HELPING TOOLS FOR ITEM BANK CALIBRATION AND DEVELOPMENT OF COMPUTERIZED ADAPTIVE TESTS

Javier López-Cuadrado, Anaje Armendariz, Tomás A. Pérez, Rosa Arruabarrena
University of the Basque Country (UPV-EHU)
San Sebastian/Spain
javilo@ehu.es, jiparlea@ehu.es, tomas@si.ehu.es, rosa.arruabarrena@ehu.es

Abstract

Recently-developed testing theories and current technology make possible a fast and easy generation of computerized adaptive tests (CATs), which emulate the intelligent behaviour of human evaluators. In fact, they dynamically select and administer the most appropriate items depending on the previous responses given by the examinees. However, to choose the proper item from the bank, the CAT algorithm needs to know the values of some psychometric characteristics (called parameters) that feature the items. This means that the item bank must be calibrated according to some psychometric model.

When the inexperienced user wants to design this kind of assessments, which are more efficient than traditional paper and pencil trials, three main problems arise: (1) the comprehension of statistical and psychometrical concepts, (2) the understanding of computers at a technical level, and (3) the need of many people, such as experts in the area and individuals to form the statistical samples needed during the calibration process. This paper presents a tandem of help tools, called CALLIE/GenTAI, which reduce the amount of knowledge required to effectively generate CATs. CALLIE guides the user through the different tasks of the item bank calibration process, helping her or him whenever making a decision is necessary, supporting the management of samples of individuals, and hiding, as much as possible, psychometrical concepts and statistical calculi at the same time. Once the item bank is calibrated, the user can employ GenTAI to administer dynamically generated CATs.

Keywords

On-line testing, calibrating tasks, help tools, test-administering tools, computerized adaptive testing.

1. INTRODUCTION

The developers of e-learning systems in general, and on-line language learning tools in particular, must be aware of the fact that every new learner has a different skill. Thus, it is necessary to place the incoming students at their stage, so they can progress properly as they interact with the e-learning system. Actually, an incorrect choice of the initial ability can make the student to be discouraged and lose interest.

There are many ways to estimate the initial skill for the student. For instance, one can let the learner choose it. In this sense, after having worked together with some Basque language teachers and experts, our experience tells us that, in most cases, our students usually selected levels below the adequate.

A second idea consists in interviewing every new student and estimating on the fly his or her knowledge. This option is suitable for learning environments with a limited number of learners and a concrete demand through time, but it might require too many resources (i.e. teachers to perform the interviews) to be feasible within an on-line e-learning system. An intermediate and common solution makes use of one or several tests of increasing difficulty that the student has to complete until he or she fails one. This test will obviously determine the starting point for the interaction with the system.

Nowadays, both alternatives are combined by the entry test to Hezinet/Boga [1], a commercial e-learning system for Basque language learning that is being used since 2000 in more than 60 Basque-adult schools. A set of 10 tests, each of them with a different difficulty, is available for the teacher,
who, after having a personal interview with the student, decides which of these tests will be administered first. When the correct responses ratio is excessively high or low, the new student has to complete several tests, as described before. We have detected that this method is appropriate for novice students, but not for latter-level learners. For instance, if one optimistically considers that a student uses 30 seconds to answer an item, the most proficient users can spend more than an hour to complete the admission tests in the worst case [2].

For this reason, we have developed GenTAI [3], an admission test generator that reduces, from about 120 to only 30 or 40, the number of items to be administered to the student, without losing accuracy and reliability. GenTAI implements Computerized Adaptive Tests (CATs) [4], which offer many advantages to those learning management systems that incorporate an assessment component; concretely, an increase of the security, a reduction of the time needed to pass the test and more precise estimations about the student's real ability. The earliest CAT systems were developed for high-stakes testing within standardized assessment programs, such as the pioneering Armed Service Vocational Aptitude Battery (ASVAB) [5], the Test Of English As a Foreign Language (TOEFL) [6] or the Graduate Record Examination (GRE) [7]. However, the number of CAT-based low-stakes test generators is growing day by day, and GenTAI is just a particular case.

CATs emulate the intelligent behaviour of human evaluators, since they dynamically select and administer the most appropriate items depending on the previous responses given by the examinee. After an initial trait level is assigned to the student, he or she must answer the set of items that dynamically are added to the test. During this period the ability of the student is progressively refined and subsequent items are selected to fit this level of aptitude. Thus, if a student, as expected, solves incorrectly an activity which is estimated to be difficult for the current ability estimate, then next item will probably be a little bit easier, so that the exact ability status will be more precisely estimated. The main idea is that each examinee must answer only items that are really informative for his or her ability, as questions that are too easy or too difficult result useless in this case. Therefore, every dynamically generated CAT will potentially be different, which makes the assessment process more individualized and secure.

To identify the proper item from the item bank during a CAT administration, the implemented algorithm makes use of some psychometric characteristics of the items, i.e. the parameters of an underlying model that is stated by the Item Response Theory (IRT) [8]. This theory provides very powerful techniques to carry out the evaluation, particularly when using CATs, but it imposes significant constraints. The most important one is that the item bank for the admission tests has to be calibrated, which means that some parameters of the evaluation items, such as their difficulty, must be estimated to fit the requirements of the IRT model being used. GenTAI uses the 3-parameter logistic (3PL) model [9], which features each item by means of the parameters of difficulty, discrimination and guessing factor.

The tasks required to achieve the calibration of an item bank are not complicated but they might consume many time and resources, especially if one wants to generate CATs but has limited means. Actually, obtaining the entry data sample for the item parameter estimation is a long process that requires the collaboration of several experts and/or hundreds of individuals, and a minimum knowledge of computing, psychometrics and statistics as well. GenTAI can smooth the technical needs since it automatically generates CATs just by being feeded with a calibrated item bank, while the help tool CALLIE allows any inexperienced user to perform item bank calibration processes.

The paper is organized as follows. Sections 2 and 3 discuss the most used procedures for calibrating an item bank, i.e. by experts and respectively by the psychometrical methodology. We have calibrated the item bank used by GenTAI to perform adaptive assessments on Basque language learners following both approaches [10], so it has been possible to detect common issues. In this sense, Section 4 presents the design of a general calibration process, in terms of tasks to be carried out, as well as the number and implication grades of the individuals needed for this purpose. Section 5 describes the system CALLIE, the calibration help tool that guides the user through the different tasks of the process, helping her or him whenever it is necessary to make a decision, supporting the management of samples of individuals, and hiding, as much as possible, any psychometrical concept or statistical calculus at the same time. Finally, Section 6 concludes with a discussion of future work.
2. CALIBRATION MADE BY EXPERTS

The calibration of an item bank consists in estimating values for the parameters that feature the items. When authoring and calibration processes are done using separate roles, it is usual to inquire one or more experts in the field about their personal, subjective, estimation of the parameters for the items. When using the experience of professionals it is recommendable to estimate only the difficulty of the item, since the other parameters (discrimination and guessing factor) can be extremely problematic to measure. In this case, one can use either the Rasch IRT model [11], which handles only the difficulty of the items, and can be considered as a variation of the already presented 3PL in which the discrimination power and the guessing factor are both constant.

The best way to obtain estimation for the difficulty of the items is to prepare several questionnaires that the experts should complete. Either if they are computerized or based on paper and pencil (p&p), [12] and even if the items are distributed among different questionnaires, it is preferable to get at least a minimum of 5 assessments per item [13].

Every questionnaire has to explain very clearly the objective of the task that the experts are ready to begin, as well as the instructions to perform it properly. For instance, they should be aware of what they are expected to contribute: should they give just the difficulty for every item contained in the survey or should they also solve them? What is the scale used to measure the difficulty? The inclusion of some examples is also highly recommended. There are more factors to consider before designing the questionnaires, such as the profile of the experts; if they are volunteers, for example, the amount of items to include in every sheet should not be excessively large, in contrast to the case where experts have been paid for their work. During the expert-based calibration of the item bank used to generate admission CATs for the Hezinet/Boga system, we asked several Basque teachers and philologists to complete a questionnaire [14]. Since they were volunteers, we tried to prepare forms that one could complete in less than one hour. However, the 15% of the experts who had agreed to take part in the calibration process abandoned [15].

Since the starting item bank contained 252 multiple-choice items, we distributed them among several different questionnaires, so that every expert had to answer only a part of the whole item bank. To help them, we included an equivalence-table containing the different Basque language level categorizations (HABE, IVAP, EOI, The Basque Government...). We asked the experts not only to locate each item in this difficulty framework, but also to give their correct response and to determine the skill they assess (verbs, syntax, declension, vocabulary, suffix, spelling, connectors, written expressions, or other). This way, the first thing we checked for each item was whether the experts had given the correct answer and also if they had agreed in the evaluated skill. Actually, some items were removed from the item bank because they presented significant differences among the experts.

At the end, several algorithms can be applied to estimate the values of the difficulty that will be assigned to an item, and at the same time to determine which items are valid. One of them is the maximum consensus algorithm, which, after gathering the opinions of a set of experts, shows them the results for those items with consensus, to finally let them discuss and change their minds when there is no clear agreement. It is also possible to use some estimators, such as the statistical mode or the mean. In the case of Hezinet/Boga, we combined both alternatives. Firstly, 55 items with too heterogeneous distribution of estimations were discarded. Later, we defined an ad-hoc estimator, which was similar to a bounded mean, to fix the difficulty level of each item. This estimator only utilized the most frequent estimations of difficulty given by the experts, so it avoided the influence of extreme judgements (i.e. the outliers values) as it favoured the consensus. Concretely, it calculated, for each item, the mean of the experts’ estimations that, summing at least the 75% of the gathered judgements, fell in a 30% of the whole scale.

3. PSYCHOMETRIC CALIBRATION

The psychometric calibration allows to obtain not only the difficulty of every item in the scale used by the IRT, but also their discriminative power and guessing factor. First of all, we need to collect the responses given by a large group of examinees that has to be representative of the population that will later use the final item bank [16]. To perform such a dense task (many items, many individuals), and also because of security matters, it is recommended to distribute the evaluation items into several test
forms (subtests) and apply them separately. The problem in partitioning both item and individual sets is that every subtest will be administered independently, that is, without any relationship with the rest. Therefore, the values of the item parameter estimates will not share a common scale; they will probably be identified in a different range for each test form. An anchor design can solve this situation [17]. The most typical approach consists in using different (not necessarily equivalent) groups, with the intention that each of them answers a different subtest, but having some items in common with other groups. Then, the estimates for the common items (which form the anchor item set) will be compared, providing the key to equate the different test form scales and, consequently, to get a common scale for the parameter estimates of the whole item bank.

Once the anchor design is ready, one can administer the subtests in p&p format or by computerized means. Each alternative has its advantages and inconveniences. Concretely, it can be easier to organize and supervise a p&p subtest administration, but it might require somebody to transcribe the results as they are collected to feed the statistical software. In any case, the next stage of the calibration process consists in carrying out some reliability analysis, which are intended to detect and rectify existing anomalies. At this point it is also usual to verify that the item bank is one-dimensional, in other words, to confirm that every item assesses the same (one and only one) latent trait.

After revising and debugging the response matrix, and eventually even removing some items from the bank (for example, because they do not satisfy the one-dimensionality constraint), one has to obtain statistical estimates for both item parameters and individual abilities, using as input the responses given to all the previously administered subtests. During the measurement of the model-to-data fit, one must confirm that the selected IRT model and the parameter estimates empirically fit. Concretely, it is necessary to verify that the estimated values correspond to the observed ones, that is, to those obtained during the administration stage. If the IRT model and the item bank do not match, then any IRT property is lost: information about the items will not be reliable and, as a result, one will not trust in the ability estimates provided by any CAT that is generated from the item bank. It is important not to forget that a CAT applies less items than a traditional test, so the effects of defective items can be critical. So, as a result of the model-fit assessment, it is very common some items to be removed from the bank because their characteristics (i.e. their parameter estimates) do not match the IRT model.

The calibration finishes with the equating process. At this moment, the scales that measure the item parameters will surely be different for each subtest, but, thanks to the anchor design, it is possible to use the anchor item set as a link to linearly transform these scales. As a consequence, the whole item bank will use a common scale that will be the same that states the ability estimates given by any CAT created from it [17].

During the IRT-based calibration of the Hezinet/Boga item bank, the set of 252 items was divided into 6 subtests, containing each of them 60 items, 22 of which were common to all the subtests. Each subtest was administered to a sample of at least 640 volunteers, thanks to a web-based tool (Fig. 1) that was developed for this purpose [18]. It not only allowed the researchers to manage and organize the administrations needed during the item bank calibration, but also stated the basis for the development of CALLIE, the item bank calibration component that will be discussed in section 5.

The application lied on a web-server that was on duty on a 24-7-365 basis (24 hours a day, 7 days a week, 365 days a year), thus everybody could visit the site and complete a Basque language subtest anytime, anywhere. The use of an identification code, rather than an access code, let us take advantage of the anonymous volunteers that unselfishly wanted to complete a test form. In order to know if the administrations had been carried out in acceptable conditions, we validated the identification codes by telephone or e-mail, and then decided whether admit them or not. At the end, a total of 3976 subtests were completed, 2268 of which correspond to supervised sessions, 976 to non-supervised but validated administrations, and 732 to test forms that have been refused. Besides rejecting those non-supervised administrations that could not be confirmed, we decided to discard test forms accomplished in more than 50 minutes, those finished in less than 5 minutes, and those that included at least one item that had taken more than 200 seconds to be answered.
4. THE GENERAL CALIBRATION PROCESS

After have performed both expert-based and psychometric calibrations for the Hezinet/Boga item bank, we have detected that they have many points in common. Thus, we present a unique serialization of the process, which is implemented by CALLIE and shown in Fig. 2, and includes also a third type of calibration: the manual mode, in which the responsible of the calibration (RoC), that is, the user of CALLIE responsible for the process, enters the parameter estimates directly into the system. This situation may occur, for instance, if the item bank has been already calibrated by other means.

The RoC will need a different number of involved individuals, whose number and profile will vary depending on the type of calibration. Thus, during an experts-based calibration on-line experts will give their estimations through the web, while off-line experts will complete p&p traditional questionnaires. In the same way, during a psychometric calibration, one can distinguish among on-line and off-line examinees.

First of all, the RoC has to establish the set of items that will be calibrated, and then select the calibration mode to be used.

- For a manual calibration (Fig. 2), there is only one task for the RoC, consisting in typing the values for each item parameter.

![Fig. 2. Manual calibration process.](#)
The first task for an expert-based calibration (Fig. 3) is the preparation of the questionnaires. To design them correctly, the RoC has to consider the number of judgments needed for each item, the amount of experts available, and the quantity of items that every expert should assess. It is clear that these values must fit, since, for instance, a calibration of a bank containing 1000 items where every expert will not deal with more than 100 items is unfeasible, unless 10 or more experts are available.

For a psychometric calibration (Fig. 4), the first task is the planning of the subtests and, in most cases, the anchoring design. As happens with the calibration based on expert judgement, in this case the subtests must be designed according to the number of responses needed for each item, the amount of examinees available, and the quantity of items every individual should answer.

Once the forms (questionnaires or subtests) have been arranged, the RoC has to determine how many of the involved people (experts or examinees) will fill them by the p&p method and how many will give their responses by a computerized way. In the former case, the RoC should feed the system with the collected data, while in the latter there will be no need for that task. Anyhow, when the system gathers all the data, the RoC will be able to filter them, as well as performing verifications such as those regarding the consensus among the experts, the established criteria for accepting a non-supervised subtest administration, or the analysis of unidimensionality of the item bank.

After having filtered the obtained data, the calibration process finishes with the estimates of the item parameters and, if an anchoring design has been defined for the psychometric mode, with their equating to a common scale.

The idea of generating CATs for both admission and assessment purposes can be very attractive to many teachers or administrators of e-learning systems. However, the calibration of the item bank, which is an essential and not trivial process, can be enough to ruin the initiative, because it requires some experience on technology, psychometrics and statistics, as well as specific software and human
resources management. We have designed CALLIE, a decision supporting tool designed to ease the tasks defined in the previous section to those people with no specific background. Having into account the profile of the final user, the main objective for the authors is to build a very easy-to-use tool. At the same time, the designed software is intended to be integrated within the Hezinet/Boga architecture, so it will probably offer not only a user interface but also a protocol for the communication between different subsystems.

![CALLIE Diagram](image)

**Fig 4. Expert-based calibration process**

To start, the item bank must be stored in a database. To make it, the RoC can use ADISTI [19], an easy-to-use software that allows the creation of item banks following the IMS Question & Test Interoperability in a transparent way for the user.

After the RoC selects the set of items to be calibrated, he or she must choose the calibration procedure. Then, the interaction with the system follows the schema described in Figs. 2, 3 and 4, as previously explained.

### 5. CONCLUSIONS

The advantages of computerized adaptive tests are well known. This type of tests could be more extended if there were fewer drawbacks when calibrating the items. We have presented the process needed for a decision support system to help people without much technical and psychometrical knowledge to complete that task. CALLIE is our prototype. Our plan is to add this system to ELSA, an e-learning system we have developed. Until now, we have used ELSA to cooperatively work with high schools around us assessing the student skills of Basque Language (the language they use on classes). So far, we have provided the calibrated items, and have generated reports about student skills, but our aim is that, sometime in the future, the high school staff could do either the calibration or the reports by itself using CALLIE.

Our presented work is centred in two aspects: (1) *online calibration*, where the system itself detects which items are not calibrated and administers them gradually to different users, in order to gather some data to be used later to calibrate those items; and (2) *offline calibration*, where the user
participates actively in the process of gathering those data. The former is a valid approach when a system has a volume of users, enough to complete the calibration. The latter, however, is intended to fill the gap when the number of users is already small.

CALLIE is in the test phase. We are now supervising how the system performs its decisions and how users understand the pieces of advice that CALLIE offers.

References


