

Alma Mater Studiorum Università di Bologna  
Archivio istituzionale della ricerca

Chance-Constrained Automated Test Assembly

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

*Published Version:*

Giada Spaccapanico Proietti, M.M. (2024). Chance-Constrained Automated Test Assembly. JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS, 49(1), 92-120 [10.3102/10769986231169039].

*Availability:*

This version is available at: <https://hdl.handle.net/11585/927913> since: 2024-01-11

*Published:*

DOI: <http://doi.org/10.3102/10769986231169039>

*Terms of use:*

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).  
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

**Giada Spaccapanico Proietti, Mariagiulia Matteucci, Stefania Mignani, Bernard P. Veldkamp. (2024). "Chance-Constrained Automated Test Assembly". *Journal of Educational and Behavioral Statistics*, Vol. 49, No. 1, pp. 92-120.**

The final published version is available online at:

<https://doi.org/10.3102/10769986231169039>

#### Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

**When citing, please refer to the published version.**

### Abstract

Classical automated test assembly (ATA) methods assume fixed and known coefficients for the constraints and the objective function. This hypothesis is not true for estimates of item response theory parameters which are crucial elements in test assembly classical models.

To account for uncertainty in ATA, we propose a chance-constrained version of the maximin ATA model, which allows maximizing the  $\alpha$ -quantile of the sampling distribution of the test information function obtained by applying the bootstrap on the item parameter estimation. A heuristic inspired by the simulated annealing optimization technique is implemented to solve the ATA model. The validity of the proposed approach is empirically demonstrated by a simulation study. The applicability is proven by using the real responses to the TIMSS 2015 science test.

*Keywords:* automated test assembly; uncertainty; chance-constrained; simulated annealing

### Chance-Constrained Automated Test Assembly

In educational measurement, tests should be designed and developed providing evidence of fairness, reliability, and validity (American Educational Research Association et al., 2014). To meet these requirements, a *test assembly* process should be employed to perform an optimal selection of items from an item bank. In addition to producing test forms that conform to the content and psychometric specifications, a test assembly process ensures that the resulting ability measurements can be trusted and interpreted in a transparent way. Moreover, it can produce comparable measurements in operational settings where various parallel versions of tests are needed. Furthermore, test assembly plays a crucial role in ability assessment as it lies at the basis of the entire test production process: from the earlier stages of item creation to the selection of items for building the test forms. In detail, the requirements of the final tests specified in the test assembly model not only determine the structure of the test forms but also define the composition of the item pool (Ariel & van der Linden, 2006), guiding the item writing process.

In the last decades, the simplified access to modern digital resources such as sophisticated item banking systems opened the possibility of improving the manual test assembly process through *automated test assembly* (ATA). The introduction of ATA dramatically improved the quality of the test forms and simplified the test assembly process, especially for large testing programs.

ATA differs from the manual process because the item selection is performed by optimizing mathematical models through specific software called *solvers*. Automation has brought many advantages over manual test assembly. First of all, a rigorous definition of test specifications reduces the need to repeat some phases of the test development. Secondly, ATA is the only way to find multiple optimal or near-optimal combinations of items starting from large item banks, despite the computational complexity of the task. Thus, ATA is fundamental to making measurements comparable while simultaneously reducing operational costs.

In ATA, mathematical optimization models such as 0-1 linear programming (LP) models (see van der Linden, 2005) are usually applied. These classical models use the item information functions (IIFs) as linear coefficients for the decision variables which are kept fixed throughout the entire optimization process. However, it is well known that the IIFs are derived from the item parameters estimated within the item response theory (IRT) framework. Consequently, the IIFs should be considered uncertain inputs in the ATA models. Many papers (e.g., Mislevy et al., 1994; Patton et al., 2014; Tsutakawa & Johnson, 1990; Xie, 2019; Zhang et al., 2011; Zheng, 2016) discussed the consequences of uncertainty in item parameters on several aspects of educational measurement, such as the accuracy of ability estimation. However, relatively few studies focused on this issue in the ATA research field. In particular, De Jong et al. (2009), Veldkamp (2013), Veldkamp et al. (2013), Veldkamp and Paap (2017), and Veldkamp and Verschoor (2019) proposed robust alternatives to the classical optimization models. These papers focus on the assembly of single test forms only.

In this paper, we propose incorporating the uncertainty in the optimization model for simultaneous multiple test assembly, which is the most applied and discussed ATA model in the literature (Ali & van Rijn, 2016; Debeer et al., 2017; van der Linden, 2005). In more detail, we suggest a test assembly model based on the chance-constrained (CC) approach (see Charnes & Cooper, 1959; Charnes et al., 1958), namely the *CCATA* model, by which the  $\alpha$ -quantile of the sampling distribution of the test information function (TIF) is maximized. The proposed model extends the classical maximin ATA model (van der Linden, 2005, p. 69-70). The sampling distribution of the TIF is obtained by applying the bootstrap technique (Bradley & Tibshirani, 1993) during the estimation of item parameters, i.e., the item calibration. In this way, we ensure that, independently of the calibration conditions, we have a high probability of having a certain, possibly low error in the ability estimation (or conversely, a high TIF). The main novelty of our model is to take into account the observed structure of uncertainty of the item parameters and, in this light,

produce optimal tests with the highest accuracy of the ability estimates. The validity of the proposal is assessed by comparing our method with other existing approaches in a simulation study.

For solving the *CCATA* model, we developed an algorithm based on simulated annealing (SA), a stochastic meta-heuristic proposed by Goffe (1996). The added value of this technique is represented by the possibility of handling large-sized models, characterized by many optimization variables and constraints, and non-linear functions. All the proposed algorithms have been coded in the open-source framework *Julia* (Bezanson et al., 2017) and are free to use as they do not rely on commercial software.

The paper is organized as follows. First, the key elements of IRT and ATA are reviewed. The following section discusses the issues arising from uncertainty in IRT and ATA models. Subsequently, an introduction to the CC approach for solving optimization problems with uncertainty is provided. Then, a CC version of the maximin ATA model is proposed. The retrieval of the TIF empirical distribution and the development of a heuristic based on SA for solving the model are discussed in the same section. Afterward, the results of a simulation study are presented in order to compare our proposal to the existing approaches solved by the *Cplex 12.10.0 Optimizer* (IBM, 2019). An application of our approach to real data taken from the 2015 Trends in International Mathematics and Science Study (TIMSS) data is shown. Some concluding remarks and suggestions for applying the *CCATA* model end the paper.

### **Item Response Theory and Test Assembly Models**

In educational and psychological measurement, IRT modeling provides several methods to estimate the item parameters. Intending to produce test forms with the highest accuracy in ability estimation, IRT is a solid foundation for ATA methods because the Fisher information function, which is a key object in test assembly, is derived from the item parameter estimates. Given an IRT model, once the items have been calibrated, it is

possible to evaluate how informative the test is at various ranges of the latent ability using the TIF, which is defined as the sum of the item Fisher information of all the items in the test (or the inverse of the variance of the maximum likelihood estimator of the ability  $\theta$ ). Hence, the TIF has very favorable properties: the additivity (i.e., the linearity) over the test items and its easiness of interpretation. Formally, for a given test with  $n$  items and ability  $\theta \in (-\infty, \infty)$ , the TIF is equal to

$$TIF(\theta) = \sum_{i=1}^n \mathbb{I}_i(\theta), \quad (1)$$

where  $\mathbb{I}_i(\theta)$  is the IIF for item  $i$  computed at  $\theta$ . Expressions for the IIFs can be easily derived within the framework of IRT. For example, if we assume binary response data, where the probability  $P_i(\theta)$  of item  $i$  endorsement follows the two-parameter logistic (2PL) model, the IIF of item  $i$  is equal to

$$\mathbb{I}_i(\theta) = a_i^2 P_i(\theta)(1 - P_i(\theta)) = a_i^2 \frac{\exp(a_i \theta + b_i)}{[1 + \exp(a_i \theta + b_i)]^2}. \quad (2)$$

The item parameters  $a_i$  and  $b_i$  represent the discrimination and the intercept for item  $i$ , respectively<sup>1</sup>.

From a general point of view, an ATA model is an optimization model consisting of an objective function to be maximized or minimized and a set of constraints to be satisfied. Specific objective functions may be related to psychometric features of the test, such as the maximization of the TIF at given cutoff scores, or to test content or other test requirements, such as the minimization of the total testing time. Examples of constraints include the test length, the restriction on the number of items of a certain type, test overlap, and so on. Altogether, they represent the test specifications, which should be defined in the standard form of Table 1 (van der Linden, 2005, p. 40), before being translated into an ATA model.

---

<sup>1</sup> The slope-intercept parametrization is used.

**Table 1***Standard Form of a Test Assembly Problem*

optimize	<i>Objective function</i>
subject to	
	<i>Constraint 1</i>
	<i>Constraint 2</i>
	$\vdots$
	<i>Constraint J</i>

Only one objective can be optimized at a time. If we have more than one function to optimize, some tricks can be applied to transform the objectives into constraints (Veldkamp, 1999), such as the maximin paradigm. On the other hand, there is no upper limit for the number of constraints, provided that the solver can handle the problem (Spaccapanico P. et al., 2020). If at least one combination of items that meets all the constraints does exist, then the set of these combinations is the feasible set; otherwise, if this set is empty, the model is said to be infeasible. The subset of the feasible set that optimizes the objective function represents the optimal feasible solution.

Tests can be assembled merely through the selection of appropriate items out of an item bank. One way to do so is to use mathematical programming techniques like 0-1 linear programming (LP) or mixed integer programming (MIP) models and optimize them with commercial solvers such as CPLEX (IBM, 2019) or Gurobi (Gurobi, 2018). Following the mentioned approaches, it is possible to assemble a set of tests that meet some (mostly linear) constraints maximizing their TIFs (see van der Linden, 2005). For example, given an item pool of size  $I$ , a commonly used objective for ATA models maximizes the TIFs of



$T$  tests at  $K$  ability points:

$$\text{maximize } \sum_{i=1}^I \mathbb{I}_i(\theta_{kt}) x_{it}, \quad \forall t, k, \quad (\text{objective}) \quad (3)$$

with  $t = 1, \dots, T$ , and  $k = 1, \dots, K$ .  $\mathbb{I}_i(\theta_{kt})$  is the IIF for item  $i$  at abilities  $\theta_{kt}$ , the set of ability points for which we want to control the shape of the TIFs, and  $x_{it}$  is a decision variable taking value 1 if the item  $i$  is assigned to test  $t$  and 0 otherwise. Depending on the application scenario, the  $K$  ability points may be chosen within a limited set of values around the mean of the population ability. A common choice is to maximize the TIF at  $\theta = 0$ , which is generally the population's average ability.

Since the model (3) has  $T * K$  objectives, it cannot be solved without resorting to multi-objective programming methods (Deb et al., 2016). Therefore, the maximin paradigm is applied. Within this setting, given an item pool of  $I$  items, the maximin approach allows to maximize the lower bound  $y$  of the TIFs, i.e., it maximizes the minimum observed TIF among all the tests. The maximin ATA model is specified by the following objective and set of constraints:

$$\text{maximize } y \quad (\text{objective}) \quad (4a)$$

subject to

$$\begin{aligned} \sum_{i=1}^I \mathbb{I}_i(\theta_{kt}) x_{it} &\geq y, \quad \forall t, k, \\ y &\geq 0, \end{aligned} \quad (4b)$$

where  $y$  is the lower bound for the TIF, so that all the considered TIFs are equal or higher than this value. In this way, the previous objectives are transformed in  $T \times K$  constraints, and only one objective appears in the model.

In order to describe the structure of the test forms, extra inequalities must often be added to the model due to security concerns. In fact, among others, it may be required to specify a minimum number of items in a given category (e.g. content domain or item type) and the item use among the test forms.

## Uncertainty in Test Assembly

In the classical context of test assembly, the optimization models used for item selection do not consider the uncertainty of the estimates of item parameters (van der Linden, 2005). For example, the maximin ATA model is based on the TIF, which appears in the objective function, being the goal of the optimization model. The TIF is the sum of the IIFs of the items in the test form and depends on the item parameter estimates, which are generally considered fixed quantities. Nevertheless, ignoring the uncertainty derived from the estimation process may lead to several issues, such as the misinterpretation of the psychometric properties of the assembled test forms. When the calibration algorithm produces biased estimates for the item parameters, the IIFs are not accurate enough, and, consequently, the TIF of the assembled test might be underestimated or overestimated. In Veldkamp et al. (2013), the authors found that, for large uncertainties, the decrease of information in robust test assembly can reach 37%. As a consequence, the perceived accuracy of ability estimates may be compromised. Mostly regarding the latter issue, a good test assembly model would consider the variation of item parameter estimates in order to build test forms in a conservative manner, i.e., it would produce tests with a maximum plausible lower bound of the TIF.

Several attempts to incorporate uncertainty in the test assembly models have been made, mostly by proposing robust approaches. Starting from the conservative approach of Soyster (1973), where the maximum level of uncertainty is considered for 0–1 LP optimization, De Jong et al. (2009) proposed a modified version, where one posterior standard deviation is subtracted from the estimated Fisher information to take the calibration error into account. This approach was also adopted in Veldkamp et al. (2013), where the consequences of ignoring uncertainty in item parameters are studied for ATA models. In addition, Veldkamp (2013) investigated the approach of Bertsimas and Sim (2003), who developed a robust method for LP models by including uncertainty only for some parameters in the assembly of linear test forms. More recently, Veldkamp and Paap

(2017) proposed to include the uncertainty related to the violation of the assumption of local independence in ATA for testlets. Finally, Veldkamp and Verschoor (2019) discussed robust alternatives for both ATA and computerized adaptive testing.

The mentioned ATA robust approaches consider the standard error of the estimates and a protection level  $\Gamma$  that indicates how many items in the model are assumed to be changed in order to affect the solution (Bertsimas & Sim, 2003). In this sense, the uncertainty is treated in a deterministic way, and, given  $\Gamma$ , the solution is adjusted by adopting a highly conservative approach, as standard errors are the maximum expression of uncertainty of the estimates.

A reasonable solution to the mentioned problems appears to be the use of chance-constraints (or probabilistic constraints). In fact, they are among the first extensions proposed in the stochastic programming framework to deal with constraints where some parameters are uncertain (Charnes & Cooper, 1963; Krokhmal et al., 2002).

### **Chance-Constrained Modeling**

The CC approach (Charnes & Cooper, 1959; Charnes et al., 1958) is a method for optimization problems with uncertainty, where a conservative parameter  $\alpha$ , the risk level, modulates the level of fulfillment of probabilistic constraints. The CC modeling has been deeply explored in the financial field, especially in risk management and reliability applications. In this context, the decision-maker must select a combination of assets for building a portfolio by maximizing their utility function (see Chen, 1973; Freund, 1956; Rockafellar & Uryasev, 2000, 2001; Scott Jr & Baker, 1972)

More recently, this problem was formulated in terms of percentiles of loss distributions, giving rise to the theory of chance-constraints originally proposed by Charnes and Cooper (1959).

Probabilistic constraints include parameters assumed to be randomly distributed and subject to some predetermined threshold  $\alpha$ , defined in the interval  $[0, 1]$ , controlling

their fulfillment. By modifying  $\alpha$ , it is possible to relax or tighten some constraints, modulating the level of the conservativeness of the model. To introduce the formal representation of a CC model, we start with the standard form of a mixed-integer optimization model:

$$\begin{aligned} & \max_{\mathbf{x}} \quad f(\mathbf{x}) \\ & \text{subject to} \quad g_j(\mathbf{x}) \leq 0 \quad j = 1, \dots, J \\ & \quad \mathbf{x} \in \mathbb{Z}^p \times \mathbb{R}^q, \end{aligned} \tag{5}$$

where  $f(\cdot)$  is the objective function to be optimized,  $g_j(\cdot)$  is the function expressing constraint  $j$ ,  $J$  is the number of constraints, and  $\mathbf{x}$  is the vector of  $p$  integer and  $q$  continuous optimization variables. Both  $f(\cdot)$  and  $g(\cdot)$  are scalar functions.

The optimization domain is  $D = \text{dom}(f) \cap \bigcap_{j=1}^J \text{dom}(g_j)$  and the set  $\mathbf{X} = \{\mathbf{x} : \mathbf{x} \in D, g_j(\mathbf{x}) \leq 0, \forall j\}$  is the feasible set, which means that a solution  $\mathbf{x}$  is feasible if it is in the optimization domain and it satisfies all the constraints. Thus, a chance-constrained reformulation of the optimization problem adds to model (5) the following set of  $H$  probabilistic constraints:

$$\mathbb{P}[g_h(\mathbf{x}, \boldsymbol{\xi}) \leq 0] \geq 1 - \alpha, \quad h = 1, \dots, H, \tag{6}$$

where  $\boldsymbol{\xi}$  is a vector of random variables, which represent the uncertain parameters. This formulation seeks a decision vector  $\mathbf{x}$  that maximizes the function  $f(\mathbf{x})$  while satisfying the chance-constraints  $g_h(\mathbf{x}, \boldsymbol{\xi}) \leq 0$  with probability at least equal to  $(1 - \alpha)$ .

CC models represent a fully customizable robust approach to optimization. However, although they were proposed in the 1950s, they are still hard to be solved. In fact, a major issue is the general non-convexity of the probabilistic constraints. Even though the original deterministic constraints  $g_h(\mathbf{x}, \boldsymbol{\xi})$  with non-random  $\boldsymbol{\xi}$  are convex, the respective chance-constraints may be non-convex. Moreover, the chance-constraints are usually intractable because the quantiles of the random variables are difficult or impossible to compute (see Nemirovski & Shapiro, 2006) or involve non-convex functions. Several

methods of approximating the chance-constraints have been proposed in the literature (see Ahmed & Shapiro, 2008; Kataria et al., 2010; Margellos et al., 2014; Song et al., 2014; Tarim et al., 2006; Wang et al., 2011).

### Chance-Constrained Automated Test Assembly

In order to develop a conservative approach that incorporates the uncertainty of item parameters into the ATA model, we propose a stochastic optimization approach for the maximin test assembly model based on the CC method. Under this approach, the TIF is not considered a fixed quantity but a random variable. As explained further on, the distribution of the TIF is retrieved by using the bootstrap technique. Whenever a maximin principle is applied, the CC model can be seen as a percentile optimization problem (Krokhmal et al., 2002). In fact, the probability in the inequality (6) is replaced by the  $\alpha$  quantile of the distribution function of  $g_h(\mathbf{x}, \boldsymbol{\xi})$ , and this quantile is maximized. In our case,  $\boldsymbol{\xi}$  is the vector of the IIFs, and the function  $g(\cdot)$  is the summation over items.

By considering the maximin model (4a), the constraints (4b) involved in the maximization of the TIF are replaced by the chance-constrained equivalents as follows

$$\mathbb{P} \left[ \sum_{i=1}^I \mathbb{I}_i(\theta_{kt}) x_{it} \geq y \right] \geq 1 - \alpha, \quad \forall t, k, \quad (7)$$

where  $t = 1, \dots, T$  are the tests to be assembled, and  $\theta_{kt}$  are the ability points at which the TIF of the test form  $t$  must be maximized, with  $k = 1, \dots, K$ . Usually, these points are chosen within a limited set of values around the mean of the population's ability.

Commonly,  $\theta = 0$  is chosen, i.e., the TIF is peaked at  $\theta = 0$ , so that the expected standard error of ability estimates at this ability point is reduced. Finally,  $\alpha$  is a real-valued variable defined in the interval  $[0, 1]$ . In the proposed approach, the chance-constraints are optimized independently of each other. We call the model (7) *chance-constrained maximin ATA*, or briefly *CCATA*. Again, the key element of this model is the information function assumed to be random.

The *CCATA* model maximizes the expected precision of the assembled tests in estimating the latent trait values of the test-takers at the predetermined ability points with a high confidence level if the  $\alpha$  is chosen to be close to zero. In probabilistic terms, we can say that the constraints in the model (4b) must be fulfilled with a probability of at least  $(1 - \alpha)$ . By adjusting the confidence level  $(1 - \alpha)$ , it is possible to relax or tighten the attainment of the chance-constraints to reflect a specific conservative extent, e.g., a small  $\alpha$  means a high level of conservativeness. On the contrary, a large  $\alpha$  means an almost complete relaxation of the constraints. The introduction of a confidence level is one of the most relevant novelties of the *CCATA* model compared to the robust approach proposed by Veldkamp (2013) and Veldkamp et al. (2013), who, instead, performed a worst-case optimization.

Once the chance-constraints have been defined, a method to compute the probability appearing in the inequality (7) should be found. A possible solution is to make assumptions on the probability distribution of  $\xi$ , such as the multivariate normal (Kim et al., 1990). For example, Ahmed and Shapiro (2008) try to approximate the probability distribution using samples of the random variable of interest by a Monte Carlo simulation, a specific case of scenario generation<sup>2</sup> where all the scenarios have the same probability of occurrence. We decided to use the Monte Carlo method because of its flexibility and adaptability to our problem.

The proposed *CCATA* model for ATA is based on the empirical distribution of the TIFs of the assembled tests. Therefore, our random variable is the TIF of a test form. This statistic depends on the uncertain IRT item parameter estimates, such as the discrimination and the intercept. There are different ways to retrieve the distribution function of the TIF: given the standard errors of the estimates, the samples can be uniformly drawn from their confidence intervals as in the robust approach of Veldkamp (2013); otherwise, if a Bayesian

---

<sup>2</sup> The idea of scenario generation is to sample a finite number of values, the scenarios, from a reasonable distribution of  $\xi$ .

estimation is carried out, the samples in the Markov chain can be used.

In this paper, another approach is used: a bootstrap procedure is performed to resample the response data and obtain a batch of estimates for each item parameter (see next subsection). At the end of this phase, the IIF for all the items in the pool is computed at predefined ability points using the bootstrapped samples. These quantities are then used in the *CCATA* model to compute the  $\alpha$ -quantiles of the TIFs, and the model is optimized by looking for the combination of items that compose the test forms with the highest quantiles. A percentile optimization model would maximize a reasonable lower bound of the TIF: its  $\alpha$ -quantile, approximated by the  $\lceil \alpha R \rceil$ -th ranked value of the TIF computed on the  $R$  bootstrap replications of the item parameter estimates. The following sections explain the details of the retrieval of the TIF empirical distribution function by the bootstrap and the heuristic proposed to solve the model.

### **Empirical Measure of the TIF**

The test forms built using the *CCATA* model should have the maximum possible empirical  $\alpha$ -quantile of their TIFs. The optimality in this sense will ensure that the assembled tests are conservative in terms of accuracy of ability estimation (indeed, the TIF), taking into account the uncertainty in the item parameter estimates. A standard approach to extract the uncertainty could be to sample many plausible values of the item parameters from the confidence intervals built using the standard errors and, subsequently, compute the related IIFs at  $\theta$  target points. The latter may be an optimal starting point to assemble robust tests (see Veldkamp, 2013; Veldkamp et al., 2013), but it has its own downsides as a uniform interval of plausible values is assumed. Another attempt to account for the influence of sampling error in the Bayesian framework has been made by Yang et al. (2012). They proposed a multiple-imputation approach with the aim of better measuring the latent ability of a respondent.

Our approach is based on bootstrapping the calibration process. In particular, the

observed vectors of responses coming from the full sample (one vector for each test-taker) are resampled with replacement  $R$  times, and the item parameters are estimated for each sample. In this way, it is possible to preserve the natural relationship of dependence between the items, and, given the ability targets, it is possible to compute their IIFs. After that, given a set of items, we can build a test form and compute its TIF for each of the  $R$  replications. The resulting sample constitutes the empirical distribution function of the TIF.

More formally, let  $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_R$  be an independent and identically distributed (iid) sample of  $R$  realizations of the random vector  $\boldsymbol{\xi}$ , and  $\hat{F}_R := R^{-1} \sum_{r=1}^R \Delta \boldsymbol{\xi}_r$  be the respective empirical measure. Here  $\Delta(\boldsymbol{\xi})$  denotes the measure of mass one at point  $\boldsymbol{\xi}$ , i.e.,  $\Delta \boldsymbol{\xi}_r(A) = 1$  if  $\boldsymbol{\xi}_r \in A$ . Hence  $\hat{F}_R$  is a discrete measure assigning probability  $1/R$  to each sample. In this way, we can approximate the probability in the left-hand side of the inequality (7) by replacing the true cumulative distribution function of  $\boldsymbol{\xi}$  with  $\hat{F}_R$ .

### The Approximated Model

The retrieved empirical distribution function of the TIF is now incorporated into the *CCATA* model in the following way. Let  $\mathbf{1}_{(-\infty, 0]} \{x\} : \mathbb{R} \rightarrow \{0, 1\}$  be the indicator function of  $x$  in the interval  $(-\infty, 0]$ , i.e.,

$$\mathbf{1}_{(-\infty, 0]} \{x\} = \begin{cases} 0, & \text{if } x > 0 \\ 1, & \text{if } x \leq 0. \end{cases} \quad (8)$$

Thus, given a specific chance-constraint  $h$ , a known set of optimization variables  $\mathbf{x}$  and samples  $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_R$  of our random vector, we can rewrite

$$\begin{aligned} \mathbb{P}[g_h(\mathbf{x}, \boldsymbol{\xi}) \leq 0] &= \mathbb{E}_F [\mathbf{1}_{(-\infty, 0]} \{g_h(\mathbf{x}, \boldsymbol{\xi})\}] \\ &\approx \mathbb{E}_{\hat{F}_R} [\mathbf{1}_{(-\infty, 0]} \{g_h(\mathbf{x}, \boldsymbol{\xi})\}] \\ &= \frac{1}{R} \sum_{r=1}^R \mathbf{1}_{(-\infty, 0]} \{g_h(\mathbf{x}, \boldsymbol{\xi}_r)\}. \end{aligned} \quad (9)$$



Equation (9) means that the chance-constraint is approximated by the fraction of the  $R$  bootstrap samples in which  $g_h(\mathbf{x}, \boldsymbol{\xi}_r) \leq 0$ .

Adopting the same principle to the left-hand side of the chance-constraints in the inequality (7), the *CCATA* model can be approximated by

$$\begin{aligned}
& \max_{\mathbf{x}} \quad y \\
& \text{subject to} \quad \frac{1}{R} \sum_{r=1}^R \mathbf{1}_{[y, \infty)} \{ \vec{\mathbb{I}}_r(\theta_{kt})' \mathbf{x}_t \} \geq 1 - \alpha, \quad \forall t, k, \\
& \quad \quad \quad g_j(\mathbf{x}_t) \leq 0 \quad \quad \quad \forall j, t, \\
& \quad \quad \quad \mathbf{x}_t \in \{0, 1\}_I, y \in \mathbb{R}^+, \quad \quad \quad \forall t.
\end{aligned} \tag{10}$$

where  $\vec{\mathbb{I}}_r(\theta_{kt}) = \mathbb{I}_{1r}(\theta_{kt}), \dots, \mathbb{I}_{Ir}(\theta_{kt})$ . The following issues characterize model (10): it is non-convex because of the indicator function used in the chance-constraints (see Rockafellar & Uryasev, 2000, 2001, for the demonstrations), and commercial solvers do not well handle the indicator function. To overcome these problems, we propose to solve the model by the heuristic described below.

### The Heuristic

Since a linear formulation cannot effortlessly approximate the proposed *CCATA* model, a heuristic based on simulated annealing (SA) (Goffe, 1996) has been developed. This technique can handle large-sized models and non-linear functions. The theory of SA is derived from the physics of annealing substances. Briefly, we adapted the annealing process to our ATA model by replacing the random selection of a decision variable with the random selection of an item from the item bank. The perturbation of the decision variables is done by adding, removing or switching the chosen item with another available item. At each modification, the objective function is evaluated and the solution is accepted in accordance with an exponential function based on a parameter called temperature. The higher the temperature, the higher the probability of accepting a worse solution. The temperature is decremented until only better solutions are accepted. The way the temperature is

controlled is referred to as the cooling schedule. If there are no further improvements in the area (neighborhood) around the current solution, the process is stopped. At the end, the re-annealing phase is actuated if the global stopping criteria have not been reached. In this phase, the best solution obtained is perturbed, the temperature is heated (set to its initial value) and another area is explored. Consequently, more than one neighborhood of the solution space is explored by adopting the SA algorithm, avoiding being trapped in a local optimum. More information about the implementation of the SA algorithm can be found in Spaccapanico P. (2020) and in the pseudo code in Section A of the Appendix.

Unfortunately, the SA algorithm is not able to deal with the constraints, so they are incorporated into the objective function using the hinge function and the Lagrange relaxation, as in Stocking and Swanson (1993). Moreover, the SA has the disadvantage that it can hardly find the feasible space for a problem. Thus, we decided to start our heuristic with a fill up sequential phase: the worst performing test, both in terms of optimality and feasibility, is filled up with the best item available in the item pool. After the selected item has been assigned, the process is repeated until all the tests have reached their maximum length, i.e., they are all filled up. Once the first step is performed, we process the solution with the SA principle. The result of the heuristic is a set of solutions with a length equal to the number of neighborhoods explored. Finally, the solution with the best objective function is selected.

### Simulation Study

The performance and advantages of the *CCATA* test assembly model (10) are investigated through a simulation study. Our specific scenario is the on-the-fly test assembly for individualized testing. In fact, we will focus on the average examinee with  $\theta = 0$ , at which the TIF should be maximized. In this way, the estimation error of the population average ability is reduced. This setting allows us to evaluate the effects of using probabilistic methods in the field of ATA models and to control the conservativeness of the

produced tests. To assess under which conditions our proposal is preferable, the true TIFs of the tests assembled by our *CCATA* model are compared to those obtained with four alternative models under several conditions. The alternative models are: the classical maximin model (*classical*, see Equation (4a)), the mean minus 3 standard deviations model (*3sd*, Soyster, 1973), the mean minus 1 standard deviation model (*1sd*, De Jong et al., 2009), and the robust model (*robust*, Veldkamp et al., 2013). The mean and the standard deviations of the IIFs used in models *3sd* and *1sd* are computed on the bootstrap samples. For the *robust* model, the protection level  $\Gamma$ , which indicates how many items in the model should be changed in order to affect the solution, is set equal to 40, following the suggestion in Veldkamp et al. (2013), so 41 sub-models are solved, and the solution which produced the highest objective is retained.

All the models are solved using the `ATA.jl` Julia package (Spaccapanico P., 2021a). For the *classical*, *1sd*, *3sd*, and *robust* models, the CPLEX solver interfaced by `JuMP.jl`<sup>3</sup> is chosen. On the other hand, the *CCATA* model is solved by our heuristic. The data needed for assembling the CC tests consists of the sample of the IIFs computed at  $\theta = 0$ , for each item in the pool, namely the vector  $\vec{\mathbb{I}}_r(0)$ , for  $r = 1 \dots R$ . These quantities are obtained by estimating the item parameters by bootstrapping where the 2PL model is assumed. The item parameters  $a$  (discrimination) and  $b$  (intercept) are sampled from the following distributions:  $a \sim LN(0, 0.25)$ ,  $b \sim N(0, 1)$ .

The results are compared in terms of the true TIFs averaged across tests and replications. Other benchmarks used to compare the model performances are the relative BIAS and relative RMSE between the true and observed TIFs. The true TIF is the reciprocal of the real expected ability estimation error; higher values indicate that the test will produce on average more accurate ability estimates. Moreover, by comparing the values of the relative BIASes and RMSEs, we can evaluate the accuracy and conservativeness of the models under the specified conditions. In particular, the BIAS asserts if the observed

---

<sup>3</sup> <http://www.juliaopt.org/JuMP.jl/0.18/>

TIF underestimates (negative values) or overestimates (positive values) the true one. Moreover, as the RMSE approaches zero, the model's capability to estimate the true TIF increases. On the other hand, high absolute values of the RMSE and BIAS indicate that the observed TIF is not reproducing the real expected ability estimation error of the test.

## Simulation Design

The optimization has been performed on a personal computer with an AMD Ryzen 7 PRO 4750U processor and 16 GB of RAM. Two `Julia` packages have been used for the computational tasks: `Psychometrics.jl` for calibration and bootstrap (Spaccapanico P., 2021b), and `ATA.jl` for the ATA models (Spaccapanico P., 2021a). The steps addressed in the simulation study are described below:

1. A pool of  $I = 250$  true items with contents:  $\text{content\_A} = \{\text{type1}, \text{type2}, \text{type3}\}$ ,  $\text{content\_B} = \{\text{type4}, \text{type5}, \text{type6}\}$  is simulated.
2. For each replication  $m = 1, \dots, M$ , the responses of  $N = 3000$  subjects with  $\theta \sim N(0, 1)$  are generated. Then, the items are calibrated with the marginal maximum likelihood estimation approach with an unbalanced design of 500-1000 responses per item.  $M = 10$  replications are performed. To investigate the validity of the methods in multiple scenarios, we also implemented the cases  $N = 1200$  and  $N = 6000$ , where each item gets 200-400 and 2000-4000 responses, respectively.
3. The items are re-calibrated  $R = 500$  times on  $N^* = N$  respondents sampled with replacement (bootstrap). The  $\vec{\mathbb{I}}_r(0)$  for  $r = 1, \dots, R$  are computed.
4. The test specifications (see Table 2) are added to the models, and the optimization hyperparameters are set as explained in the next paragraph.
5. For each combination of sample size and set of test specifications, the models *classical*, *3sd*, *1sd*, *robust*, and *CCATA* are solved. The *CCATA* model is solved both with  $\alpha = 0.01$  and  $\alpha = 0.05$ .

Performing the bootstrap procedure on the item calibration and solving each ATA model is computationally intensive. In detail, each model requires about 500 seconds to approach its theoretical upper bound of the objective, and the bootstrap procedure takes about 6-7 hours, depending on the sample size.

## Test Specifications

The mentioned models are solved under different settings, such as the number of test forms and confidence levels. The assembly is performed in a parallel framework, i.e., the  $T$  tests must meet the same constraints. Two fictitious categorical variables, *content\_A*, and *content\_B*, with three possible categories each, are simulated to constrain the tests to have certain content validity. The following specifications replicate realistic ATA applications where feasibility is the main concern, along with the search for the optimal set of tests in terms of the TIF. The complete set of test specifications is summarized in Table 2.

**Table 2**

*Test Specifications*

Case	T	Max item use
1	10	4
2	10	2
3	20	4
4	25	4
Case	Variable	Bounds
All	Test length	[38, 40]
All	<i>content_A</i>	[6, 10], [9, 12], [18, 25]
All	<i>content_B</i>	[9, 12], [15, 19], [9, 12]
All	Maximum overlap between tests	11

For example, the constraints described for variable *content\_A* require that tests have 6 to 10 items of the first category of the variable *content\_A*, 9 to 12 items of the second

category, and so forth. For *classical*, *3sd*, *1sd*, and *robust* models, different combinations of the specifications in Table 2 create four cases to be investigated in increasing order of complexity. For the *CCATA* model, 8 cases are investigated (4 cases for each  $\alpha$  level).

Moreover, the hyperparameters for the heuristic are chosen as follows. The starting temperature is equal to 0.1, so the solver does not check solutions too far from the last explored neighborhood, while the geometric cooling parameter is set equal to 0.1. At the beginning of the optimization, we perform one fill up phase, only taking into account the feasibility of the model. Then, we proceed to look for the most optimal combination of items by randomly selecting one item in all the tests to be added, removed, or switched. A Lagrange multiplier equal to 0.1 is chosen to balance the model's feasibility and optimality. The amount of time needed to solve the model is imposed as the termination criterion, and it is set equal to 500 seconds. This stopping criterion is also valid for the other models.

## Results

In Table 3 and Figure 1, the mean of the true TIFs computed at  $\theta = 0$ ,  $\overline{TIF^\dagger(0)}$ , is reported. It is obtained by averaging the true  $TIF_{tm}^\dagger$  across the  $t = 1, \dots, T$  tests and  $m = 1 \dots, M$  replications, as follows:

$$\overline{TIF^\dagger(0)} = M^{-1}T^{-1} \sum_{m=1}^M \sum_{t=1}^T TIF_{tm}^\dagger(0). \quad (11)$$

Table 4 and Figure 2 show the results for the relative BIAS between the observed and true TIF, while Table 5 and Figure 3 for the corresponding relative RMSE. Relative measures are chosen to make the results of the different conditions comparable. The two indicators are obtained as follows. First, for each test  $t$  and replication  $m$ , the observed TIF,  $TIF_{tm}^{\dagger\dagger}(0)$ , and the true TIF,  $TIF_{tm}^\dagger(0)$ , are computed; then, they are averaged with respect to the  $T$  tests, getting  $\overline{TIF_m^{\dagger\dagger}(0)}$  and  $\overline{TIF_m^\dagger(0)}$ , respectively. Finally, BIAS and RMSE are computed as:

$$\text{BIAS} = M^{-1} \sum_{m=1}^M \left[ \left( \overline{TIF_m^{\dagger\dagger}(0)} - \overline{TIF_m^\dagger(0)} \right) / \overline{TIF_m^\dagger(0)} \right], \quad (12)$$

$$\text{RMSE} = \sqrt{M^{-1} \sum_{m=1}^M \left( \overline{TIF_m^{\dagger\dagger}(0)} - \overline{TIF_m^{\dagger}(0)} \right)^2 / \overline{TIF^{\dagger}(0)}}. \quad (13)$$

Clearly, the observed TIFs are different for each model. For example, the observed TIF for a particular test under the *CCATA* model corresponds to the  $\alpha$  quantile of its empirical distribution function. In contrast, for the *3sd* and *1sd* models, the observed TIF is the sum of the mean of the  $R = 500$  IIFs values obtained with the bootstrap, minus 3 or 1 bootstrap standard deviations, respectively. Finally, for the *classical* and *robust* ATA models, the observed TIF is the sum of the IIFs computed on the item parameters estimated on the full sample, following the classical approach.

**Table 3**

$\overline{TIF^\dagger(0)}$ , true TIF at  $\theta = 0$  averaged across  $T$  tests and  $M$  replications.

Case	<i>CCATA</i> ( $\alpha = 0.01$ )	<i>CCATA</i> ( $\alpha = 0.05$ )	<i>classical</i>	<i>3sd</i>	<i>1sd</i>	<i>robust</i>
$N = 1200$						
1	12.6274	12.7517	13.0457	9.8415	13.0052	12.9474
2	10.3974	10.4868	10.5739	9.4460	10.5651	10.5706
3	9.7330	10.2992	10.3712	9.3373	10.3727	-
4	9.4686	9.4603	8.9815	8.9878	8.9815	-
$N = 3000$						
1	13.4907	13.5446	13.5599	13.2208	13.5487	13.3787
2	10.6187	10.6286	10.6792	10.5404	10.6741	10.6781
3	10.5389	10.5506	10.3151	10.0580	10.6204	-
4	9.4715	9.4700	8.9815	8.9775	8.9815	-
$N = 6000$						
1	13.6271	13.3121	13.5403	13.4735	13.6565	13.6889
2	10.6664	10.5988	10.7347	10.7079	10.7310	10.7332
3	10.5452	10.3929	10.5362	10.5249	10.3805	-
4	9.4684	9.4516	8.9815	8.9815	8.9815	-

Note. The *robust* model could not be solved for cases 3 and 4.



**Table 4***Relative BIAS of the TIF.*

Case	<i>CCATA</i> ( $\alpha = 0.01$ )	<i>CCATA</i> ( $\alpha = 0.05$ )	<i>classical</i>	<i>3sd</i>	<i>1sd</i>	<i>robust</i>
<i>N</i> = 1200						
1	0.1191	0.1658	0.2248	-0.8973	-0.1553	0.2218
2	0.0239	0.0699	0.1228	-0.9298	-0.2429	0.1231
3	-0.0182	0.0604	0.1187	-0.9319	-0.2462	-
4	-0.0360	0.0148	0.0733	-0.9368	-0.2811	-
<i>N</i> = 3000						
1	-0.0063	0.0196	0.0667	-0.6081	-0.1448	0.0666
2	-0.0412	-0.0174	0.0312	-0.6725	-0.1847	0.0310
3	-0.0474	-0.0204	0.0263	-0.6819	-0.1852	-
4	-0.0752	-0.0449	0.0078	-0.7004	-0.2082	-
<i>N</i> = 6000						
1	-0.0211	-0.0032	0.0378	-0.4268	-0.1103	0.0374
2	-0.0369	-0.0219	0.0191	-0.4679	-0.1344	0.0192
3	-0.0408	-0.0246	0.0177	-0.4703	-0.1379	-
4	-0.0574	-0.0349	0.0064	-0.4901	-0.1494	-

Note. The *robust* model could not be solved for cases 3 and 4.

**Table 5***Relative RMSE of the TIF.*

Case	<i>CCATA</i> ( $\alpha = 0.01$ )	<i>CCATA</i> ( $\alpha = 0.05$ )	<i>classical</i>	<i>3sd</i>	<i>1sd</i>	<i>robust</i>
<i>N</i> = 1200						
1	0.1212	0.1674	0.2262	0.8980	0.1564	0.2235
2	0.0280	0.0721	0.1248	0.9297	0.2435	0.1251
3	0.0269	0.0624	0.1222	0.9318	0.2463	-
4	0.0382	0.0209	0.0757	0.9369	0.2813	-
<i>N</i> = 3000						
1	0.0207	0.0253	0.0693	0.6083	0.1455	0.0689
2	0.0443	0.0235	0.0356	0.6725	0.1853	0.0353
3	0.0507	0.0260	0.0343	0.6817	0.1857	-
4	0.0769	0.0475	0.0183	0.7005	0.2086	-
<i>N</i> = 6000						
1	0.0237	0.0121	0.0392	0.4269	0.1108	0.0391
2	0.0374	0.0229	0.0206	0.4679	0.1345	0.0208
3	0.0413	0.0256	0.0199	0.4703	0.1378	-
4	0.0578	0.0356	0.0098	0.4901	0.1496	-

Note. The *robust* model could not be solved for cases 3 and 4.

As can be noticed from Tables 3, 4, and 5, the *robust* model could not produce any solution under conditions 3 and 4. These models reached the termination criterion of 500 seconds before a feasible solution was found for all their sub-models. For this reason, the *robust* approach turned out to be impractical for complex, i.e., large-sized, ATA models. Specifically, large-sized ATA models are characterized by having several optimization variables and constraints. This condition occurs especially when overlap constraints are imposed, because many auxiliary optimization variables are needed to linearize the model.

Looking at Table 3 and Figure 1, the results on the mean TIF are very similar for all the approaches. However, some patterns have been detected and explained afterward. Lower values of the true TIFs are observed for the *3sd* model mainly for the smallest sample size and for  $N = 6000$ , in case 1. The *CCATA* model never produces the worst results and outperforms the other approaches in case 4, where the underlying ATA model is more constrained and has a higher number of decision variables. In particular, the configuration with  $\alpha = 0.05$  seems to behave slightly better than the configuration with  $\alpha = 0.01$ . Overall, we can say that our approach is stable and reliable. Also, the heuristic is able to find satisfying optimal solutions for our model.

Likewise, the relative BIASes and RMSEs shown in Tables 4 and 5 and depicted in Figures 2 and 3 are very interesting. As expected, the relative BIAS and RMSE tend to approach zero as the sample size increases for all the approaches. This behavior is more evident for the classical ATA model. Previous findings about the classical ATA maximin model are confirmed by this simulation. In detail, we observe that the mean TIF obtained with this method overestimates the mean true TIF for all the cases under inspection. The positive bias goes from 0.6%, if the responses per item are 1000 or 2000 ( $N = 6000$ ) to 22.48%, if each item gets from 200 to 400 responses ( $N = 1200$ ). This aspect highlights the importance of using a more conservative ATA model in order to keep the results interpretable. Otherwise, the expected measurement precision of the tests is overestimated as well. On the other hand, it is evident that the *3sd* and *1sd* models produce very low estimates of the true TIFs since these approaches are too conservative. For example, the *3sd* model always generates meaningless values since the negative relative BIAS never exceeds the 42%, even with the largest sample size. Moderately better results are obtained with the *1sd* model, but the negative difference is still not above the  $-11\%$ . In general, the *3sd* and *1sd* models tend to underestimate the true TIF in all the cases. The *robust* model produces quite accurate results, very similar to the *classical* model, but given its complex structure, it requires too much time to be solved. Thus, the *robust* model can be applied to

large-sized ATA models only if large investments in equipment or cloud computing are made. Instead, the *CCATA* model always tends to produce relative BIAS and RMSE close to zero and likely negative. In particular, the *CCATA* outperforms all the other approaches for  $N = 1200$  in all the cases in terms of BIAS and RMSE, turning out to be a powerful method when the sample size is small. The advantage of the *CCATA* solution compared to the other approaches is noticeable, especially for cases 2, 3, and 4. For  $N = 3000$ , the *CCATA* method outperforms the other approaches for cases 1, 2, and 3 in terms of BIAS and RMSE. However, with this sample size, the results are pretty similar, especially to the *classical* and the *robust* methods. Finally, for  $N = 6000$ , the *CCATA* solution outperforms the other approaches for case 1. The results are again very similar to the *classical* and the *robust* solutions.

The *CCATA* approach significantly improves the interpretation of the test's expected precision, which can be expressed as "the tests have a  $1 - \alpha$  probability of having a mean TIF higher than..." denoting the level of conservativeness of the solution. Furthermore, for the *CCATA* model, the relative BIAS decreases when  $\alpha$  passes from 0.05 to 0.01, showing that the risk level of the solution is, as expected, positively correlated with  $\alpha$  and hence, customizable. In other words, we could say that, to increase the probability of having a true TIF higher than the observed one, i.e., a lower risk level,  $\alpha$  should be decreased.

### Application to Real Data

The data used in this application come from the 2015 TIMSS survey, a large-scale standardized student assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA). Since 1995, this project has monitored mathematics and science achievement trends in 39 countries every four years, in the fourth and eighth grades and in the final year of secondary school. TIMSS 2015 was the sixth of such assessments. Further information regarding this study is available on the TIMSS 2015

web page. We selected the Italian sample of grade 8 students for the science test ( $n = 4479$ ). The greater availability of science items, compared to the mathematics ones, has driven the choice of the subject. The original item pool has been filtered, removing derived<sup>4</sup> and polytomous items, retaining only original binary items. The final dataset contains 234 items with the following categorical features:

- four content domains (69 Biology items, 57 Chemistry items, 58 Physics items, and 50 Earth Science items)
- three cognitive domains (98 Applying items, 88 Knowing items, and 48 Reasoning items)
- four topics (110 items with topic 1, 80 items with topic 2, 33 items with topic 3, 11 items with topic 4)

Furthermore, a subset of these items is grouped into 27 units.

The design is unbalanced, as students are given only a subset of the items, so missing values appear in the response data. In particular, each item has from 611 to 663 responses. The item parameters were estimated according to the 2PL model. After the calibration, we performed a non-parametric bootstrap with  $R = 500$  replications on the item parameters, and we computed the IIF at  $\theta = 0$  for all the items in the pool. The two already mentioned *Julia* packages `Psychometrics.jl` and `ATA.jl` were used for calibration, bootstrap, and test assembly tasks.

In the calibrated item pool, the discrimination parameter estimates range from  $1e-05$  to 4.708, with a mean of 0.920 and a median of 0.867. There are two items with the minimum allowed value of the discrimination estimate. On the other hand, the intercept estimates range from -4.340 to 4.546, with mean and median equal to 0.071 and 0.025, respectively.

---

<sup>4</sup> According to TIMSS technical report (Foy, 2016, p. 54), derived items are created combining two or more items.

The final matrix of the IIFs contains  $234 \times 500$  samples. Subsequently, we solved the *CCATA* model by using the proposed approach and imposing the following specifications in terms of test constraints, which were based on the features of the tests administered in the TIMSS 2015. In detail, a set of  $T = 14$  tests with length from 29 to 31 items is assembled. The already mentioned friend sets are included in the assembly as constraints. We imposed the tests to have at least 6 items for each content domain (Biology, Chemistry, Physics, and Earth Science), a minimum of 8 items in the Applying and Knowing cognitive domains, and a minimum of 7 items in the Reasoning cognitive domain. The first and the second topic must be present at least 10 times in each test form. Forms must contain at least 2 items on the third topic and 1 item on the fourth topic. Each item can be used in at most 3 test forms. The overlap must be less than or equal to 15 items between adjacent forms, 5 items between forms at a distance equal to 2 (e.g., forms 1 and 3 can have at most 5 items in common), and no overlap is allowed for the pairs at a distance greater than 2. For the *CCATA* model, we chose  $\alpha = 0.05$  and a Lagrange multiplier equal to 0.01. The last choice is motivated by the high level of infeasibility of the model. We excluded from the assembly 11 items that had an IRT  $b$  parameter higher than 3 or lower than -3. Removing items with extreme difficulty parameters helped the solver assemble the tests with a TIF peaked at  $\theta = 0$ .

After we included all the specifications in the model, we ran the optimization algorithm, which implements our heuristic. We selected the same termination criteria as in the simulation study. Before the time limit was reached, the algorithm explored 4 neighborhoods: the first and the second neighborhoods were not feasible, while the third and the fourth neighborhoods produced feasible tests with a minimum 0.05-quantile of the TIF equal to 4.55 and 4.84, respectively.

Thus, the best solution is produced within the last neighborhood, where the smallest 0.05-quantile among the tests is equal to 4.843. The assembled tests fulfill all the constraints, as shown in Table 6. Also, constraints on overlap and item use are satisfied.

**Table 6***TIMSS Data, Features of the Test Forms Assembled by the CCATA Model.*

Test ( $t$ )	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Length	29	29	29	30	29	29	29	29	29	30	30	29	29	29
Content Domain														
Biology	9	6	7	6	10	10	10	10	7	7	9	8	9	10
Chemistry	6	6	8	9	6	7	6	6	6	8	9	8	7	6
Physics	8	9	8	6	7	6	7	7	8	8	6	7	7	7
Earth Science	6	8	6	9	6	6	6	6	8	7	6	6	6	6
Cognitive Domain														
Applying	12	13	10	12	12	13	12	8	11	13	12	10	12	10
Knowing	9	8	12	11	9	9	10	12	11	9	11	11	9	11
Reasoning	8	8	7	7	8	7	7	9	7	8	7	8	8	8
Topic														
1	11	10	11	11	11	10	12	15	16	15	17	13	10	15
2	10	12	10	10	10	10	10	10	10	10	10	10	13	10
3	7	6	6	7	6	8	6	3	2	4	2	2	2	3
4	1	1	2	2	2	1	1	1	1	1	1	4	4	1

The maximized  $\alpha$ -quantiles together with the TIF at  $\theta = 0$  computed on the sample are reported in Table 7. A graphical representation of the sampling distributions of the TIFs is shown in Figure 4.

**Table 7***Test Information Function of the Assembled Tests for TIMSS Data at  $\theta = 0$ .*

Test ( $t$ )	$Q(TIF_t(0), 0.05)$	$TIF_t(0)$
1	4.856	5.157
2	4.844	5.166
3	4.895	5.243
4	4.861	5.175
5	4.999	5.325
6	4.878	5.178
7	4.896	5.276
8	4.856	5.259
9	4.868	5.243
10	4.861	5.175
11	4.870	5.286
12	4.907	5.355
13	4.880	5.308
14	4.853	5.185

The resulting TIFs and quantiles do not considerably differ among the test forms, this is a signal that the model reached an optimal solution which is very proximal to the global one. However, the high complexity of the model and the low values of the IIFs at  $\theta = 0$  contributed to low TIFs. In particular, we found that the TIFs of the assembled tests have their peaks in the interval  $\theta > 0$  (Figure 4), suggesting that the item bank is appropriate to measure the ability of examinees more proficient than the Italian ones.

Analyzing the sampling distribution of the TIFs of the assembled tests illustrated in Figure 4, we can notice that the TIF computed on the full sample is consistently higher than the 0.05-quantile. Thus, we could say that there is a low possibility that test 2



produces estimates of the ability of an examinee with a true  $\theta = 0$  with a standard error of measurement greater than  $\sqrt{1/4.843} = 0.454$ .

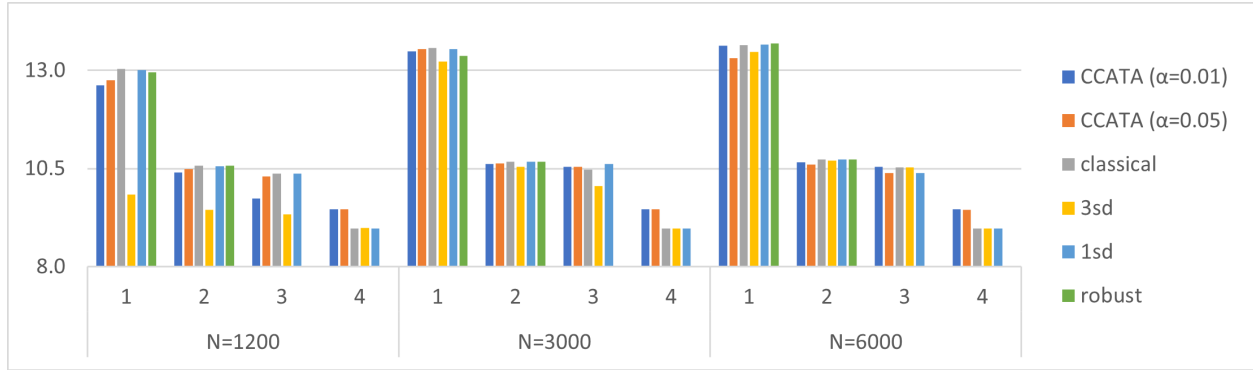
### Concluding Remarks

In this work, a chance-constrained version of the maximin ATA model, namely *CCATA*, has been introduced. This new test assembly model is able to deal with uncertainty in item parameters affected by calibration errors, which, in practice, can be relevant especially for small sample sizes where the classical approaches highly overestimate the true TIF. In particular, the proposed approach can take into account the structure of the uncertainty observed in the response data used in the calibration phase, with the aim of reducing the risk of misinterpreting the test accuracy in estimating the examinee's ability. This goal is achieved by approximating the distribution function of the TIF using the bootstrapped replicates of the item parameter estimates. The new model reformulates the classical maximin ATA model in a percentile optimization problem a sub-category of CC models. To deal with the non-linear formulation of the proposed *CCATA* model, we developed a heuristic based on the SA principle for finding the optimal conservative tests. In this way, unlike classical and robust optimization techniques, it is also possible to handle large-sized models.

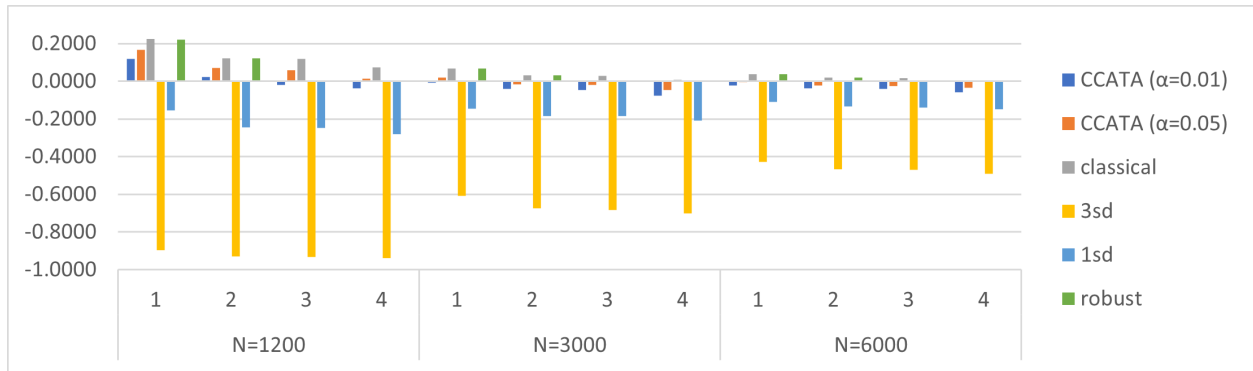
The results of a simulation study in the context of on-the-fly assembly for individualized testing show that the *CCATA* model, together with our heuristic, maximizes an adjustable conservative version of the TIF, i.e., its  $\alpha$ -quantile, where  $\alpha$  can be arbitrarily chosen from the test assembler. In particular, it has been empirically proven that these quantiles are lower bounds to the true TIF for small  $\alpha$ s, such as 0.05 or 0.01. Thus, using the sampling distribution function of the TIF along with the CC formulation gives a better idea of the accuracy of the tests in estimating future abilities and reduces the potential side effects of calibration errors. In contrast, with alternative methods, the observed TIF is often higher or excessively lower than the true one giving dangerous

misinterpretations. An application on real data from the TIMSS survey demonstrated that our approach is replicable in real-world situations.

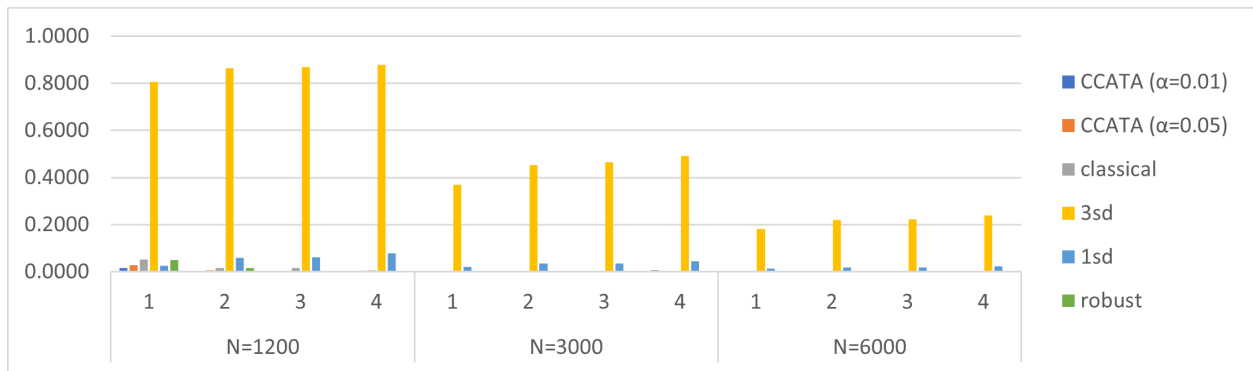
The results are encouraging, especially for complex and large-sized ATA models and for small sample sizes. A further contribution to the ATA research field is the development of two open-source *Julia* packages `Psychometrics.jl` and `ATA.jl` (Spaccapanico P., 2021a, 2021b), which do not rely on commercial solvers and can be used free of charge. However, further studies are needed to consider different test constraints and more Monte Carlo replicates. Moreover, unlike other robust ATA models, the CCATA model requires the availability of response data for the application of the bootstrap technique. To perform the study under these conditions, it would be useful to reduce the computational effort required for item calibration.

**Figure 1**

True TIF averaged across tests and replications. Plots are grouped by Case =  $\{1, 2, 3, 4\}$  and by  $N = \{1200, 3000, 6000\}$ .

**Figure 2**

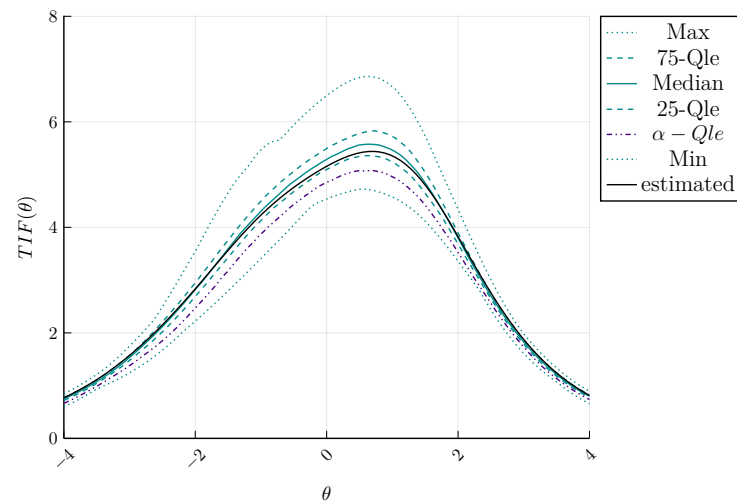
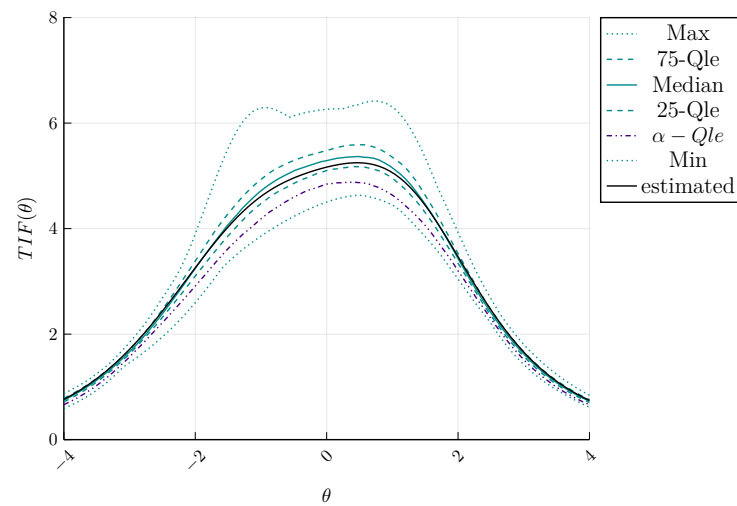
Relative BIAS of the TIF. Plots are grouped by Case =  $\{1, 2, 3, 4\}$  and by  $N = \{1200, 3000, 6000\}$ .

**Figure 3**

Relative RMSE of the TIF. Plots are grouped by Case =  $\{1, 2, 3, 4\}$  and by  $N = \{1200, 3000, 6000\}$ .

**Figure 4**

Examples of TIFs of the Assembled Tests 1 and 2. TIF Estimated on the Full Sample (solid black) against Quantiles.

*Test 1**Test 2*

## References

- Ahmed, S., & Shapiro, A. (2008). Solving chance-constrained stochastic programs via sampling and integer programming. In *2008 tutorials in operations research: State-of-the-art decision-making tools in the information-intensive age* (pp. 261–269). Informs.
- Ali, U. S., & van Rijn, P. W. (2016). An evaluation of different statistical targets for assembling parallel forms in item response theory. *Applied Psychological Measurement*, *40*(3), 163–179.
- American Educational Research Association, American Psychological Association, & National Council on Measurement Education. (2014). *Standards for educational and psychological testing*.
- Ariel, A., & van der Linden, B., W.J. and Veldkamp. (2006). A strategy for optimizing item-pool management. *Journal of Educational Measurement*, *43*(2), 85–96.
- Bertsimas, D., & Sim, M. (2003). Robust discrete optimization and network flows. *Mathematical Programming*, *98*(1-3), 49–71.
- Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98.
- Bradley, E., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall/CRC.
- Charnes, A., & Cooper, W. W. (1959). Chance-constrained programming. *Management Science*, *6*(1), 73–79.
- Charnes, A., & Cooper, W. W. (1963). Deterministic equivalents for optimizing and satisficing under chance constraints. *Operations Research*, *11*(1), 18–39.
- Charnes, A., Cooper, W. W., & Symonds, G. H. (1958). Cost horizons and certainty equivalents: An approach to stochastic programming of heating oil. *Management Science*, *4*(3), 235–263.

- Chen, J. T. (1973). Quadratic programming for least-cost feed formulations under probabilistic protein constraints. *American Journal of Agricultural Economics*, 55(2), 175–183.
- De Jong, M. G., Steenkamp, J.-B. G. M., & Veldkamp, B. P. (2009). A model for the construction of country-specific yet internationally comparable short-form marketing scales. *Marketing Science*, 28, 674–689.
- Deb, K., Sindhya, K., & Hakanen, J. (2016). Multi-objective optimization. In *Decision sciences* (pp. 161–200). CRC Press.
- Debeer, D., Ali, U. S., & van Rijn, P. W. (2017). Evaluating statistical targets for assembling parallel mixed-format test forms. *Journal of Educational Measurement*, 54(2), 218–242.
- Foy, P. (2016). *TIMSS 2015 user guide for the international database*. Boston College, Chestnut Hill, MA: TIMSS & PIRLS International Study Center.  
[https://timssandpirls.bc.edu/timss2015/international-database/downloads/T15\\_UserGuide.pdf](https://timssandpirls.bc.edu/timss2015/international-database/downloads/T15_UserGuide.pdf)
- Freund, R. J. (1956). The introduction of risk into a programming model. *Econometrica*, 24(3), 253–263.
- Goffe, W. L. (1996). SIMANN: A global optimization algorithm using simulated annealing. *Studies in Nonlinear Dynamics & Econometrics*, 1(3), 169–176.
- Gurobi. (2018). The gurobi optimizer [version 8.0].
- IBM. (2019). Ibm ilog cplex optimization studio [version 12.10.0].
- Kataria, M., Elofsson, K., & Hasler, B. (2010). Distributional assumptions in chance-constrained programming models of stochastic water pollution. *Environmental Modeling and Assessment*, 15, 273–281.
- Kim, C. S., Schaible, G. D., & Segarra, E. (1990). The deterministic equivalents of chance-constrained programming. *Journal of Agricultural Economics Research*, 42(2), 30–39.

- Krokhmal, P., Palmquist, J., & Uryasev, S. (2002). Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk*, 4, 43–68.
- Margellos, K., Goulart, P., & Lygeros, J. (2014). On the road between robust optimization and the scenario approach for chance constrained optimization problems. *IEEE Transactions on Automatic Control*, 59(8), 2258–2263.
- Mislevy, R. J., Wingersky, M. S., & Sheehan, K. M. (1994). Dealing with uncertainty about item parameters: Expected response functions. *ETS Research Report Series*, 1994(1), i–20.
- Nemirovski, A., & Shapiro, A. (2006). Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17, 969–996.
- Patton, J. M., Cheng, Y., Yuan, K.-H., & Diao, Q. (2014). Bootstrap standard errors for maximum likelihood ability estimates when item parameters are unknown. *Educational and Psychological Measurement*, 74(4), 697–712.
- Rockafellar, R. T., & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2, 21–42.
- Rockafellar, R. T., & Uryasev, S. (2001). *Conditional value-at-risk for general loss distributions*. ISE Dept., University of Florida.
- Scott Jr, J. T., & Baker, C. B. (1972). A practical way to select an optimum farm plan under risk. *American Journal of Agricultural Economics*, 54(4), 657–660.
- Song, Y., Luedtke, J. R., & Küçükyavuz, S. (2014). Chance-constrained binary packing problems. *INFORMS Journal on Computing*, 26(4), 735–747.  
<https://doi.org/https://doi.org/10.1287/ijoc.2014.0595>
- Soyster, A. L. (1973). Convex programming with set-inclusive constraints and applications to inexact linear programming. *Operations Research*, 21, 1154–1157.
- Spaccapanico P., G. (2020). <https://doi.org/10.6092/unibo/amsdottorato/9217>
- Spaccapanico P., G. (2021a). ATA.jl: Automated test assembly made easy [Computer software]. <https://github.com/giadasp/ATA.jl>

Spaccapanico P., G. (2021b). Psychometrics.jl [Computer software].

<https://github.com/giadasp/Psychometrics.jl>

Spaccapanico P., G., Matteucci, M., & Mignani, S. (2020). Automated test assembly for large-scale standardized assessments: Practical issues and possible solutions. *Psych*, 2(4), 315–337. <https://www.mdpi.com/2624-8611/2/4/24>

Stocking, M. L., & Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17(3), 277–292.

Tarim, S. A., Manandhar, S., & Walsh, T. (2006). Stochastic constraint programming: A scenario-based approach. *Constraints*, 11(1), 53–80.

Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter estimation on ability estimates. *Psychometrika*, 55(2), 371–390.

van der Linden, W. J. (2005). *Linear models for optimal test design*. Springer.

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., Matteucci, M., & de Jong, M. G. (2013). Uncertainties in the item parameter estimates and robust automated test assembly. *Applied Psychological Measurement*, 37(2), 123–139.

Veldkamp, B. P., & Paap, M. C. S. (2017). Robust automated test assembly for testlet-based tests: An illustration with analytical reasoning items. *Frontiers in Education*, 2(63), 1–8.

Veldkamp, B. P., & Verschoor, A. J. (2019). Robust computerized adaptive testing. In *Theoretical and practical advances in computer-based educational measurement* (pp. 291–305). Springer.



- Wang, Q., Guan, Y., & Wang, J. (2011). A chance-constrained two-stage stochastic program for unit commitment with uncertain wind power output. *IEEE Transactions on Power Systems*, *27*(1), 206–215.
- Xie, Q. (2019). *The impact of collateral information on ability estimation in an adaptive test battery* (Doctoral dissertation) [<https://doi.org/10.17077/etd.njvy-42a6>]. University of Iowa.
- Yang, J. S., Hansen, M., & Cai, L. (2012). Characterizing sources of uncertainty in item response theory scale scores. *Educational and Psychological Measurement*, *72*(2), 264–290.
- Zhang, J., Xie, M., Song, X., & Lu, T. (2011). Investigating the impact of uncertainty about item parameters on ability estimation. *Psychometrika*, *76*(1), 97–118.
- Zheng, Y. (2016). Online calibration of polytomous items under the generalized partial credit model. *Applied Psychological Measurement*, *40*(6), 434–450.

## Appendix

### The Heuristic Pseudo Code

The heuristic we developed is inspired by the work of Stocking and Swanson (1993) where the constraints are treated as part of the loss function using the Lagrange relaxation. The algorithm is based on the separability of the problem. This means that the fulfillment of the constraints and the objective function are computed separately for each test. In this way, it is possible to evaluate the optimality  $opt_t(\mathbf{x}_t)$ , and the feasibility,  $feas_t(\mathbf{x}_t)$ , at the test level and hence determine which test should be modified in order to obtain a better solution for the ATA model. To simplify the notation, we define the quality of a test as the sum of its optimality and feasibility weighted by the Lagrange multiplier  $\beta$ , i.e.,  $f_t(\mathbf{x}_t) = \beta opt_t(\mathbf{x}_t) - (1 - \beta)feas_t(\mathbf{x}_t)$ , where  $\mathbf{x}_t$  is the portion of objective variables related to test  $t$ . Given these quantities, the fulfillment of the ATA model global constraints is given by the sum of the feasibility of each test. In contrast, the global optimality depends on the objective of the ATA model. In the maximin ATA model, the global objective is the value of the lowest test information function among all tests.

Along with the iterations of the SA algorithm, each modification to the tests (e.g., an item is added to a test form) is accepted with a probability given by the Boltzmann factor  $\mathbb{P}[\Delta f] = e^{\frac{-\Delta f(\mathbf{x})}{T}}$  which is a function of the temperature  $T$  and of the variation,  $\Delta f(\mathbf{x})$ , of the global objective function produced by the test modification. In detail, the heuristic is described through the following pseudo code:

Define the constraints and the objective function of the ATA model.

Initialize the decision variables and objective function to zero.

Set the hyperparameters:

$G$ : number of neighborhoods to explore

$T_0$ : starting temperature

$g$ : temperature geometric decreasing rate

$\tau$ : maximum time

$C$ : convergence

$\beta$ : Lagrange multiplier (for balancing feasibility and optimality)

**for** g=1 to G **do**

**procedure** FILL UP

**repeat**

            Find and select the test with the lowest quality.

            Add the best item to the selected test, i.e., the item that increases test quality the most.

**until** All tests are filled up

**end procedure**

**procedure** SIMULATED ANNEALING

$c = 0$

**repeat**

            Find and select the test with the lowest quality.

**procedure** ADD ITEM

**for all** Items available in the item bank **do**

                Accept to add the item with probability equal to the Boltzmann factor.

**if** The add is accepted **then**

                    Go to CHECK CONVERGENCE

**end if**

**end for**

**end procedure**

**procedure** SWITCH ITEMS

**for all** Items already in the selected test **do**

**for all** Items available in the item bank **do**

                    Accept to switch the items with probability equal to the Boltzmann factor.

**if** The switch is accepted **then**

                        Go to CHECK CONVERGENCE

**end if**

**end for**

```

    end for
  end procedure

  procedure CHECK CONVERGENCE
    if The value of the objective function is increased then
       $c = 0$ , decrease temperature geometrically by  $g$ .
    else
       $c+ = 1$ 
    end if
  end procedure

  until ( $c < C$ )
  end procedure

  Increase the temperature to  $T_0$ 

  if Elapsed time is greater than  $\tau$  then
    Stop the algorithm.
  end if
end for

Save the best solution

```