New developments on automated standardized test assembly
Research project – Supervisor Prof. Stefania Mignani

Brief introduction
Especially in the field of educational measurement, computer-based testing is gradually replacing classical paper & pencil for test administration. For this reason, there is an increasing need to produce different parallel forms of standardized tests. Typically, manual or automated test assembly is used to select items to be included in the forms that meet some pre-specified targets. Classical automated test assembly (ATA) procedures resort to linear programming and mixed-integer programming and include only linear functions to express the specifications and the objectives at item and test levels. Most importantly, classical ATA models rely on the assumption of fixed or known variables for the item parameters coming from an item response theory (IRT) model estimation (see, e.g., Matteucci, Mignani, and Veldkamp, 2009) and, consequently, for the item information functions (IIF). Motivated by the need to overcome these limitations, the research project focuses on the inclusion of uncertainty in the ATA models and on the extension to potential nonlinear models.

Background and statement of the problem
Within the context of educational measurement, a test is a collection of items developed to measure students’ abilities. In order to make different measurements of the same ability, comparable tests should be standardized i.e. the test procedures must be fixed in such a way that differences among testing conditions do not influence the scores (Verschoor, 2007). Test assembly consists in selecting items from an item pool to build parallel test forms that meet some desiderata also called specifications. Test assembly procedures satisfy the need of fulfilling several specific requirements such as reducing the test length, maximizing the precision of the ability estimates, or building tests with the same difficult level. For these reasons, test assembly plays a fundamental role in test development.

Nowadays, several methods are used for test assembly based on either classical test theory (CTT) or item response theory (IRT). Larger testing organizations have better access to resources like sophisticated item banking systems, opening the possibility to improve their test assembly process by means of automated test assembly (ATA). ATA has several advantages over manual test assembly. First, the test specifications should be defined rigorously early on reducing the need to repeat some phases of the test development. More importantly, ATA is the only way to find optimal or near-optimal solutions starting from large item banks, for which manual assembly is not feasible due to the large number of possible combinations of items. As a consequence, ATA is fundamental
to make measurements comparable while reducing operational costs. From a practical point of view, the ATA models are not always easy to solve because they may involve a very large number of decision variables and constraints.

Modeling a test assembly problem implies the definition of objectives and constraints. The objectives deal with maximizing or minimizing some attributes, such as minimizing the gap between the test information function at some ability point to its target or maximizing the reliability of the test. On the other hand, the constraints impose a bound on test/item attributes, such as limiting the difficulty of the test, fixing the number of items having particular characteristics (e.g. type, domain) or considering enemy/friend sets. Decision variables represent the possible combination of items that compose the test form in a mathematical formulation. The standard form in these problems is an objective function to be optimized subject to a number of constraints, where the latter define a possibly feasible set of tests for a given item bank, and the former expresses our preferences for the tests in this feasible set.

Classical ATA models work with linear functions and resort to linear programming and mixed-integer programming (van der Linden, 2005). The main limitations of classical ATA models are: (a) the impossibility to define any part of the model in a non-linear manner, also with the standard form suggested in van der Linden (2005); (b) the assumption of fixed or known variables. With respect to (b), IRT item parameters, which describe the psychometric properties of the items in the bank, and subsequently the item information functions (IIF), which are a key element in ATA models, are assumed as known. This assumption is very strong as the IRT item parameters and the IIFs are estimated based on a specific IRT model.

These restrictions motivate the need to find alternative ways to specify and solve ATA models. Therefore, the problem addressed will deal with the inclusion of the sources of uncertainty into the ATA model, allowing for the use of nonlinear functions.

**Research questions or hypothesis, aim, objectives and deliveries**

The need of considering the shape of the uncertainty of the IIFs in ATA arises to ensure an almost sure reliability of the assessment in certain areas of the ability continuum. In this context, the main research question deals with the possibility to improve the item selection in ATA so that the test reliability and the precision of the ability estimates are increased by taking advantage of existing sources of uncertainty.

The first aim of the research project is to define a new approach to deal with structured uncertainty in the parameters of mixed-integer linear programming models applied to ATA, by considering possible nonlinear functions. This can be done exploiting the potential of the recent developments in the mixed-integer programming framework that have solid mathematical
properties, such as the Branch and Bound method when the model is not linear (e.g. with second-order conic constraints). Specifically we will refer to percentile optimization and models with chance-constraints.

Secondly, since there is the need to work on the range of variability of the IIF, several methods to retrieve its probability distribution will be investigated.

A suitable approach would consider the uncertainty associated to the estimates of the item parameters, and hence to the IIF, by incorporating the source of uncertainty directly into the optimization model used to assemble the tests. A promising category of models are the ones addressed in Kall and Wallace (1994). Moreover, this source must, somehow, be retrieved. Several methods are available, for example one can optimize the model allowing the item parameter estimates to range in an interval given by their standard errors (see Veldkamp, Matteucci, and de Jong; 2013) or, in a more structured way, following the posterior distribution estimated by a Bayesian model. Depending on which method is used the optimization model will be different.

The first part of the project will consist in a literature review on programming models that incorporate uncertainty, focusing on those applied to mixed-integer variables. In particular on those suitable for percentile optimization, such as the approach described in Delage and Mannor (2010). In fact, by using stochastic optimization models it is tried to ensure that, independently on the situation in which the calibration has been made, there is a high probability to have a certain error (possibly low) in the ability estimation. Digging into the recent studies which developed new algorithms for estimating item parameters will be also of primary importance because the calibration step should be integrated to the test assembly process.

The second step of the project will be devoted to find the best combination of calibration algorithms and test assembly models and coding them in a suitable programming language, and compare the results with the ones of a non-probabilistic optimization model, i.e. a classical ATA model. The chosen method will be tested by several simulation studies and by an application on testing real data, likely coming from the Italian National Institute for the Evaluation of the Education and Training system (INVALSI). In fact, starting from 2018 the same INVALSI adopted computer based testing for the administration of standardized tests at different grades that relies on ATA models.

Expected outcomes from the project are: (a) to show that the measurement properties of the assembled tests are better when the sources of uncertainty are taken into account in comparison to traditional ATA methods by using both simulated and real data; (b) to develop an open-source suite (package) written in a suitable programming language to assemble the tests according to the research developments.
**Participants in the study and the role they play**

The successful applicant will be part of the STAT department research group working on item response theory models and on the automated test assembly for the INVALSI tests (Stefania Mignani and Mariagiulia Matteucci). Other researchers from the STAT department with specific competencies (latent variable models, optimization methods, fuzzy mathematical programming) could be involved in the project as well.

Thanks to the three-years agreement between the STAT department and the INVALSI on the test assembly models for measuring student abilities (signed in 2018), the INVALSI team working on ATA will collaborate to the research project by providing ad hoc data and giving support in the interpretation of results.

Moreover, international collaborations are active with Prof. Dr. Bernard P. Veldkamp (Department of Research Methodology, Measurement and Data Analysis, University of Twente, The Netherlands) and Dr. Angela Verschoor (CITO, The Netherlands). Both the University of Twente and CITO are research hubs of excellence in the field of educational measurement and automated test assembly. The first results of the research collaboration with Bernard Veldkamp and Angela Verschoor have been presented at the 11th International Test Commission (ITC) Conference (Montreal, Canada, 2-5 July 2018) and will be presented at the International Association for Computerized Adaptive Testing (IACAT) Conference on 10-13 June 2019 in Minneapolis (USA).

**References**


Activity plan

The project is structured in the following steps:

1) Review of the literature about traditional automated test assembly methods, 0-1 linear programming, and recent extensions incorporating uncertainly working with mixed-integer variables.

2) Development of a proposal combining a proper calibration algorithm and a test assembly model including uncertainty. Comparison of the proposal to traditional non-probabilistic optimizations models by using simulation and real data studies.

3) Development of an open-source suite (package) written in a suitable programming language to assemble the tests according to the research developments.