these tests, it is reasonable to assume that similar or even higher cheating rates may be observed. According to Wollack and Fremer (2013), "*With tests now acting as*

gatekeepers to numerous professions, the temptation to cheat is at an all-time high".

There are likely multiple valuable interpretations of the concept of cheating. In the context of examinations, Cizek (2012) has provided a definition of cheating as "*Any action undertaken before, during, or after a test or assignment that aims to gain an unfair advantage or produce inaccurate results*".

Firstly, cheating may occur at nearly any stage of the test development, administration, and grading process. Test-takers might try to inappropriately access test content even before the actual exam, such as by participating in unauthorized test preparation courses or sharing secure test material with others. They may also attempt to obtain test items through electronic hacking, theft of paper test booklets, or similar means. Test candidates could arrange for someone else, ostensibly more capable, to take the test on their behalf, leading to fraudulent results. Test-takers may also seek to gain information from other participants, engage in copying or collusion, introduce forbidden materials into the testing session, or gather information during scheduled breaks or other moments throughout the test. Some times are even the teachers that try to cheat, improving answers or marking incorrect answers as correct ones. As these examples illustrate, there are numerous opportunities for cheating.

Secondly, cheating is deliberate; it is done with the intention of achieving a test score that doesn't accurately reflect the test-taker's true knowledge or ability.

In fact, the essence of educational tests is to gather as much information as possible, in the form of answers to questions and queries. This information can then be used to estimate a summary value, such as a score, of the real level of ability of the test-takers. Without delving too deeply into the specifics of how such tests are devised and constructed, and returning to the premises of

Chapter 2 of focusing solely on models (IRT) and tests that capture one ability at a time, what these test have in common is the goal of intercepting the real test-taker's ability as effectively as possible. However, whenever cheating occurs, in any of its form, the resulting test scores are unlikely to provide an accurate measurement of the test-taker's true level of knowledge, skill, or ability. In essence, beyond the moral and ethical implications, concern about cheating can be seen as a psychometric issue related to the validity or interpretive accuracy of test scores.

What has been said so far is applicable to any type of educational test, be it pen-and-paper or computerized. To be more specific, even oral tests are not exempt from cheating practices, but the focus of all this work has always been on written tests. In particular, great attention has been paid to CAT, and indeed, the topic of cheating and methods to deal with it will be analyzed in more detail for this type of computerized test.

In fact, the implementation of CAT programs brought forth new challenges. High-stakes testing programs faced item security issues, as test-takers had a great ability to memorize and share test items. Indeed, there are a number of websites known as *brain-dump sites* that gather all stolen test items and can be accessed by malicious users. For example, in 2002, there was a massive use of information from these 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 testsCizek and Wollack (2016)

Furthermore, the great need to calibrate items to create efficient item banks always carries the risk of overexposing those items that will later be used to construct actual tests.

But CAT has also brought solutions to deal with cheating. Given its adaptive nature, various practices such as copying from nearby test-takers or getting information from them are much more challenging to implement,

because each test-taker is likely to have a different test. For this reason, this thesis focuses on a specific type of cheating, where test-takers gain advanced knowledge of correct answers, referred to as *pre-knowledge*, and use it to answer test questions accurately, regardless of their actual ability. Such items are termed *compromised items*. Item compromise occurs when an item performance changes over time due to its content being distributed beyond its valid usage boundaries (Zara and Pearson, 2006). So, item compromise and the use of pre-knowledge are closely linked, and often analyzing one means analyzing the other.

Another reason why this work focuses only on this specific type of cheating is that the type of test under consideration is multiple-choice tests. The items that make up this type of test are among the easiest to memorize and perhaps also among the easiest to share. Therefore, in multiple-choice CATs it is expected that the use of item pre-knowledge is a prevalent form of cheating.

The urgent need to detect cheating, has resulted in the development and implementation of quantitative methods for detecting pre-knowledge cheatings. In Section 3.2, the main methods used in the literature will be briefly described.

3.2 Solutions in the literature

Following the work of Cizek and Wollack (2016), the methods to identify cheaters in the case of item pre-knowledge can be organized into four categories:

- Methods to identify cheaters at the individual level.
- Methods to identify items that may have been compromised.
- Methods to identify both cheaters and items that may have been compromised.

- Methods to identify groups of cheaters.

As already mentioned in the Introduction and as will be discussed in the Section 3.3, this work aims to analyze and find a solution to the problem of cheaters during the test administration. For this reason, the last three categories will not be explored in depth, as they encompass methods that, by their very nature, propose post-test interventions, which use information collected at different time points (for the identification of compromised items) and aggregated information (for the identification of groups of cheaters) that are not available while the test is still ongoing.

In summary, the second category includes methods that seek to understand if a group of items has been compromised by looking for significant differences over time in the number of test-takers who correctly answer those items. An increase in this number could indicate potential item compromise. These methods include the Simpson-Hatter (SH) method for controlling item exposure (Sympson and Hetter, 1985), the Moving Averages method (Han, 2003) and the Log Odds Ratio Statistic method (Obregon, 2013). In the third category, there are methods to detect groups of test-takers whose answers exhibit unusual similarity. Some of these methods are the Detection of Collusion Using Kullback-Leibler Divergence (Belov, 2012, 2013, 2016), the Detection of Collusion Using Cluster Analysis (Wollack and Maynes, 2016), and the Detection of Collusion Using Factor Analysis (Zhang et al., 2011). In the last category, the methods aim to simultaneously identify both compromised items and cheaters. Some of these methods include the FLOR Log Odds Ratio Index (McLeod, 2006) and the joint using of Differential Person Functioning and Differential Item Functioning (O'Leary and Smith, 2016).

As for the first category, regarding the methods to identify cheaters with pre-knowledge, some of the methods proposed in the literature are:

- The Deterministic Gated Item Response Theory Model (DGM) (Shu et al., 2013). This is a method used when possible compromised items have already been identified. After checking the level of exposure of the items, they are labeled as secure ($I_k = 0$) or compromised ($I_k = 1$). The idea behind this method is that the probability of a correct response to an item k depends not only on the true ability θ_n of a test-taker and the psychometric parameters of the item ($\boldsymbol{\psi}_k = (a_k, b_k, c_k)$) but also on whether the item is compromised ($I_K = 1$) and the cheating ability of the test-taker θ_{n_C} (which is estimated based on the test-taker's performance on the items specified as compromised), as follows

$$\begin{aligned} P(Y_{nk} = 1|\theta_n, \theta_{n_C}, \boldsymbol{\Psi}_k, T_n, I_k) = \\ = P(Y_{nk} = 1|\theta_n, \boldsymbol{\Psi}_k)^{1-T_n} \times \\ \times \left[(1-I_k)P(Y_{nk} = 1|\theta_n, \boldsymbol{\Psi}_k) + I_k P(Y_{nk} = 1|\theta_{n_C}, \boldsymbol{\Psi}_k)\right]^{T_n}, \end{aligned} \tag{3.1}$$

where $P(Y_{nk} = 1|\theta_{n_C}, \boldsymbol{\Psi}_k)$ is the probability of answering correctly to a compromised item using the cheating ability θ_{n_C} (this probability depends on the chosen IRT model) and $T_n = 1$ when $\theta_n < \theta_{n_C}$.

In Equation (3.1), if an item is considered secure ($I_k = 0$), both test-takers with pre-knowledge ($T_n = 1$) and those without pre-knowledge ($T_n = 0$) provide answers based on their true abilities θ_n. However, when an item is compromised ($I_k = 1$), the responses of test-takers who did not have pre-knowledge of the item ($T_n = 0$) are still determined by their true abilities, while the answers of test-takers who had pre-knowledge of the item ($T_n = 1$) are based on their cheating abilities θ_{n_C}.

Using MCMC estimation, the model assigns each test-taker to one of two latent classes in each iteration. One class is for those who perform better on the items specified as compromised by the user ($T_n = 1$), and the other is for test-takers who perform equally well or better on the

items specified as secure ($T_n = 0$). Over all iterations following the burn-in phase of the MCMC run, the proportion of posterior samples in which test-taker n is assigned to the pre-knowledge class is denoted as T_n^*. Test-takers with T_n^* values exceeding a user-defined threshold (e.g., 0.95) are classified as individuals with pre-knowledge, while those with T_n^* values below the threshold are categorized as test-takers without pre-knowledge. By adjusting the user-specified threshold (e.g., ranging from 0.95 to 0.99), the model allows for a trade-off between false positives and false negatives.

Clearly, this method requires knowing if and which items have been compromised, and its effectiveness relies on the accuracy of this information. As highlighted by Cizek and Wollack (2016), this exposes the method to classification and estimation errors. Furthermore, as pointed out by Eckerly et al. (2015) and Shu et al. (2013), even when there is no mis-specification of the set of compromised items, the method is subject to two sources of bias in estimating the difficulty b_k of the compromised items.

The first source of bias arises from the fact that the DGM, in estimating the difficulty of compromised items, removes all answers provided by those flagged as cheaters. This results in an upward bias in the items difficulty estimates, leading to an increase in the estimates of both the true abilities θ_n and the cheating abilities θ_{n_C}.

The second source of bias is due to the fact that the true ability θ_n is always considered to be different from the one used for responding, as there is the possibility that the response was aided by the cheating ability θ_{n_C} (*scale drift*). This results in a down bias in the estimates of the item difficulty. The magnitude of the bias depends on the percentage of test-takers with pre-knowledge.

- Scale Purified Deterministic Gated Item Response Theory Model (scale purified DGM). To overcome the limitation of the DGM, Eckerly et al. (2015) proposed a modification of the DGM, to purify the person and item parameter estimates. This modification involves an iterative scale purification procedure, which includes the following steps:
  - *Step 1*: Estimate item difficulty parameters using the Rasch model for all response data.
  - *Step 2*: Instead of estimating item difficulty parameters in the DGM, fix them to the parameter estimates obtained in Step 1, and then run the DGM.
  - *Step 3*: Remove the response data of the flagged test-takers and re-estimate item difficulty parameters using the Rasch model.
  - *Step 4*: Run the DGM again with the response data of all test-takers, using the purified item difficulty estimates from Step 3.

  By fixing the item difficulties of each iteration and subsequently re-estimating them, the scale purification procedure eliminates item difficulty estimation bias caused by the omission of response data from honest examinees and minimizes bias due to scale drift. Eckerly et al. (2015) conducted simulations to compare the scale-purified DGM to the original DGM. They found that false positive rates (rates of cheaters incorrectly classified) significantly decreased and true detection rates (rates of cheaters and non cheaters correctly classified) increased when using the scale-purified DGM, especially in cases with a high base rate of test-takers benefiting from item pre-knowledge.

- Person Fit Statistic for response patterns. Marianti et al. (2014) and Fox and Marianti (2017), following the work of Levine and Rubin (1979) and Drasgow et al. (1985), starting from the log-likelihood of a response pattern, have defined an index called Person Fit Statistic for response

pattern, l_n^y, which, as the name suggests, is a statistic that pertains to each test-taker and can be used to determine whether the test-taker's response pattern aligns with the general response behaviour. In concept, it is not very different from the T_n statistic in DGM, but it only uses information related to the response pattern.

The person fit statistic is defined as

$$\begin{aligned} l_n^y &= -\ln L(\theta_n|\overrightarrow{Y_n}) = \\ &= -\sum_{k=1}^{K} \left[Y_{nk} \ln \left[P(Y_{nk}) \right] + (1 - Y_{nk}) \ln \left(1 - P(Y_{nk}) \right) \right], \end{aligned} \tag{3.2}$$

where

$$P(Y_{nk}) = P(Y_{nk} = 1 \mid \theta_n, \boldsymbol{\Psi}_k). \tag{3.3}$$

In Equation (3.2), the person-fit statistic l_n^y is used to assess the fit of an individual test-taker's answers to a set of items. This statistic is constructed based on the sum of logarithms of the probabilities of correct responses for each item. When a test-taker with low ability θ_n begins to answer difficult items correctly, because they have low probability to do that, the sum of logarithms of these low probabilities will be a small negative value. However, when this value is inverted, it becomes a very large positive value of the person-fit statistic.

The same goes for those who have a very high θ_n and start to make mistakes even on very simple items.

In light of this behaviour, it is appropriate to define a limit value known as the *extreme value*, denoted by the letter C. This value is set such that any person-fit statistic greater than C corresponds to an aberrant pattern, indicating a significant misfit between the test-taker's answers and the expected answers based on their estimated ability.

To find this extreme value C, it is advisable to standardize the person-fit statistic. By standardizing the statistic, it can be compared to a standard normal distribution, allowing for the definition of the limit value C based on defined significance level α.

The standardized statistic is

$$l_{s_n}^y = \frac{l_n^y - E(l_n^y)}{\sqrt{\text{Var}(l_n^y)}}, \tag{3.4}$$

where:

$$\begin{aligned} E(l_n^y) = -\sum_{k=1}^{K} [P(Y_{nk} = 1 \,|\, \theta_n, \mathbf{\Psi}_k) \ln [P(Y_{nk} = 1 \,|\, \theta_n, \mathbf{\Psi}_k)] + \\ + (1 - P(Y_{nk} = 1 \,|\, \theta_n, \mathbf{\Psi}_k)) \ln (1 - P(Y_{nk} = 1 \,|\, \theta_n, \mathbf{\Psi}_k))], \end{aligned} \tag{3.5}$$

and

$$\begin{aligned} \text{Var}(l_n^y) = \sum_{k=1}^{K} [P(Y_{nk} = 1 \,|\, \theta_n, \mathbf{\Psi}_k)(1 - P(Y_{nk} = 1 \,|\, \theta_n, a\mathbf{\Psi}_k)) \times \\ \times \left(\ln \left(\frac{P(Y_{nk}=1 \,|\, \theta_n, \mathbf{\Psi}_k)}{1-P(Y_{nk}=1 \,|\, \theta_n, \mathbf{\Psi}_k)} \right) \right)^2]. \end{aligned} \tag{3.6}$$

Given a specific significance level α, the corresponding threshold value C of the normal distribution is calculated. If $l_{s_n}^y$ in Equation (3.4) is greater than C, then the test-taker is classified in the group of individuals with pre-knowledge, otherwise is classified in the group of test-takers who have no pre-knowledge. This classification is expressed by a dichotomous variable F_n^y, which takes the value of 1 in the first case and 0 in the second.

$$F_n^y = \begin{cases} 1 & \text{if } P\left(l_{s_n}^y(\theta_n, \boldsymbol{\Psi}) > C\right) \\ 0 & \text{if } P\left(l_{s_n}^y(\theta_n, \boldsymbol{\Psi}) \leq C\right) \end{cases} \tag{3.7}$$

Using MCMC estimation, the method creates the dichotomous variable F_n^y for each test-taker $n = 1, \ldots, N$ at each iteration. Then, as the DGM, over all iterations following the burn-in phase of the MCMC run, the proportion of posterior samples in which test-taker n is assigned to the pre-knowledge class is denoted as $F_n^{y^*}$. Test-takers with $F_n^{y^*}$ values exceeding a user-defined threshold (e.g., 0.95) are classified as individuals with pre-knowledge, while those with $F_n^{y^*}$ values below the threshold are categorized as test-takers without pre-knowledge. Also in this case, by adjusting the user-specified threshold (e.g., ranging from 0.95 to 0.99), the model allows for a trade-off between false positives and false negatives.

As previously mentioned and as explicitly expressed in Equations (3.2, 3.4), l_n^y and $l_{s_n}^y$ are defined using only the information from the response pattern. This may be seen by some as a limitation of the method, because it fails to leverage potential auxiliary information that can be obtained during the test or that was available previously (for example, the DGM uses information regarding possible compromised items). A resolution to this limitation is proposed by Marianti et al. (2014) and Fox and Marianti (2017), who suggest both a method to identify potential cheaters using only the information from RTs and a method that combines both response pattern and RT information.

- Person-Fit Statistic for RTs (Marianti et al., 2014, Fox and Marianti, 2017). The idea behind the development of this statistic is the same as behind the development of the $l_{s_n}^y$ statistic. In this case as well, the

method starts with a log-likelihood function to eventually arrive at a standardized statistic with a known distribution. However, in this case, the method works with the log-likelihood of RTs. The distribution underlying RTs is the one defined by Fox and Marianti (2016) and reported in Section (2.4.3) of this work, for which the probability formula is provided in Equation (2.45) and with a density function equal to:

$$f(rt_{nk}, \zeta_n, \phi_k^*, \lambda_k, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2 \ln rt_{nk}}} \exp\left[-\frac{1}{2\sigma_k^2}\phi_k^* \left(\ln rt_{nk} - (\lambda_k - \zeta_n)\right)^2\right]. \quad (3.8)$$

The non-standardized statistic is defined as minus two times the log-likelihood of RTs:

$$\begin{aligned} -2\ln L(\zeta_n|\overrightarrow{RT_n}) &= \sum_{k=1}^{K}\left[\left(\frac{\ln RT_{nk}-\mu_{nk}}{\sigma_k}\right)^2 + \ln(2\pi\sigma_k^2)\right] = \\ &= \sum_{k=1}^{K}\left[Z_{nk}^2 + \ln(2\pi\sigma_k^2)\right], \end{aligned} \quad (3.9)$$

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 a χ_K^2, with K degrees of freedom, where K is the number of items. So, the likelihood-based person-fit statistic for RTs, l_n^t, is defined as:

$$l_n^t = \sum_{k=1}^{K} Z_{nk}^2 = \sum_{k=1}^{K}\left(\frac{\ln RT_{nk} - \mu_{nk}}{\sigma_k}\right)^2, \quad (3.10)$$

where

$$\mu_{nk} = \lambda_k - \phi_k^* \zeta_n. \tag{3.11}$$

As for l_n^y, an unusually large statistic values indicate a misfit. In fact, a test-taker that having a larger (smaller) RT than the average for their speed (μ_{nk}), will have a large and positive (negative) $(\ln RT_{nk} - \mu_{nk})$. Once raised to the square, it will give a positive much larger value than those who had a RT close to the average for their speed. This time, unlike what happened with response patterns, it is not necessary to standardize l_n^t, because it is already chi-squared distributed. Therefore, it is possible to know the extreme value C associated with a certain significance level α, directly from the χ_K^2 distribution. Also in this case, if l_n^t in Equation (3.10) is greater than C, then the test-taker is classified in the group of individuals with pre-knowledge, otherwise is classified in the group of test-takers who did not have pre-knowledge. 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 follows

$$F_n^t = \begin{cases} 1 & \text{if } P\left(l_n^t(\zeta_n, \boldsymbol{\lambda}, \boldsymbol{\phi}^*, \boldsymbol{\sigma^2}) > C\right) \\ 0 & \text{if } P\left(l_n^t(\zeta_n, \boldsymbol{\lambda}, \boldsymbol{\phi}^*, \boldsymbol{\sigma^2}) \leq C\right) \end{cases}. \tag{3.12}$$

Finally, as DGM and the person fit statistic for response patterns, using MCMC estimation, a 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 pre-knowledge class is denoted as $F_n^{t^*}$, and is used as an estimate of the posterior probability of aberrant RT pattern.

This statistic can be used either independently or jointly with the previous statistic to identify cheaters by utilizing information from both response patterns and RTs.

- Person fit statistic for joint response patterns and RTs. First, the two statistics, $l^y_{s_n}$ and l^t_n , are calculated separately following the formulas in Equation (3.4) and Equation (3.10), respectively. Then, two separate dichotomous variables, F^y_n and F^t_n, are defined. Only at this point a new dichotomous variable, $F^{(y,\,t)}_n$, is defined, which will have a value of 1 when both F^Y_n and F^t_n are equal to 1, and 0 otherwise as follows

$$F^{(y,\,t)}_n = \begin{cases} 1 & \text{if } P\left(l^y_{s_n}(\theta_n, \boldsymbol{\Psi}) > C,\ l^t_n(\zeta_n, \boldsymbol{\lambda}, \boldsymbol{\phi}^*, \boldsymbol{\sigma^2}) > C\right) \\ 0 & \text{if } 1 - P\left(l^y_{s_n}(\theta_n, \boldsymbol{\Psi}) > C,\ l^t_n(\zeta_n, \boldsymbol{\lambda}, \boldsymbol{\phi}^*, \boldsymbol{\sigma^2}) > C\right) \end{cases}. \tag{3.13}$$

In other words, a test-taker will be classified in the group of individuals with pre-knowledge, only if they are classified in this group by both statistics. In all other cases, they will be classified in the group of test-takers who did not have pre-knowledge.

Finally, also in this case, the status of $F^{(y,\,t)}_n$ can be computed at each MCMC iteration. The average over MCMC iterations is used as an estimate of the posterior probability of jointly aberrant pattern for answers and RT.

In summary, all the methods shown for identifying test-takers with pre-knowledge share a common approach. They aim to identify specific behaviours of test-takers and summarize them in a statistic that quantifies, through a numerical value, how much this behaviour deviates from the expected. However, another common element among these approaches is their reliance on information that is only available at the end of the test (such as

ability or speed estimates). Therefore, as they have been presented, these methodologies cannot be entirely used to address the problem of cheating while the adaptive test is still ongoing.

The next Section (3.3), starts facing this limitation and present a proposal to overcome it.

3.3 New proposal

In this section, a method is presented 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 peculiarity 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, as mentioned in Section (3.1), to compromise the validity of the test scores of those individuals. For this reason, the method proposed here is divided into two main parts.

The first step concerns the identification of a statistic capable of identifying cheaters with pre-knowledge directly during the test, using the *partial* information that the adaptive algorithm is collecting and analyzing it in real-time.

The second step involves developing a method that can fully use this statistic to *rebalance* the test for those suspected of cheating, without affecting those who are not engaged in misconduct.

The aim of the method is to detect aberrant behaviours of the test-takers, so it tends to focus more on test-takers' behaviours, rather than on identifying compromised items.

In relation to the first objective, the idea is to start with a statistic already present in the literature and find a way to use it by leveraging the partial information that the algorithm is collecting in real time. The new statistic

should be calculable in real-time during the test administration, and it should be updated as the test-taker give their answers to the new administered items.

Among the methods proposed in the literature, since the aim is to make the most of the potential that the CAT can offer, the focus was placed on the person fit statistic for the RTs (Marianti et al., 2014, Fox and Marianti, 2017), described in Section (3.2) and reiterated here (Equation 3.14). Since what is about to be proposed is an experimental method, the decision was made to use only the information related to the RTs. This means not adopting the joint approach with the response patterns (Section 3.2).

Therefore, the person fit statistic to be used in cheating detection is

$$l_n^t = \sum_{k=1}^{K} \left(\frac{\ln RT_{nk} - \mu_{nk}}{\sigma_k} \right)^2 \sim \chi^2_{\mathbf{K}}, \tag{3.14}$$

where

$$\mu_{nk} = \lambda_k - \phi_k^* \zeta_n. \tag{3.15}$$

As already mentioned, the statistic l_n^t is defined as a sum of differences that includes all the test items K. These differences are calculated with respect to the estimate of the response speed, that is derived at the end of the test using the Gibbs sampling algorithm (Fox et al., 2021). However, implementing this methodology during the test is challenging due to its time-consuming nature, thereby compromising one of the primary advantages of CAT, which relies on the nearly instantaneous selection of the next item.

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). 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:

$$\begin{gathered} E[RT_{nk}|\hat{\zeta}_{n_{MLE_m}}] = \exp\left(\lambda_k - \hat{\zeta}_{MLE_m n} + \frac{1}{2\phi_k^2}\right), \\ \sigma_k = \frac{1}{\phi_k}, \end{gathered} \tag{3.16}$$

where $\hat{\zeta}_{MLE_m n}$ is the maximum-likelihood estimator for the person speed (Fan et al., 2012),

$$\hat{\zeta}_{n_{MLE_m}} = \max_k L(\zeta_n) = \frac{\sum_{k \in R_m}[\phi_k^2(\lambda_k - \ln rt_{nk})]}{\sum_{k \in R_m}[\phi_k^2]}. \tag{3.17}$$

As can be seen from Equation (3.17), the value of $\hat{\zeta}_{n_{MLE_m}}$ 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 maximum-likelihood estimator in real-time during the test. Once the value of $\hat{\zeta}_{MLE_m n}$ is calculated, obtaining the expected value $E[RT_{nk}|\hat{\zeta}_{n_{MLE_m}}]$ is straightforward (Equation 3.16).

Therefore, by replacing μ_{nk} and σ_k with $E[RT_{nk}|\hat{\zeta}_{n_{MLE_m}}]$ and $\frac{1}{\phi_k}$, respectively, a new statistic was defined that was called the *interim person fit statistic* (IPS):

$$l^t_{n_m} = \sum_{k=1}^{m} \left(\frac{\ln RT_{nk} - \ln E[RT_{nk}|\hat{\zeta}_{n_{MLE_m}}]}{\frac{1}{\phi_k}} \right)^2 \sim \chi^2_m. \tag{3.18}$$

The statistic in Equation (3.18) follows a χ^2 distribution with m degrees of freedom, where m indicates the current iterative step, as it represents the standardized error of normally distributed logarithms of RT. The validity of

this property has been tested and confirmed in Chapter (4), through Q-Q plot analysis and the Kolmogorov-Smirnov test.

Thanks to this property, if a significance level α is chosen, the threshold value C_m can be easily found from the χ^2_m distribution. If the threshold C_m is exceeded, the test taker is identified as a cheater. This is because the rejection region of the null hypothesis is reached, where the null hypothesis is that RT follows a log-normal distribution. Henceforth, individuals with an IPS lower than C_m, falling within the non-rejection region of the null hypothesis, will be referred to as *honest respondent*, while the remaining individuals, as previously mentioned, will be classified as cheaters.

From now on, all those test-takers who have not (or are presumed not to have) used item pre-knowledge will be referred to as honest respondents, while those who have used item pre-knowledge will generally be called 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 behaviour of cheaters. At this point, it is important to emphasize that the underlying 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 had undermined, and its fairness. So, 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 (STA) (Section 2.5.1), will start to administer the next items from a *more secure item database*, in order to reduce the probability that the cheater has pre-knowledge on those items. In fact, the more secure item database is an item bank that has a very low exposure rate and will be more frequently updated.

The use of STA allows for the introduction of constraints for the Item Selection Criterion (ISC) at each step m. In this case, the constraint pertains

to the choice of the database from which all subsequent items that will compose the shadow test m are selected. If $l^t_{n_m}$ is lower than C_m, then the subsequent items will be selected from the main database, just as in the absence of constraints. However, when $l^t_{n_m}$ is greater than C_m, indicating that the respondent is flagged as a cheater, then all subsequent items that will compose the shadow test m will be selected from the more secure database. It is important to note that since the statistic can be calculated and updated at each step m, the shadow test $m+1$ may differ from the previous one because of the respondent's answers and the updated $l^t_{n_{m+1}}$. In fact, if at step m, $l^t_{n_m}$ did not have enough strength to reject the null hypothesis, it is possible that at step $m+1$, the situation has changed, $l^t_{n_{m+1}}$ is higher than C_{m+1} and the subject has been flagged as a cheater.

The proposed procedure is then been called CHeater identification using Interim Person fit Statistics (CHIPS).

However, a priori consideration that can be made about the CHIPS is that the proposed $l^t_{n_m}$ statistic, being entirely dependent on the RTs, may react differently depending on the amount of pre-knowledge that cheaters have regarding the items in the main database. Indeed, having the same properties as the l^t_n statistic, it might be susceptible to fluctuations in RTs. This would make it vulnerable to cases where cheaters consistently respond quickly to all questions, which could happen when they have pre-knowledge of all the items in the main database. For this reason, a modified version was considered, the modified CHIPS (M-CHIPS), which, before starting to calculate $l^t_{n_m}$, administers some items from the more secret database to all those who have a very high estimate of speed according to $\hat{\zeta}_{n_{MLE_m}}$.

Regarding the more secure database, there are several considerations to be done. Firstly, the secure database should be constructed in such a way that it closely mimics the psychometric characteristics of the items in the main database items. This ensures that the distribution of item characteristics

in the secure database is as similar as possible to that of the main database, maintaining the integrity and fairness of the test. By analyzing the problem of cheaters in multiple-choice tests, in some cases, this objective can be achieved simply by changing some item content and the response options. Furthermore, it is crucial to keep the secure database regularly updated to incorporate new items and ensure the diversity and relevance of the questions. This helps maintain the security of the test and reduces the chances for cheaters to obtain pre-information about specific items.

All of this has a dual advantage. The first is to improve the estimation of the cheater's real ability. In fact, when the cheater is faced with a question for which they have no pre-knowledge, they will answer based on their own real ability, ensuring that the estimation of ability is no longer distorted due to pre-knowledge. The second result is that if an honest respondent is mistakenly classified as a cheater, they will not be penalized in any way, as the questions they will get, will be similar to those that would have been administered to them normally. However, it is still not desirable to have many honest respondents wrongly classified as cheaters, as this would result in overexposure of the secure database items, making it more susceptible to information leakage.

In order to assess the performance of the CHIPS and M-CHIPS procedures, a simulation study was conducted in the Chapter 4.

Chapter 4

Simulation study

In this Chapter, the CHIPS and M-CHIPS will be tested in a simulation study, whose main characteristics are described in Section 4.1, and the results obtained will be presented and discussed in Section 4.2.

Firstly, an *introductory investigation* aiming to identify the most critical features to be replicated in the simulation was done. The goal was to simulate a scenario that could best reflect real-life testing situations where some individuals engage in cheating behaviors.

Subsequently, this scenario was used for a *preliminary analysis* comparing a CAT using the CHIPS approach to a CAT with the traditional IRT approach. The comparison was based on various performance indices, such as the BIAS and the Root Mean Square Error (RMSE) of the ability estimator, and graphs (scatter plots). 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. The same analyses were also repeated for the M-CHIPS.

Based on the results of this preliminary analysis, several research questions were formulated. To address the questions, one or more characteristics of the preliminary scenario were modified.

Following that, secondary questions were explored using simulation analyses to do in-depth analysis.

Finally, the obtained results were discussed holistically to assess the entire analysis, while also understanding its limitations and potential future developments.

4.1 Simulation setup

First and foremost, the simulation setup that could best mirror the real-life situation was sought. To achieve this, various essential points were analyzed to construct the preliminary simulation. Indeed, these points constitute both the general structure of the CAT (Chapter 2.5) and the ones required for the implementation of the CHIPS and M-CHIPS.

- **The population size**. It was decided to use a population size of $N = 100$. The reason behind this choice is that the estimates of abilities for the simulated subjects are independent of each other, so there is no need to simulate a large number of students. Additionally, 100 subjects are more than sufficient to cover the most plausible values for both abilities and speeds. Moreover, having a moderately sized population allows for faster computation of the simulation, especially because replications are needed.
- **Interim and final ability estimation methods**. To estimate the interim abilities, the MAP method was employed. For the estimation of the final ability, the MLE method was used. Both methods were utilized to compare the classical CAT and the CAT with the CHIPS approach.
- **Starting items**. For each test taker, the first 5 starting items were randomly selected.
- **Item selection rule**. The MIC was chosen as item selection rule.
- **Stopping rule**. Each test is terminated once it reaches the maximum length of $K = 35$ items.

- **Main and secret database**. The main database 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 mirrored to the main one, therefore it contains items with the same psychometric characteristics.

- **Ability and speed distribution**. For each student θ_n and ζ_n were simulated from a bivariate normal distribution (van der Linden, 2007, Fox and Marianti, 2016, Fox et al., 2021) 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 obtained from the joint analysis of the INVALSI data, as discussed in Section 2.4.4. This approach ensures that the estimates closely resemble those of real students.

- **Ability and speed distribution of cheaters**. 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 honest respondents ($N_H = 80$). 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.

- **Simulated answers**. For each item $k = 1, \cdots, K$ and for each test taker $n = 1, \cdots, N$, the answer is simulated based on the 2PL model (Equation 2.5). However, if the test taker n is a cheater with pre-knowledge on item k, the answer will always be correct ($Y_{nk} = 1$).

- **Item pre-knowledge**. The number of items in the main database on which the cheaters have pre-knowledge depends on the scenario and can

be 50%, 75% or 100% of the total. It is also assumed that cheaters have no item pre-knowledge for the more secure database.

- **Simulated RT**. For each item $k = 1, \cdots, K$, and for each test taker $n = 1, \cdots, N$ the RT is simulated based on the equation

$$\ln RT_{nk} = \lambda_k - \phi_k^* \zeta_n + \varepsilon_{nk}, \quad \varepsilon_{nk} \sim N(0, \sigma_{\varepsilon_k}^2). \tag{4.1}$$

  In this way, even the randomness in response times is taken into account.

- **Cheater's RT-divider**. To simulate the speed at which cheaters respond to items they have pre-knowledge about, a two-step process was adopted. Firstly, the normal RT they would take based on their speed was estimated. Then, this RT was divided by a predetermined value known as the RT-divider. For instance, if a cheater with a speed of ζ_n should take 100 seconds to answer to a given item k under normal circumstances (Equation 4.1), if they have pre-knowledge on that item, and with an RT-divider set to 4, they would instead take 25 seconds to answer. This approach maintains the randomness of response time (since is retained the random component in Equation (4.1) while preserving its connection to the individual's speed. For the choice of RT-dividers, there are no explicit references in the literature. Therefore, in order to choose values that can best represent a plausible situation, a graphical comparison was performed between the estimated speed distributions of 100 honest respondents and 100 cheaters. This comparison was conducted for 9 different values of the RT-divider (ranging from 2 to 10). Each graph in Figure 4.1 displays the density curve of the estimated speed (Equation 3.16) for both honest respondents (yellow) and cheaters (blue). Since the modifications to the RT-divider do not affect honest respondents in any way, the only curve that changes is that of the cheaters. In general, as we are assuming a scenario in which cheaters behave differently

from honest respondents, we are looking for a setup that allows for a substantial but plausible difference.

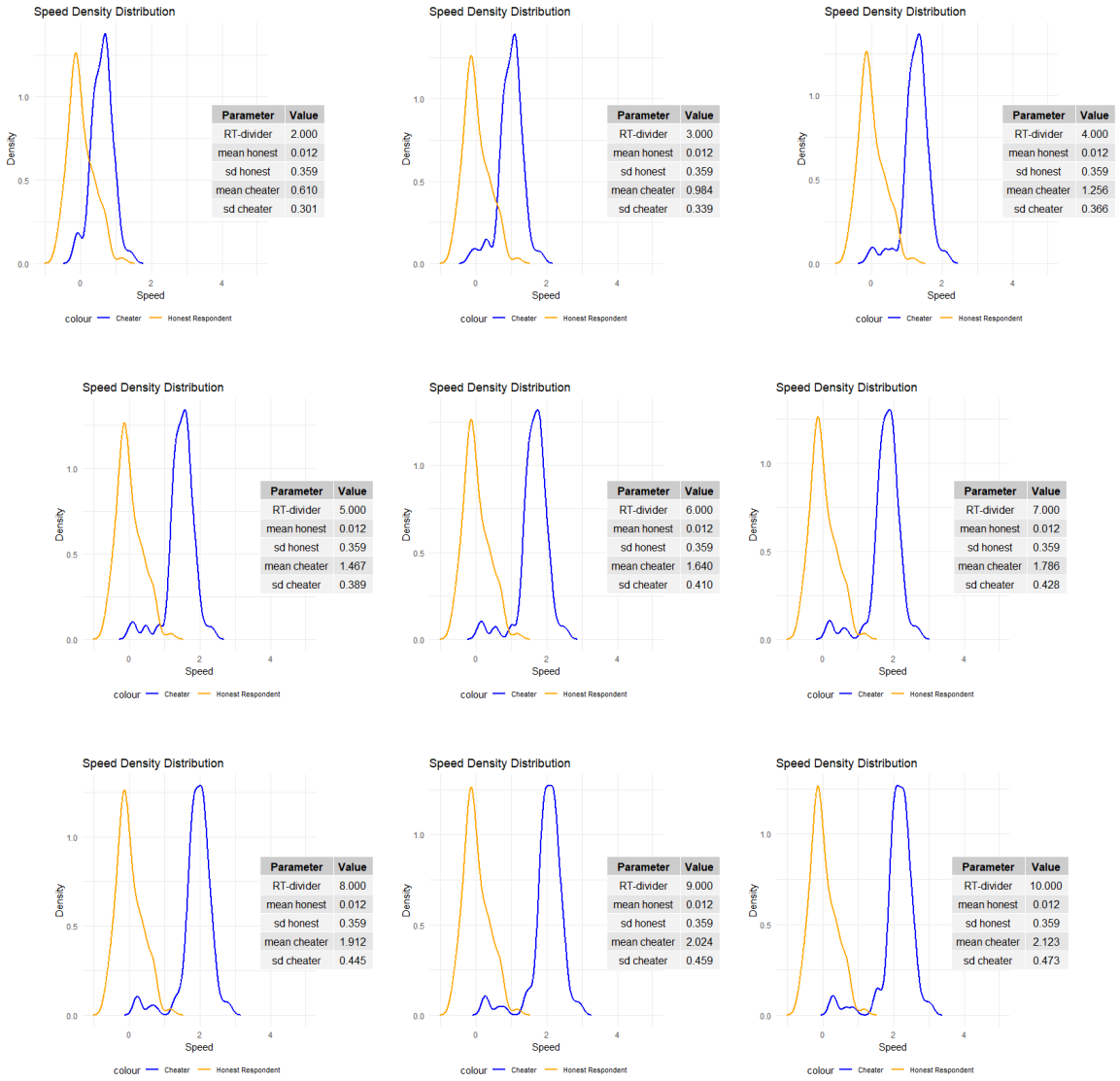

Fig. 4.1 Expected speed comparison for honest respondents and cheaters.

For RT-divider values of 2 and 3, the two curves show little difference and have a high level of overlap. However, starting from an RT-divider value of 4, the overlap between the two curves reduces, and only the last percentiles of the honest respondents' curve lie below the cheaters' curve. Additionally, the mean of cheaters' RT is approximately 1, which

is a plausible value for individuals responding very quickly to items. As the RT-divider increases, the two curves diverge further, with decreasing overlap, and the mean RT of cheaters increases, eventually reaching the limit value of 2 for an RT-divider of 8. Based on these observations, it was decided to use an RT-divider of **4**, as it is the first value to show plausible overlap and mean value.

- **Significance level**. A significance level $\alpha = 0.05$ was chosen for the CHIPS test.
- **The number of simulation replications**. To minimize the influence of random errors on the obtained results, each simulation was replicated **100** times. The results presented in the following section are the averages of the outcomes from these 100 repetitions.

The simulations were performed on *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 implement the CHIPS.

4.2 Results

Regarding the results of the preliminary analysis, the first step was to verify whether, even within a simulated scenario, the $l^t_{n_m}$ statistic is distributed as a χ^2_m. Only for this verification, the population dimensionality was increased to 1000 units and, to reflect the hypothesized distribution pattern, the simulated percentage of cheaters was reduced to 5%. The statistic was evaluated for $m = 35$ items and a pre-knowledge percentage of 75%.

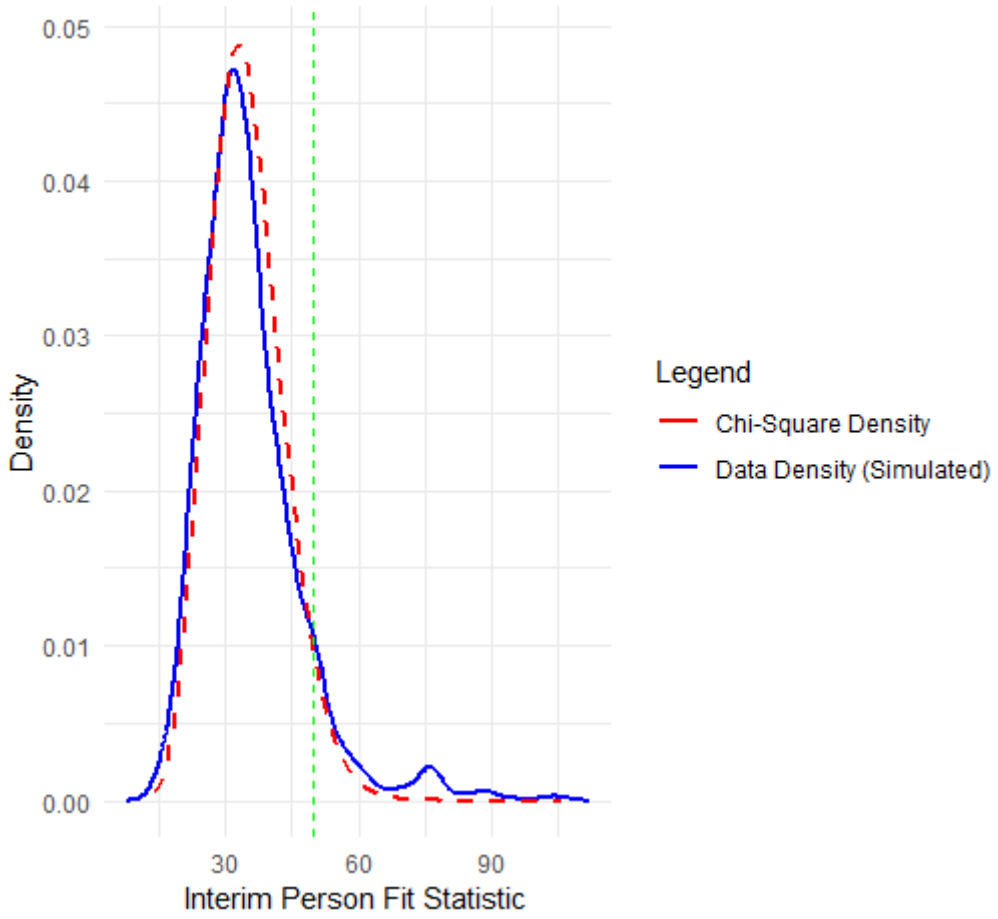

Fig. 4.2 Simulated $l^t_{n_m}$ density distribution compared with the χ^2_{35} density distribution.

Figure 4.2 displays a distribution trend of the simulated data (solid blue line) closely adhering to that of a χ^2_{35} (dashed red line). Only at the extreme right tail, a deviation can be observed. This discrepancy is attributed to the fact that this area encompasses most of the cheaters' statistics, which are correctly identified as such, being located to the right of the threshold C (dashed green vertical line). In fact, from Figure 4.3, representing the distribution of the statistic in the two groups of cheaters and honest respondents, it is evident that honest respondents consistently follow a χ^2_{35} distribution (solid orange line), while the cheaters, as desired, follow a normal distribution (solid blue line) with a mean greater than C.

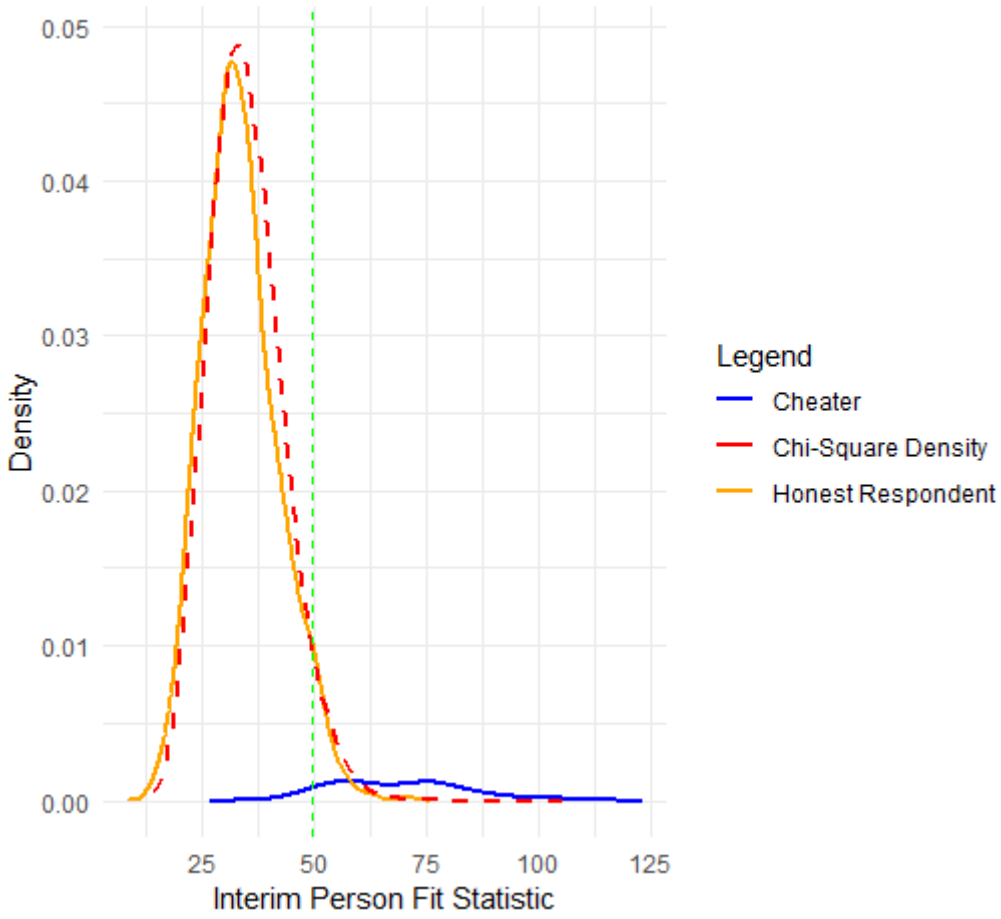

Fig. 4.3 $l^t_{n_m}$ density distribution comparison for honest respondents and cheaters.

Those results are also supported by both, a Q-Q plot of the honest respondents' statistic and the χ^2_{35} distribution (Figure 4.4), and the one-sample ($N = 100$) Kolmogorov-Smirnov test that does not reject the null hypothesis that the honest respondents' statistic belong to a χ^2_{35} distribution, with a statistic $D = 0.101$ and a p-value $= 0.27$.

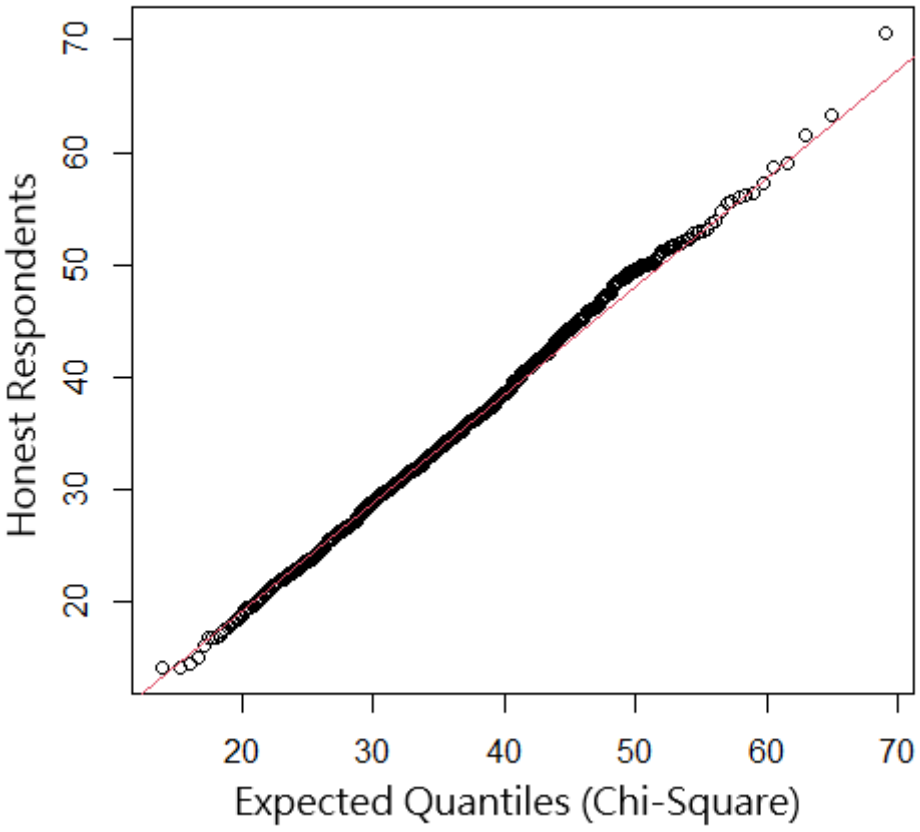

Fig. 4.4 Q-Q plot of honest respondents' $l^t_{n_m}$ and χ^2_{35} distribution.

Once verified that $l^t_{n_m}$ follows its theoretical distribution, the results of the preliminary study can be analyzed.

Table 4.1 presents BIAS and RMSE of the ability estimator for both classic IRT and CHIPS methods for each level of pre-knowledge.

As can be observed, with an increase in pre-knowledge, both the IRT and CHIPS methods exhibit an increase in both BIAS and RMSE. Furthermore, for any level of pre-knowledge, both indices are lower for the CHIPS method compared to the IRT based method, indicating superior performance in accurately estimating the test takers' true abilities in the former. However, if

the absolute values of both indices are considered, while CHIPS provides acceptable values for the first two pre-knowledge levels, when pre-knowledge reaches 100%, notably higher values are observed, albeit lower than those of the IRT based method. Moreover, across all levels of pre-knowledge, the CHIPS method outperforms the IRT-based method, demonstrating enhanced accuracy in estimating the true abilities of test-takers. Nonetheless, at 100% pre-knowledge, notably high values for BIAS and RMSE are observed, albeit still lower than those of the IRT-based method. So, from the analysis of these initial results, it appears that the CHIPS performs very well for the first two levels of pre-knowledge, but doesn't greatly improve the performance compared to the IRT-based method when pre-knowledge (P-K) reaches 100%.

	BIAS		**RMSE**	
P-K	**IRT**	**CHIPS**	**IRT**	**CHIPS**
50%	0.196	0.066	0.288	0.100
75%	0.398	0.115	0.968	0.185
100%	1.212	1.112	7.238	6.554

Table 4.1 BIAS and RMSE of ability for IRT and CHIPS. Preliminary analysis.

The same results are graphically presented using scatter plots depicting the estimated against the true values of θ_n (Figure 4.5).

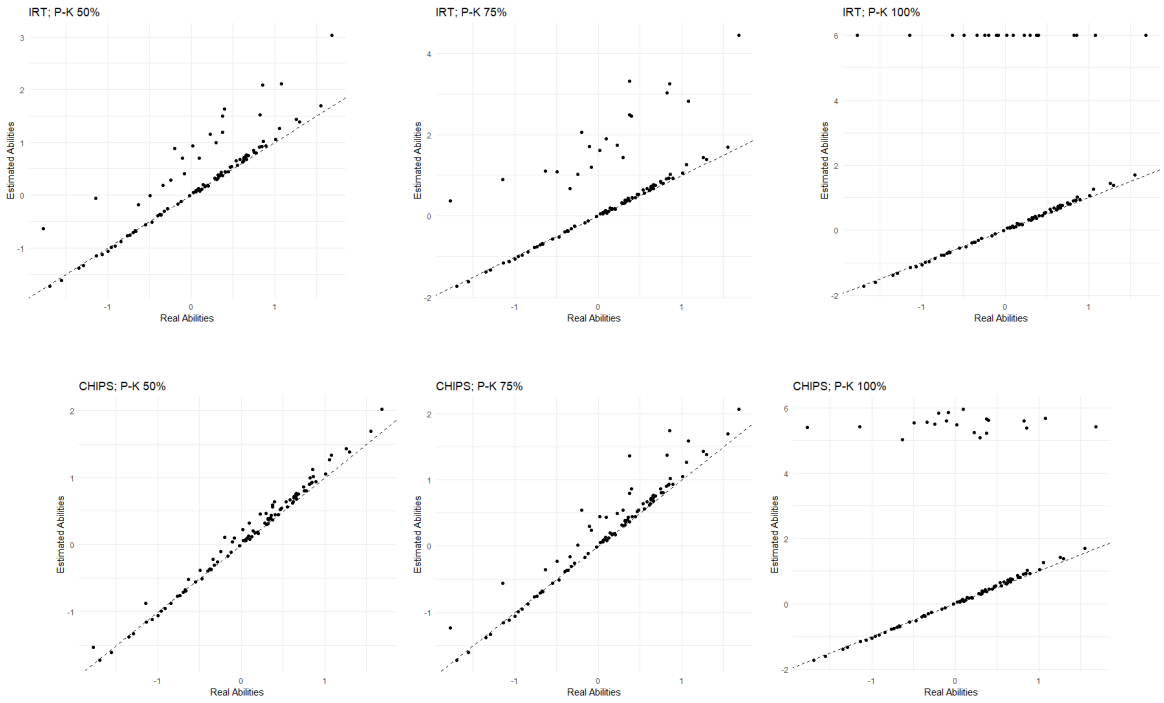

Fig. 4.5 Scatter plot of real and estimated abilities. Preliminary analysis.

As evident, some point tend to deviate from the bisecting line as pre-knowledge increases, aligning with the earlier discussion. Nevertheless, it is also noticeable that a notable portion of points remains relatively clustered near the bisecting line, regardless of the method used or the level of pre-knowledge, with only a minor fraction scattering.

To gain a clearer understanding, honest respondents were examined separately from cheaters, using both methods, and by evaluating BIAS (Table 4.2) and RMSE (Table 4.3).

	HONEST RESPONDENTS		**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

Table 4.2 BIAS of ability for cheaters and honest respondents. Preliminary analysis.

	HONEST RESPONDENTS		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 4.3 RMSE of ability for cheaters and honest respondents. Preliminary analysis.

In this case as well, both measures highlight the same results. Specifically, when considering only the cheaters, CHIPS performs better than IRT, but both BIAS and RMSE will increase with growing pre-knowledge. On the other hand, when considering only the honest respondents, these indices not only remain consistent across different levels but are also identical for both methods. Clearly, the fact that the indices do not increase is due to the fact that pre-knowledge has no influence on honest respondents. It affects only the cheaters, as the results confirm.

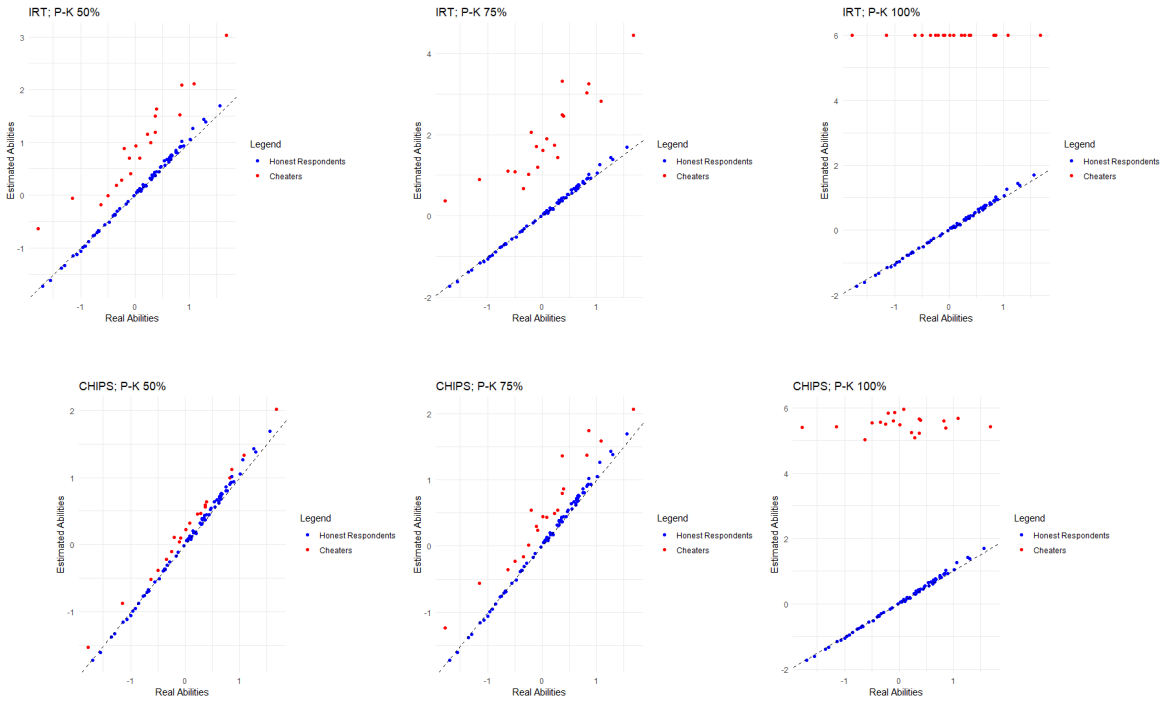

Fig. 4.6 Scatter plot of real and estimated abilities. Honest respondents and cheaters. Preliminary analysis.

Conversely, the equality in both BIAS and RMSE between the two methods demonstrates how CHIPS is able to improve the ability estimation for cheaters while it does not affect the one for honest respondents. Additionally, by examining the absolute values, it is evident that the values for honest respondents are very low for both indices, while those for cheaters are higher and increase with higher pre-knowledge. This implies that, although the percentage of cheaters is lower compared to honest respondents ($N_C = 20$, $N_H = 80$), it is the variations in the indices for cheaters that significantly impact the overall indices. In fact, concerning only the cheaters, the conclusions mirror those drawn in Table 4.1.

To better assess the impact of using CHIPS on cheaters' ability estimation performance, Table 4.4 displays the percentage reduction in both BIAS and RMSE, using CHIPS instead of IRT:

$$\Delta BIAS_C = 100\ \frac{BIAS_{CHIPS_C} - BIAS_{IRT_C}}{BIAS_{IRT_C}},$$
$$\Delta RMSE_C = 100\ \frac{RMSE_{CHIPS_C} - RMSE_{IRT_C}}{RMSE_{IRT_C}}. \tag{4.2}$$

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

Table 4.4 Percentage variation of cheaters' BIAS and RMSE using CHIPS instead of IRT. Preliminary analysis.

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

Furthermore, the results of the test analysis seem to lead to the same conclusions. In fact, CHIPS can also be interpreted as a *hypothesis test*, with the null hypothesis (H_0) stating that $\ln RT_{nk} \sim N\left(\lambda_k - \phi_k\zeta_n,\ \sigma_k^2\right)$, a test statistic $l^t_{n_m}$ with a known distribution (χ^2_m), and a critical value C. Moreover, not rejecting H_0 is equivalent to identifying the subject as an honest respondent, while rejecting H_0 is equivalent to identifying the subject as a cheater. From this perspective, the Type I error rate is simply the proportion of times an honest respondent is incorrectly classified as a cheater, making the true negative rate (1 − Type I error rate) represents the proportion of times an honest respondent is correctly classified. Consequently, the Type II error rate represents the proportion of times a cheater is incorrectly classified as an honest respondent, making (1 − Type II) error rate, i.e., the power of the test, the proportion of times a cheater is correctly classified.

Before delving into the results analysis, it is worth noting the Type I and Type II error rates in this context. Misclassifying a cheater (Type II error) means not modifying the test for that test taker, with the high risk of overestimating their actual ability. On the other hand, misclassifying an honest respondent (Type I error) means slightly exposing the more secure database, without penalizing that test-taker in any way, as the questions maintain the same psychometric properties. Therefore, it is preferable to keep the Type II error rate lower, thereby increasing the power of the test, as long as excessively high values of Type I error rate are not reached.

Tables 4.5, 4.6, 4.7 present test outcomes, for the three levels of pre-knowledge.

P-K = 50%		**Decision about H_0**	
		Fail to reject	**Reject**
H_0	**True**	0.957	0.043
is	**False**	0.043	0.957

Table 4.5 Decision table. Pre-knowledge 50%. Preliminary analysis.

P-K = 75%		**Decision about H_0**	
		Fail to reject	**Reject**
H_0	**True**	0.957	0.043
is	**False**	0.040	0.960

Table 4.6 Decision table. Pre-knowledge 75%. Preliminary analysis.

P-K = 100%		**Decision about H_0**	
		Fail to reject	**Reject**
H_0	**True**	0.957	0.043
is	**False**	0.894	0.107

Table 4.7 Decision table. Pre-knowledge 100%. Preliminary analysis.

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, as indicated in the simulation setup. Such a low value of Type I error rate explains why CHIPS is capable of not negatively influencing the estimation of honest respondents' abilities. In further support of this, it can be observed that the Type I error rate (and also the true negative rate) remains unchanged with greater pre-knowledge, just like the BIAS and RMSE for honest respondents. 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. Therefore, understandably, as the Type II error rate increases, the BIAS and RMSE for cheaters also increase, since CHIPS fails to correctly identify the majority of cheaters.

Regarding the reasons for the increase in the Type II error rate, they can be traced back to the formulation of the *interim person fit statistic* (IPS) (Equation 3.18). In fact, $l^t_{n_m}$, much like l^t_n, is highly dependent on the divergence between the expected and actual RT, rather than on the absolute value of the estimated speed. Even individuals with an extreme estimated speed, if they do not change their speed during the test, will have a value of the statistic which does not lead to the rejection of H_0. Therefore, in the case of 100% pre-knowledge, cheaters will consistently respond with an extreme speed, never encountering a question for which they do not know the answer. The 10% of correctly classified cheaters are attributed to the random component ε_{nk}, as specified in Equation (2.34).

To overcome this limitation, CHIPS has been slightly modified. Indeed, the algorithm has been adjusted to administer items from the more secure database to those who have a high interim speed estimate during the initial stages of the test. Specifically, after the algorithm administers the first 5 random items, the IPS is not calculated with $m = 5$, as was done for the initial analysis, but only the MLE of speed (Equation 3.17) is computed. At this point, for those with a value for $\hat{\zeta}_{MLE_m n}$ greater than a certain value U, the

subsequent 4 items (items 6-9) are selected from the more secret database. Only after the ninth items is answered, the Modified-CHIPS (M-CHIPS) returns to work as CHIPS. In this way, cheaters who will face a question of which they do not know in advance the answer will tend to slow down and the IPS will be able to better identify them.

Clearly, choosing a value of U 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 honest respondents, 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 fact, $\exp(0.694) = 2$. Hence, it represents a very high but plausible speed that only a few honest respondents should possess. Nevertheless, it is still lower than the speed exhibited by a cheater responding to questions they have pre-knowledge of.

To verify the effectiveness of this modification, a second simulation was conducted using the same setup as the previous one.

Table 4.8 illustrates how the M-CHIPS not only outperforms the IRT but also effectively overcomes the CHIPS limitation for 100% pre-knowledge.

	BIAS			**RMSE**		
P-K	**IRT**	**CHIPS**	**M-CHIPS**	**IRT**	**CHIPS**	**M-CHIPS**
50%	0.196	0.066	0.064	0.288	0.100	0.097
75%	0.398	0.115	0.074	0.968	0.185	0.109
100%	1.212	1.112	**0.155**	7.238	6.554	**0.463**

Table 4.8 BIAS and RMSE of ability for IRT, CHIPS and M-CHIPS. Preliminary analysis.

As highlighted in Tables 4.9, 4.10 and in Figure 4.7, the improvement is primarily due to the reduction in BIAS and RMSE for cheaters. In fact, the M-CHIPS manages to enhance the estimation of cheaters' abilities compared to CHIPS, without negatively impacting honest respondents' 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.

	HONEST RESPONDENTS			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**

Table 4.9 BIAS of ability for cheaters and honest respondents. IRT, CHIPS and M-CHIPS. Preliminary analysis.

	HONEST RESPONDENTS			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 4.10 RMSE of ability for cheaters and honest respondents. IRT, CHIPS and M-CHIPS. Preliminary analysis.

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

Table 4.11 Percentage variation of cheaters' BIAS and RMSE using M-CHIPS instead of IRT. Preliminary analysis.

Moreover, by comparing Table 4.11 with Table 4.4, 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 also exhibit a considerable difference when the last level is reached.

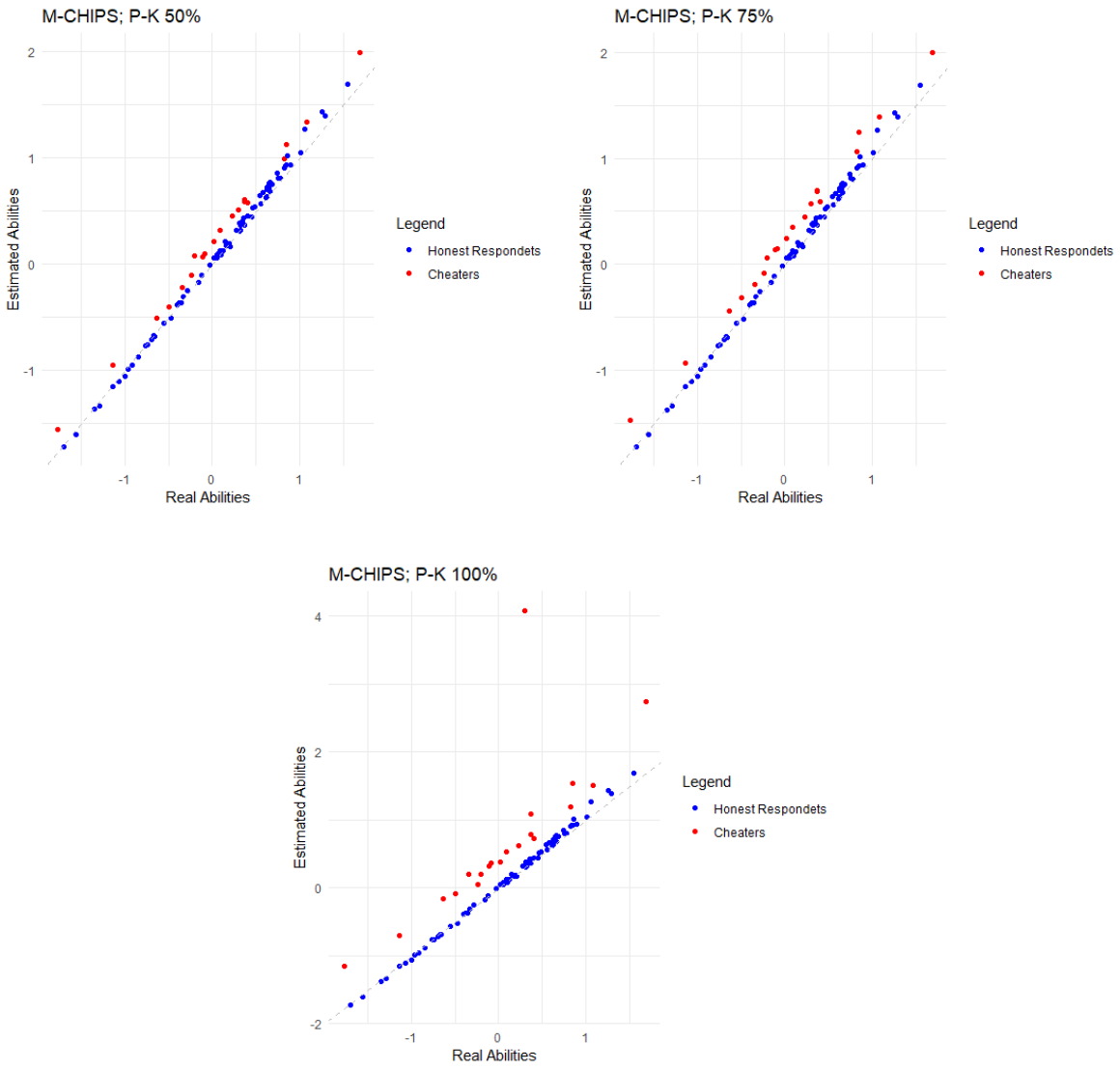

Fig. 4.7 Scatter plot of real and estimated abilities. M-CHIPS.

Finally, upon closer examination of Figure 4.7, it becomes evident that for 100% pre-knowledge, the points on the plot corresponding to cheaters (red points), though slightly distant from the bisector of the first quadrant (gray dashed line), seem to be aligned, except for two outliers. These outliers correspond to cheaters who were not identified as such by the M-CHIPS. 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) that ones displayed in Tables 4.9, 4.10. 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. Indeed, Tables 4.12, 4.13, 4.14 not only show that the Type I error rate has consistently remained below 0.05 but also demonstrate that, for 100% pre-knowledge, the test power has considerably increased compared to CHIPS, rising from 0.107 (Table 4.7) to 0.885 (Table 4.14).

P-K = 50%		**Decision about H_0**	
		Fail to reject	**Reject**
H_0	**True**	0.957	0.043
is	**False**	0.039	0.961

Table 4.12 Decision table. M-CHIPS. Pre-knowledge 50%. Preliminary analysis.

P-K = 75%		**Decision about H_0**	
		Fail to reject	**Reject**
H_0	**True**	0.957	0.043
is	**False**	0.030	0.970

Table 4.13 Decision table. M-CHIPS. Pre-knowledge 75%. Preliminary analysis.

P-K = 100%		**Decision about H_0**	
		Fail to reject	**Reject**
H_0	**True**	0.957	0.043
is	**False**	**0.115**	**0.885**

Table 4.14 Decision table. M-CHIPS. Pre-knowledge 100%. Preliminary analysis.

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

Given the achieved results, the subsequent analyses were conducted using M-CHIPS instead of CHIPS. Furthermore, these results were utilized to design the subsequent simulations. Indeed, the analysis of these findings brought forth several research questions, to which answers were sought through new simulations. The subsequent simulations were based on the same framework outlined in Section 4.1, but one condition was manipulated each time. The main questions are:

- What happens if α is increased or decreased?
- What happens if the correlation between θ_n and ζ_n is positive rather than negative?
- What happens if different parameters for the ability distribution are assumed between cheaters and honest respondents?
- What happens if, instead of fixed-length tests, variable-length tests are used?

What happens if α is increased or decreased? Regarding this first research question, the only changed setup variable was the significance level α, which was set at **0.1** and **0.01**.

	HONEST RESPONDENTS			**CHEATERS**		
P-K	**α = 0.01**	**α = 0.05**	**α = 0.1**	**α = 0.01**	**α = 0.05**	**α = 0.1**
50%	0.031	0.031	0.030	0.248	0.197	0.175
75%	0.031	0.031	0.030	0.352	0.249	0.209
100%	0.031	0.031	0.030	1.000	0.653	0.508

Table 4.15 BIAS of ability for cheaters and honest respondents. M-CHIPS. α = (0.01, 0.05, 0.1).

	HONEST RESPONDENTS			CHEATERS		
P-K	**α = 0.01**	**α = 0.05**	**α = 0.1**	**α = 0.01**	**α = 0.05**	**α = 0.1**
50%	0.088	0.087	0.087	0.171	0.135	0.125
75%	0.088	0.087	0.087	0.308	0.195	0.153
100%	0.088	0.087	0.087	2.681	1.964	1.500

Table 4.16 RMSE of ability for cheaters and honest respondents. M-CHIPS. α = (0.01, 0.05, 0.1).

Tables 4.15, 4.16 confirm what is easily hypothesized, but they also provide a result that might not be as expected. They show not only how an increase in the Type I error rate, regardless of the pre-knowledge level, leads to an increase in both BIAS and RMSE for cheaters, but also how this increase has almost no effect on the indices of honest respondents. The first result was anticipated, as an increase in the significance level α should facilitate the identification of cheaters, resulting in improved estimates of their abilities. On the other hand, an increase in α should also lead to more honest respondents being incorrectly classified as cheaters, which might potentially worsen the estimates. However, this does not seem to be the case. To obtain a more comprehensive and precise view, the results of the test analysis are also examined.

		α = 0.01		α = 0.05		α = 0.1	
P-K = 50%		**Decision about H_0**		**Decision about H_0**		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.993	0.007	0.957	0.043	0.914	0.086
is	**False**	0.063	0.937	0.039	0.961	0.032	0.968

Table 4.17 Decision table. Pre-knowledge 50%. M-CHIPS. α = (0.01, 0.05, 0.1).

P-K = 75%		α = 0.01		α = 0.05		α = 0.1	
		Decision about H_0		Decision about H_0		Decision about H_0	
		Fail to reject	Reject	Fail to reject	Reject	Fail to reject	Reject
H_0	True	0.993	0.007	0.957	0.043	0.914	0.086
is	False	0.047	0.953	0.030	0.970	0.021	0.979

Table 4.18 Decision table. Pre-knowledge 75%. M-CHIPS. α = (0.01, 0.05, 0.1).

P-K = 100%		α = 0.01		α = 0.05		α = 0.1	
		Decision about H_0		Decision about H_0		Decision about H_0	
		Fail to reject	Reject	Fail to reject	Reject	Fail to reject	Reject
H_0	True	0.993	0.007	0.957	0.043	0.914	0.086
is	False	0.234	0.766	0.115	0.885	0.073	0.927

Table 4.19 Decision table. Pre-knowledge 100%. M-CHIPS. α = (0.01, 0.05, 0.1).

Tables 4.17, 4.18, 4.19 show that for any level of pre-knowledge, as the significance level α increases, not only the Type I error rate does increase, but its value tends to align with the chosen value of α. Thus, indeed, as α increases, a higher percentage of honest respondents are incorrectly classified as cheaters. However, as seen in Tables 4.15, 4.16, this increase in Type I error rate does not lead to a worsening in the estimates of honest respondents' abilities. This confirms the earlier anticipation about the difference between the two error types. Specifically, an increase in the Type I error rate doesn't negatively impact honest respondents. Instead, it results in the overexposure of the database itself. On the other hand, concerning the Type II error rate, an increase in α leads to a decrease in the Type II error rate and therefore an increase in the power of the test. In line with what was observed in Tables 4.15, 4.16, having more cheaters correctly classified, positively impacts their ability estimates. Furthermore, this reduction in the Type II error rate is noticeable for all three levels of pre-knowledge, albeit being modest for the first two levels, and becoming more prominent when pre-knowledge reaches 100%. Similarly, a decrease in α appears to lead to a reduction in Type I error

rate, resulting in a reduced exposure of the more secure database. However, this reduction does not translate into a relevant improvement in the ability to estimate the performance of honest respondents. Conversely, the Type II error rate increases, resulting in a decrease in the ability to estimate the performance of cheaters.

In summary, in response to the question "*what happens if α is increased or decreased?*", it can be stated that the Type I error rate tends to align with the chosen value of α. This variation leads to a different overexposure of items within the more secret database, but it does not affect the quality of estimates for honest respondents. Conversely, an increase in α leads to a decrease in the Type II error rate, which positively affects the quality of estimates for cheaters, especially when the pre-knowledge level is 100%. For this reason, when choosing the value of α to use, two factors should be primarily evaluated: how important it is not to overexpose the items in the more secret database, and how much pre-knowledge is assumed among the cheaters for the items in the main database. If items from the more secret database are easily replaceable and high pre-knowledge among cheaters is presumed, then a higher α (0.1) might be preferred. Conversely, if maintaining minimal exposure of the more secret database is a priority and the main database is assumed to be less compromised, a smaller α (0.01) would be a better choice.

What happens if the correlation between θ_n and ζ_n is positive rather than negative? Regarding this next research question, α was set equal to 0.05 as in the baseline setup and the changed condition was the correlation between θ_n and ζ_n, that from -0.5 to 0.5.

Before delving into the analysis of the results, it is important to note that this change in correlation led to the simulation of a population with different abilities and speeds. Therefore, it is essential to keep in mind that the values of the indices do not refer to the same population, even though the two

populations share many distributional characteristics despite the correlation between the two latent traits.

P-K	HONEST RESPONDENTS		CHEATERS	
	NEGATIVE CORRELATION	POSITIVE CORRELATION	NEGATIVE CORRELATION	POSITIVE CORRELATION
50%	0.031	0.031	0.197	0.209
75%	0.031	0.031	0.249	0.261
100%	0.031	0.031	0.653	0.706

Table 4.20 BIAS of ability for cheaters and honest respondents. M-CHIPS. Positive and negative correlation.

P-K	HONEST RESPONDENTS		CHEATERS	
	NEGATIVE CORRELATION	POSITIVE CORRELATION	NEGATIVE CORRELATION	POSITIVE CORRELATION
50%	0.087	0.085	0.135	0.135
75%	0.087	0.085	0.195	0.190
100%	0.087	0.085	1.964	2.641

Table 4.21 RMSE of ability for cheaters and honest respondents. M-CHIPS. Positive and negative correlation.

The results from Tables 4.20, 4.21 suggest that the differences in estimation accuracy, due to the change in the correlation between latent traits, are quite minimal. There seems to be a slight improvement in the indices for honest respondents for a positive correlation, and a slight improvement for cheaters for a negative correlation. The only measure that appears to show high differences between the two cases is the RMSE for a pre-knowledge of 100%. However, the difference between the two BIAS values is not as pronounced, so this difference could be attributed to the fact that two different populations were simulated.

The test analysis results from Tables 4.22, 4.23, 4.24 also show very similar values for both the Type I error rate and the Type II error rate, regardless of the type of correlation, and this holds true across different levels of pre-knowledge.

One plausible explanation for these findings is that the IPS primarily utilizes information related to response time and does not consider whether the response is correct or not. Therefore, the change in correlation might not have a significant impact on the cheating detection process and subsequently on the estimation process. As a future development, modifying the IPS to incorporate the type of response given, and then evaluating if the sign of the correlation affects the method performance significantly, could be a potential avenue to explore.

P-K = 50%		**NEGATIVE CORRELATION**		**POSITIVE CORRELATION**	
		Decision about H_0		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.957	0.043	0.963	0.037
is	**False**	0.039	0.961	0.060	0.94

Table 4.22 Decision table. Pre-knowledge 50%. M-CHIPS. Positive and negative correlation.

P-K = 75%		**NEGATIVE CORRELATION**		**POSITIVE CORRELATION**	
		Decision about H_0		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.957	0.043	0.963	0.037
is	**False**	0.030	0.970	0.035	0.965

Table 4.23 Decision table. Pre-knowledge 75%. M-CHIPS. Positive and negative correlation.

P-K = 100%		NEGATIVE CORRELATION		POSITIVE CORRELATION	
		Decision about H_0		Decision about H_0	
		Fail to reject	Reject	Fail to reject	Reject
H_0	True	0.957	0.043	0.963	0.037
is	False	0.115	0.885	0.108	0.892

Table 4.24 Decision table. Pre-knowledge 100%. M-CHIPS. Positive and negative correlation.

In summary, the answer to the question "*what happens if the correlation between θ_n and ζ_n is positive rather than negative?*" could be that the M-CHIPS performs similarly, without undergoing significant changes. This reinforces the notion that the IPS primary reliance on response time information might mitigate the impact of correlation changes on the cheating detection process.

What happens if different parameters for the ability distribution are assumed between cheaters and honest respondents? For the honest respondents, the characteristics of the distribution remained unchanged, while those of the cheaters were generated from a bivariate normal distribution where the only difference was a mean ability of -1 instead of 0. This modification aimed to create a realistic scenario where cheaters may be less skilled than honest respondents and resort to cheating to compensate for this deficit.

Since the distribution of abilities for honest respondents was the same of the baseline setup, Table 4.25 displays the BIAS and RMSE values for cheaters only, with average abilities of 0 and -1, respectively. This setup allows for a direct comparison of the estimation performance under these different conditions.

	$BIAS_C$		$RMSE_C$	
P-K	$\mu_{\theta_C} = 0$	$\mu_{\theta_C} = -1$	$\mu_{\theta_C} = 0$	$\mu_{\theta_C} = -1$
50%	0.197	0.217	0.135	0.102
75%	0.249	0.291	0.195	0.172
100%	0.653	0.650	1.964	1.800

Table 4.25 BIAS and RMSE of ability for cheaters. M-CHIPS. μ_{θ_C} = -1; μ_{θ_C} = 0.

Similarly to the scenario with positive correlation, the change in setup does not appear to have had a significant impact of the method performance. In fact, the BIAS is slightly higher for $\mu_{\theta_C} = -1$ compared to $\mu_{\theta_C} = 0$ for the first two levels of pre-knowledge. However, this is offset by a lower RMSE. For a pre-knowledge level of 100%, both BIAS and RMSE are slightly lower for $\mu_{\theta_C} = -1$. This suggests that even when the abilities of cheaters are drawn from a distribution with a lower mean, the M-CHIPS method is able to provide reasonably accurate estimates of their abilities.

Once again, the results presented are supported by the test analysis (Table 4.26), which mirrors the previous results with very similar Type II error rate values for all three levels of pre-knowledge. Of course, the Type I error rate remains the same since the population of honest respondents has remained unchanged.

P-K = 50%		$\mu_{\theta_C} = 0$		$\mu_{\theta_C} = -1$	
		Decision about H_0		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.957	0.043	0.957	0.043
is	**False**	0.039	0.961	0.043	0.957

Table 4.26 Decision table. Pre-knowledge 50%. M-CHIPS. μ_{θ_C} = -1; μ_{θ_C} = 0.

P-K = 75%		$\mu_{\theta_C} = 0$		$\mu_{\theta_C} = -1$	
		Decision about H_0		Decision about H_0	
		Fail to reject	Reject	Fail to reject	Reject
H_0	True	0.957	0.043	0.957	0.043
is	False	0.030	0.970	0.030	0.970

Table 4.27 Decision table. Pre-knowledge 75%. M-CHIPS. μ_{θ_C} = -1; μ_{θ_C} = 0.

P-K = 100%		$\mu_{\theta_C} = 0$		$\mu_{\theta_C} = -1$	
		Decision about H_0		Decision about H_0	
		Fail to reject	Reject	Fail to reject	Reject
H_0	True	0.957	0.043	0.957	0.043
is	False	0.0115	0.885	0.100	0.900

Table 4.28 Decision table. Pre-knowledge 100%. M-CHIPS. μ_{θ_C} = -1; μ_{θ_C} = 0.

Certainly, before moving on to the final conclusions regarding this research question, it is worth noting that changing the mean ability of cheaters has a direct impact on their speed. This is because the assumption of negative correlation between the two implies that cheaters, in this case, are characterized by low abilities but high speeds. This opens up a *secondary question*:

What happens when μ_{θ_C} is decreased in the case where the correlation between latent factors is positive instead of negative? In this case as well, Table 4.29 presents the BIAS and RMSE for cheaters only, as no modifications were made to the distribution of honest respondents. Furthermore, the comparison is directly proposed against the preliminary setup.

P-K	$BIAS_C$		$RMSE_C$	
	$\mu_{\theta_C} = 0$ Negative Correlation	$\mu_{\theta_C} = -1$ Positive Correlation	$\mu_{\theta_C} = 0$ Negative Correlation	$\mu_{\theta_C} = -1$ Positive Correlation
50%	0.197	0.219	0.135	0.107
75%	0.249	0.31	0.195	0.231
100%	**0.653**	**1.021**	**1.964**	**5.369**

Table 4.29 BIAS and RMSE of ability for cheaters. Positive and negative correlation. M-CHIPS. μ_{θ_C} = -1; μ_{θ_C} = 0.

Table 4.29 shows that, for the first two levels of pre-knowledge, the M-CHIPS performs almost similarly in both proposed scenarios, with a slightly worse performance in the case of positive correlation and less able cheaters. However, when pre-knowledge increases to 100%, the worsening is more evident. To better understand the reasons behind this result, the BIAS and RMSE for correctly and incorrectly classified cheaters were reviewed. Tables 4.30, 4.31 demonstrate that, even for the second setup, the worsening of both indices is largely attributed to the incorrectly classified cheaters. In fact, for the correctly classified cheaters, the values of both BIAS and RMSE in the two setups are quite similar to each other.

	$\mu_{\theta_C} = 0$ Negative Correlation		$\mu_{\theta_C} = -1$ Positive Correlation	
	CORRECT CLASSIFIED CHEATER	INCORRECT CLASSIFIED CHEATER	CORRECT CLASSIFIED CHEATER	INCORRECT CLASSIFIED CHEATER
P-K 100%	**0.331**	2.266	**0.374**	3.288

Table 4.30 BIAS of ability for correct and incorrect classified cheaters. Pre-knowledge 100%. M-CHIPS.

	$\mu_{\theta_C} = 0$ Negative Correlation		$\mu_{\theta_C} = -1$ Positive Correlation	
	CORRECT CLASSIFIED CHEATER	**INCORRECT CLASSIFIED CHEATER**	**CORRECT CLASSIFIED CHEATER**	**INCORRECT CLASSIFIED CHEATER**
P-K 100%	**0.361**	8.109	**0.386**	17.61

Table 4.31 RMSE of ability for correct and incorrect classified cheaters. Pre-knowledge 100%. M-CHIPS.

Finally, the test analysis also supports these findings. From Tables 4.32, 4.33, 4.34, it's apparent that, for the first two levels of pre-knowledge, the two scenarios share a very similar value of Type II error rate, while the difference becomes more significant for a pre-knowledge of 100%.

P-K = 50%		$\mu_{\theta_C} = 0$ Negative Correlation		$\mu_{\theta_C} = -1$ Positive Correlation	
		Decision about H_0		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.957	0.043	0.963	0.037
is	**False**	0.039	0.961	0.050	0.95

Table 4.32 Decision table. Pre-knowledge 50%. Positive and negative correlation. M-CHIPS. μ_{θ_C} = -1; μ_{θ_C} = 0.

P-K = 75%		$\mu_{\theta_C} = 0$ Negative Correlation		$\mu_{\theta_C} = -1$ Positive Correlation	
		Decision about H_0		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.957	0.043	0.963	0.037
is	**False**	0.030	0.970	0.027	0.973

Table 4.33 Decision table. Pre-knowledge 75%. M-CHIPS. Positive and negative correlation. μ_{θ_C} = -1; μ_{θ_C} = 0.

P-K = 100%		$\mu_{\theta_C} = 0$ Negative Correlation		$\mu_{\theta_C} = -1$ Positive Correlation	
		Decision about H_0		Decision about H_0	
		Fail to reject	Reject	Fail to reject	Reject
H_0	True	0.957	0.043	0.963	0.037
is	False	0.115	0.885	0.133	0.867

Table 4.34 Decision table. Pre-knowledge 100%. M-CHIPS. Positive and negative correlation. μ_{θ_C} = -1; μ_{θ_C} = 0.

A plausible explanation for these findings could be that, when there is a positive correlation between latent factors, cheaters with lower abilities exhibit lower speeds. For moderate to high levels of pre-knowledge (50%, 75%), the M-CHIPS seems to perform exceptionally well, accurately identifying cheaters almost 100% of the time. However, for pre-knowledge values of 100%, the slow speed of the cheaters keep the M-CHIPS from efficiently identifying them. This is somewhat corroborated by Table 4.28, where it can be observed that, conversely, when the average speed of cheaters is high (attributed to lower ability but in the case of negative correlation), the Type II error rate is 3% lower. Moreover, as the cheaters possess average lower abilities, when the M-CHIPS fails to correctly identify a cheater, both the BIAS and RMSE tend to be higher, as the disparity between the actual and estimated values becomes greater.

In conclusion, addressing the main question "*What happens if different parameters for the ability distribution are assumed between cheaters and honest respondents?*" reveals that the M-CHIPS performs nearly the same for moderately high pre-knowledge levels and even performs better when pre-knowledge is at 100%. This is due to the M-CHIPS being more effective at identifying cheaters when their response speeds are high. Consequently, if the correlation between latent factors becomes positive, a worsening in the capability to accurately classify cheaters is observed. This impact, however,

is primarily seen in the ability estimates of cheaters who have not been accurately classified.

What happens if, instead of fixed-length tests, variable-length tests are used? To address this question, the stopping rule was modified so that the test no longer terminated at a fixed length, but instead stopped when the variance of the test taker's estimate fell below a certain *target value*. Three different target values were chosen (after some preliminary analysis): **0.15**, **0.10**, and **0.05**. Furthermore, to mirror a realistic scenario, both a minimum value of 25 and a maximum value of 45 were set for the test length.

	HONEST RESPONDENTS			**CHEATERS**		
P-K	**T.V. = 0.15**	**T.V. = 0.10**	**T.V. = 0.05**	**T.V. = 0.15**	**T.V. = 0.10**	**T.V. = 0.05**
50%	0.024	0.008	0.015	0.240	0.212	0.176
75%	0.024	0.008	0.015	0.292	0.270	0.215
100%	0.024	0.008	0.015	0.684	0.656	0.596

Table 4.35 BIAS of ability for cheaters and honest respondents. M-CHIPS. Target value = (0.15, 0.10, 0.05).

	HONEST RESPONDENTS			**CHEATERS**		
P-K	**T.V. = 0.15**	**T.V. = 0.10**	**T.V. = 0.05**	**T.V. = 0.15**	**T.V. = 0.10**	**T.V. = 0.05**
50%	0.100	0.090	0.080	0.161	0.142	0.1112
75%	0.100	0.090	0.080	0.226	0.196	0.150
100%	0.100	0.090	0.080	1.894	1.858	1.786

Table 4.36 RMSE of ability for cheaters and honest respondents. M-CHIPS. Target value = (0.15, 0.10, 0.05).

From the analysis of Tables 4.35, 4.36, it can be observed that, regardless of the level of pre-knowledge, a decrease in the target value leads to a slight improvement in estimation performance. This holds true for both honest respondents and cheaters. This result is quite promising because it aligns with what typically happens in variable-length CAT. Therefore, M-CHIPS seems to work well in scenarios like these. Furthermore, to corroborate this, it is noticeable that these results tend to be very close to those of M-CHIPS at a fixed length, as shown in Tables 4.9, 4.10, especially for cheaters.

Regarding the observed test length for the three target values, Table 4.37 presents the average test length for honest respondents and cheaters.

	HONEST RESPONDENTS			**CHEATERS**		
P-K	**T.V. = 0.15**	**T.V. = 0.10**	**T.V. = 0.05**	**T.V. = 0.15**	**T.V. = 0.10**	**T.V. = 0.05**
50%	25.541	28.13	38.317	25.895	28.079	37.605
75%	25.541	28.13	38.317	26.119	28.547	38.402
100%	25.541	28.13	38.317	27.664	30.586	39.594

Table 4.37 Average test length for cheaters and honest respondents. M-CHIPS. Target value = (0.15, 0.10, 0.05).

Firstly, it can be observed that a decrease in the target value corresponds to an increase in the average test length, both for cheaters and honest respondents. This results aligns with what typically happens in CAT. Furthermore, the test length tends to increase with an increase in pre-knowledge (obviously only for cheaters). This results in the average test length for cheaters being lower than that of honest respondents for a pre-knowledge of 50%, while the opposite is true for a pre-knowledge of 100%, regardless of the target value. This outcome seems to be consistent with what has been discussed so far regarding the M-CHIPS. In fact, pre-knowledge can also be seen as the probability that a cheater is being administered a question to which they already know the

answer. When this is set at 50% (so a probability of 0.5), it means that, on average, half of the items selected from the main database are answered by the cheater with their real ability. Therefore, those responses effectively guide the estimation towards the true ability value. Conversely, a pre-knowledge value of 100% means that, without a doubt, the main database questions are not answered with the true ability, which worsens the estimation, especially if the cheater has a low ability. As confirmed by Table 4.37, this means that, on average, longer tests are needed.

Lastly, the test analysis has also been conducted.

P-K = 50%		**Target value = 0.15**		**Target value = 0.10**		**Target value = 0.05**	
		Decision about H_0		**Decision about H_0**		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.970	0.030	0.970	0.030	0.954	0.046
is	**False**	0.065	0.935	0.060	0.940	0.053	0.947

Table 4.38 Decision table. Pre-knowledge 50%. M-CHIPS. Target value = (0.15, 0.10, 0.05)

P-K = 75%		**Target value = 0.15**		**Target value = 0.10**		**Target value = 0.05**	
		Decision about H_0		**Decision about H_0**		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.970	0.030	0.970	0.030	0.954	0.046
is	**False**	0.041	0.959	0.035	0.965	0.031	0.969

Table 4.39 Decision table. Pre-knowledge 75%. M-CHIPS. Target value = (0.15, 0.10, 0.05)

P-K = 100%		**Target value = 0.15**		**Target value = 0.10**		**Target value = 0.05**	
		Decision about H_0		**Decision about H_0**		**Decision about H_0**	
		Fail to reject	**Reject**	**Fail to reject**	**Reject**	**Fail to reject**	**Reject**
H_0	**True**	0.970	0.030	0.970	0.030	0.954	0.046
is	**False**	0.116	0.884	0.099	0.901	0.094	0.906

Table 4.40 Decision table. Pre-knowledge 100%. M-CHIPS. Target value = (0.15, 0.10, 0.05)

From the analysis of Tables 4.38, 4.39, 4.40, it is possible to observe how, in this case as well, the test analysis aligns and confirms what was observed in the performance analysis. Indeed, concerning the honest respondents, the Type I error rate maintains almost identical values for all three proposed target values (and, of course, for all three levels of pre-knowledge). Also, regarding the cheaters, the Type II error rate seems to maintain similar values for all three target values, albeit showing a slight tendency to decrease as they get smaller. This appears to hold for all three levels of pre-knowledge, and as seen in the fixed-length case, an increase in pre-knowledge leads to a corresponding increase in the Type II error rate, particularly when it is set at 100%.

In summary, to the question "*what happens if, instead of fixed-length tests, variable-length tests are used?*" it can be answered that the M-CHIPS appears to maintain the same good performance it exhibits in the case of fixed-length tests. Indeed, the shadow test approach (STA), which allows applying the same constraints at each item selection phase, regardless of the overall test length, enables the use of the M-CHIPS method seamlessly. Furthermore, it preserves the properties where an increase in the target value corresponds to improved estimation performance (though not very pronounced in this case), alongside an increase in the average test length. Therefore, much like in a standard CAT, the choice of the target value should be weighed against the requirements of the test administrator (a shorter test or a more precise one).

Chapter 5

Conclusions

5.1 Concluding remarks

This thesis explores the interrelation of Computerized Adaptive Testing (CAT), Response Time (RT), and their potential in addressing cheating issues. The study begins with an extensive literature review covering Item Response Theory (IRT), RT, CAT, and cheating. It aims to integrate these topics through the proposal and examination of an innovative method for identifying and controlling cheating during CAT.

In Chapter 2, the thesis is divided into three main parts. The first part (Section 2.1 - 2.3) introduces IRT models and their estimation methods, serving as the foundation for the study. The second part (Section 2.4) delves into RT, discussing its significance in educational testing and presenting various RT distributions. The focus then shifts to specific models by van der Linden (2006) and Fox and Marianti (2016), motivated by their influence on the novel method proposed in this study. An application of RT models to real data from the Italian National Institute for the Evaluation of the Education and Training System (INVALSI) demonstrates how RT information can enhance ability estimation. The third part (Section 2.5) introduces CAT, providing an overview of its context and delving into its intricate structure. The discussion focuses on Item Selection Criteria (ISC) and highlights the integration of RT

information to enhance the phase of item selection. The chapter concludes with an original study comparing different CAT configurations, emphasizing CATs adaptability in handling content restrictions and adjusting test lengths.

Chapter 3 addresses cheating in educational tests, focusing on CAT. The first part (Sections 3.1, 3.2) reviews existing literature on cheating, both in general educational tests and specifically in CAT. The examination then narrows down to cheating involving pre-knowledge of item answers. Various methods for identifying cheaters with pre-knowledge, also incorporating RT information, are discussed. The second part (Section 3.3) proposes a new method for identifying cheaters during CAT, named CHeater identification using Interim Person fit Statistic (CHIPS). This method leverages response time to define the Interim Person fit Statistic (IPS), which helps determine whether a test-taker is a cheater. The method involves administering subsequent items from a more secure item database to suspected cheaters, improving estimates of their actual abilities. A modification of CHIPS, called M-CHIPS, is presented in the same section. M-CHIPS includes an additional step in which faster test-takers are administered items from the more secure database before the algorithm calculates the IPS.

In Chapter 4, CHIPS and M-CHIPS undergo a simulation study designed to mimic a real-world CAT scenario with varying levels of item pre-knowledge (Section 4.1). CHIPS is effective in identifying cheaters, but limitations arise at 100% pre-knowledge. M-CHIPS performs better, especially for 100% item pre-knowledge, without overexposing the more secure database. The simulation study is modified to answer specific research questions, revealing the flexibility of the proposed methods to various factors, including the test length, the significance value, the correlation between speed and ability, and the ability level of cheaters.

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. Moreover, 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. It is possible that someone might intentionally slow down to try to deceive the method.

As a final note, the proposed method is intended to integrate other existing solutions for cheater identification and treatment. The method itself is very easy to collect during a computer based environment and, as shown by the simulation results, it tends to provide good results. Clearly, this method is based on probabilistic reasoning, so the decision to rely solely on it should be considered carefully, especially depending on the type of test for which it is being used. For example, if it is a high-stakes test that typically involves more checks for cheating, then using CHIPS or M-CHIPS as a *filtering* method for an additional layer of security before employing more secure methods could be an option. Conversely, for a low-stakes test that may not involve any action against suspected cheaters, this method could be a good option. As seen from the simulation results, CHIPS importantly reduces uncertainties in estimating the true abilities of cheaters without affecting those of honest respondents. In these cases, the greatest risk is a potential overexposure of items in the more secure database and a reduced method effectiveness, without worsening the situation compared to not using the method.

5.2 Future developments

Regarding future developments, they are mostly linked to the aforementioned limitations. 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 but one of the other distributions mentioned

in Chapter 2. In essence, an increase in classification error rate is expected, with a corresponding deterioration in estimation performance. However, the extent of this deterioration, as well as whether it affects only cheaters or also honest respondents, cannot be hypothesized a priori. Moreover, it is interesting to understand if there are substantial differences depending on the models used and even if it is possible to modify the 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. Regarding the first aspect, it is plausible to assume that this significantly worsens the performance of CHIPS and M-CHIPS because these methodologies rely on administering items to cheaters on which they have no pre-knowledge, thus making them respond based on their actual abilities. Possible solutions could involve having multiple secret databases with different levels of security. As for the aspect of deliberately slowing down to try to deceive the method, although initially this might pose another significant problem, a more detailed analysis could show otherwise. In fact, as seen in the simulation study, both methods have the ability to react to deviations in observed response time from the hypothesized time. This implies that in order to successfully trick the method, a cheater must initially estimate their speed, understand the time-intensity and time-discrimination of each item they are responding to, and attempt to provide a credible response time based on their estimated speed.

Furthermore, the simulations proposed did not make assumptions about unusual behaviors of honest respondents, such as fast-guessing behaviors. which is expected to be much more common in low-stakes tests than in high-stakes ones, given the majority of test-takers who want to exert as little effort as possible and finish the test quickly. In essence, this is expected to result more in an increase in the Type I error rate, and therefore, an overexposure

of the more secret database, rather than a real decrease in the estimation performance of the abilities of honest respondents. Nonetheless, it would remain a serious issue to which attempt to find a solution. Therefore, a future development could be oriented towards addressing this, perhaps by modifying the method to leverage both RT and response pattern information.

Beyond the simulation, other developments could involve the implementation of methods capable of generating the more secure items directly during the test, 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 involve applying the method to datasets with recognized cases of cheating and also field testing the method in real-world scenarios.

References

Anders, R., Alario, F., and Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods*, 21(3):309–327.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43:561–573.

Audley, R. and Pike, A. (1965). Some alternative stochastic models of choice. *British Journal of Mathematical and Statistical Psychology*, 18(2):207–225.

Belov, D. (2012). Detection of large-scale item preknowledge in Computerized Adaptive Testing via Kullback-Leibler divergence. *Law School Admission Council Research Report Series*, pages 12–01.

Belov, D. (2016). Identification of item preknowledge by the methods of information theory and combinatorial optimization. In *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pages 164–176. Routledge.

Belov, D. I. (2013). Detection of test collusion via Kullback-Leibler divergence. *Journal of Educational Measurement*, 50(2):141–163.

Bertsimas, D. and Sim, M. (2003). Robust discrete optimization and network flows. *Mathematical Programming*, 98(1-3):49–71.

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

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F. M. and Novick, M. R., editors, *Statistical Theories of Mental Test Scores*, pages 392–479. Addison-Wesley, Reading.

Bock, R. D. and Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4):431–444.

Bolsinova, M., de Boeck, P., and Tijmstra, J. (2017). Modelling conditional dependence between response time and accuracy. *Psychometrika*, 82:1126–1148.

Chang, H.-H., Qian, J., and Ying, Z. (2001). A-stratified multistage computerized adaptive testing with B blocking. *Applied Psychological Measurement*, 25(4):333–341.

Chang, H.-H. and Ying, Z. (1999). A-stratified multistage computerized adaptive testing. *Applied Psychological Measurement*, 23(3):211–222.

Cheng, Y., Diao, Q., and Behrens, J. T. (2017). A simplified version of the maximum information per time unit method in computerized adaptive testing. *Behavior Research Methods*, 49:502–512.

Cizek, G. (2012). Ensuring the integrity of test scores: Shared responsibilities. In *Annual Meeting of the American Educational Research Association, Vancouver, British Columbia*.

Cizek, G. J. and Wollack, J. A. (2016). *Handbook of Quantitative Methods for Detecting Cheating on Tests*. Taylor & Francis Group, New York.

De Boeck, P., Chen, H., and Davison, M. (2017). Spontaneous and imposed speed of cognitive test responses. *British Journal of Mathematical and Statistical Psychology*, 70(2):225–237.

De Boeck, P. and Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology*, 10:102.

Dennis Jr, J. E. and Schnabel, R. B. (1996). *Numerical methods for unconstrained optimization and nonlinear equations*. SIAM, Philadelphia.

DiTrapani, J., Jeon, M., De Boeck, P., and Partchev, I. (2016). Attempting to differentiate fast and slow intelligence: Using generalized item response trees to examine the role of speed on intelligence tests. *Intelligence*, 56:82–92.

Drasgow, F., Levine, M. V., and Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1):67–86.

Eckerly, C., Babcock, B., and Wollack, J. (2015). Preknowledge detection using a scale-purified deterministic gated IRT model. In *Annual Meeting of the National Conference on Measurement in Education, Chicago, IL*.

Fan, Z., Wang, C., Chang, H.-H., and Douglas, J. (2012). Utilizing response time distributions for item selection in CAT. *Journal of Educational and Behavioral Statistics*, 37(5):655–670.

Fox, J.-P., Klotzke, K., and Simsek, A. S. (2021). LNIRT: An R package for joint modeling of response accuracy and times. *arXiv preprint arXiv:2106.10144*.

Fox, J.-P. and Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research*, 51(4):540–553.

Fox, J.-P. and Marianti, S. (2017). Person-fit statistics for joint models for accuracy and speed. *Journal of Educational Measurement*, 54(2):243–262.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Goldhammer, F., Naumann, J., and Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, 3(1):21–40.

Goldhammer, F., Naumann, J., Stelter, A., Tóth, K., Rölke, H., and Klieme, E. (2014). The time on task effect in reading and problem solving is moderated by task difficulty and skill: Insights from a computer-based large-scale assessment. *Journal of Educational Psychology*, 106(3):608–626.

Goldhammer, F., Steinwascher, M. A., Kroehne, U., and Naumann, J. (2017). Modelling individual response time effects between and within experimental speed conditions: A GLMM approach for speeded tests. *British Journal of Mathematical and Statistical Psychology*, 70(2):238–256.

Hambleton, R. K. and Swaminathan, H. (1985). *Item Response Theory: Principles and Applications*. Springer Science & Business Media, New York.

Han, N. (2003). Using moving averages to assess test and item security in computer based testing. *Center for Educational Assessment Research Report*, 468.

Jeon, M. and De Boeck, P. (2019). An analysis of an item-response strategy based on knowledge retrieval. *Behavior Research Methods*, 51:697–719.

Johnson, M. S., Sinharay, S., and Bradlow, E. T. (2006). 17 hierarchical item response theory models. *Handbook of Statistics*, 26:587–606.

Kang, H.-A. (2017). Penalized partial likelihood inference of proportional hazards latent trait models. *British Journal of Mathematical and Statistical Psychology*, 70(2):187–208.

Klein Entink, R. H., Kuhn, J.-T., Hornke, L. F., and Fox, J.-P. (2009). Evaluating cognitive theory: A joint modeling approach using responses and response times. *Psychological Methods*, 14(1):54.

Kroeze, K. (2017). *Multidimensional Computer Adaptive Testing with the Shadow Testing routine*. https://github.com/Karel-Kroeze/ShadowCAT.git.

Lazarsfeld, P. F. (1949). The American solidier—an expository review. *Public Opinion Quarterly*, 13(3):377–404.

Levine, M. V. and Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics*, 4(4):269–290.

Lo, S. and Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6:1171.

Loeys, T., Rosseel, Y., and Baten, K. (2011). A joint modeling approach for reaction time and accuracy in psycholinguistic experiments. *Psychometrika*, 76:487–503.

Lohman, D. F. (1989). Individual differences in errors and latencies on cognitive tasks. *Learning and Individual Differences*, 1(2):179–202.

Lord, F. M. (1952). *A Theory of Test Scores (Psychometric Monograph No. 7)*. Richmond, VA: Psychometric Corporation. Retrieved from http://www.psychometrika.org/journal/online/MN07.pdf.

Lord, F. M. (1980). *Applications of Item Response Theory to Practical Testing Problems*. Routledge, New York.

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, pages 157–162.

Lord, F. M. and Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Addison-Wesley, Reading.

Mair, P. and Gruber, K. (2022). Bayesian explanatory additive IRT models. *British Journal of Mathematical and Statistical Psychology*, 75(1):59–87.

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., and Tijmstra, J. (2014). Testing for aberrant behavior in response time modeling. *Journal of Educational and Behavioral Statistics*, 39(6):426–451.

Maris, E. (1993). Additive and multiplicative models for gamma distributed random variables, and their application as psychometric models for response times. *Psychometrika*, 58:445–469.

Matzke, D. and Wagenmakers, E.-J. (2009). Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis. *Psychonomic Bulletin & Review*, 16:798–817.

McLachlan, G. J. and Krishnan, T. (2007). *The EM Algorithm and Extensions*. John Wiley & Sons, New Jersey.

McLeod, L. D. (2006). *Detecting Items That Have Been Memorized*, volume 99. Law School Admission Council, Newtown.

Meyer, J. P. (2010). A mixture Rash model with item response time components. *Applied Psychological Measurement*, 34(7):521–538.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51:177–195.

Molenaar, D., Bolsinova, M., Rozsa, S., and De Boeck, P. (2016). Response mixture modeling of intraindividual differences in responses and response times to the Hungarian WISC-IV block design test. *Journal of Intelligence*, 4(3):10.

Molenaar, D., Bolsinova, M., and Vermunt, J. K. (2018). A semi-parametric within-subject mixture approach to the analyses of responses and response times. *British Journal of Mathematical and Statistical Psychology*, 71(2):205–228.

Molenaar, D., Tuerlinckx, F., and van der Maas, H. L. (2015). A bivariate generalized linear item response theory modeling framework to the analysis of responses and response times. *Multivariate Behavioral Research*, 50(1):56–74.

Naumann, J. and Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences*, 53:1–16.

Newton, P. M. and Essex, K. (2023). How common is cheating in online exams and did it increase during the COVID-19 pandemic? A Systematic Review. *Journal of Academic Ethics*, pages 1–21.

Novikov, N. A., Nurislamova, Y. M., Zhozhikashvili, N. A., Kalenkovich, E. E., Lapina, A. A., and Chernyshev, B. V. (2017). Slow and fast responses: Two mechanisms of trial outcome processing revealed by EEG oscillations. *Frontiers in Human Neuroscience*, 11:218.

Nydick, S. (2014). *Simulate IRT-Based Computerized Adaptive Tests*. https://github.com/swnydick/catIrt.

Obregon, P. (2013). A Bayesian approach to detecting compromised items. In *Annual Meeting of the National Council on Measurement in Education, San Francisco, CA*.

O'Leary, L. S. and Smith, R. W. (2016). Detecting candidate preknowledge and compromised content using differential person and item functioning. In Cizek, G. J. and Wollack, J. A., editors, *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pages 151–163. Routledge.

Partchev, I. and De Boeck, P. (2012). Can fast and slow intelligence be differentiated? *Intelligence*, 40(1):23–32.

Patton, J. M. (2015). *Some consequences of response time model misspecification in educational measurement*. PhD thesis, University of Notre Dame, Indiana.

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. and Kuhn, J.-T. (2012). A flexible latent trait model for response times in tests. *Psychometrika*, 77:31–47.

Ranger, J., Kuhn, J.-T., and Gaviria, J.-L. (2015). A race model for responses and response times in tests. *Psychometrika*, 80:791–810.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests.* MESA Press, Chicago.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. Springer New York, NY.

Roskam, E. E. (1987). Toward a psychometric theory of intelligence. In Roskam, E. E. and Suck, R., editors, *Progress in Mathematical Psychology*, page 151–174. Elsevier Science.

Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., and Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 80:491–513.

Schnabel, R. B., Koonatz, J. E., and Weiss, B. E. (1985). A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software (TOMS)*, 11(4):419–440.

Segall, D. O. (1997). Equating the CAT-ASVAB. In W. A. Sands, B. K. Waters, . J. R. M., editor, *Computerized adaptive testing: From inquiry to operation*, page 181–198. American Psychological Association.

Shu, Z., Henson, R., and Luecht, R. (2013). Using deterministic, gated item response theory model to detect test cheating due to item compromise. *Psychometrika*, 78:481–497.

Smith, P. L. (2000). Stochastic dynamic models of response time and accuracy: A foundational primer. *Journal of Mathematical Psychology*, 44(3):408–463.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review*, 84(4):353.

Sternberg, R. J. (1985). *Beyond IQ: A Triarchic Theory of Human Intelligence*. Cambridge University Press, Cambridge, England.

Stocking, M. L. (1996). An alternative method for scoring adaptive tests. *Journal of Educational and Behavioral Statistics*, 21(4):365–389.

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

Toland, M. D., Sulis, I., Giambona, F., Porcu, M., and Campbell, J. M. (2017). Introduction to bifactor polytomous item response theory analysis. *Journal of School Psychology*, 60:41–63.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics*, 31(2):181–204.

van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, 72(3):287.

van der Linden, W. J. (2008). Using response times for item selection in adaptive testing. *Journal of Educational and Behavioral Statistics*, 33(1):5–20.

van der Linden, W. J. (2009). Conceptual issues in response-time modeling. *Journal of Educational Measurement*, 46(3):247–272.

van der Linden, W. J. and Glas, C. A. W. (2010a). *Elements of Adaptive Testing*, volume 10. Springer, New York.

van der Linden, W. J. and Glas, C. A. W. (2010b). Statistical tests of conditional independence between responses and/or response times on test items. *Psychometrika*, 75(1):120–139.

van der Linden, W. J. and Reese, L. M. (1998). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*, 22(3):259–270.

van der Linden, W. J. and Veldkamp, B. P. (2004). Constraining item exposure in computerized adaptive testing with shadow tests. *Journal of Educational and Behavioral Statistics*, 29(3):273–291.

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., and Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, 118(2):339.

Veldkamp, B. P. (1999). Multiple objective test assembly problems. *Journal of Educational Measurement*, 36(3):253–266.

Veldkamp, B. P. (2013). Application of robust optimization to automated test assembly. *Annals of Operations Research*, 206(1):595–610.

Veldkamp, B. P. (2016). On the issue of item selection in computerized adaptive testing with response times. *Journal of Educational Measurement*, 53(2):212–228.

Verhelst, N. D., Verstralen, H. H., and Jansen, M. (1997). A logistic model for time-limit tests. In *Handbook of Modern Item Response Theory*, pages 169–185. Springer, New York.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., and Mislevy, R. J. (2000). *Computerized Adaptive Testing: A Primer*. Routledge.

Wang, C., Chang, H.-H., and Douglas, J. A. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology*, 66(1):144–168.

Wang, C. and Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology*, 68(3):456–477.

Wang, T. and Hanson, B. A. (2005). Development and calibration of an item response model that incorporates response time. *Applied Psychological Measurement*, 29(5):323–339.

Whitley, B. E. (1998). Factors associated with cheating among college students: A review. *Research in Higher Education*, 39(3):235–274.

Wollack, J. A. and Fremer, J. J. (2013). *Handbook of Test Security*. Routledge, New York.

Wollack, J. A. and Maynes, D. D. (2016). Detection of test collusion using cluster analysis. In *Handbook of Quantitative Methods for Detecting Cheating on Tests*, pages 124–150. Routledge.

Wright, B. D. and Masters, G. N. (1981). *The Measurement of Knowledge and Attitude*. University of Chicago, Chicago.

Zara, A. and Pearson, V. (2006). Defining item compromise. In *annual meeting of the National Council on Measurement in Education, San Francisco, CA*.

Zhang, Y., Searcy, C., and Horn, L. (2011). Mapping clusters of aberrant patterns in item responses. In *annual meeting of the National Council on Measurement in Education, New Orleans, LA*.