Computerized adaptive testing, the item bank calibration and a tool for easing the process

Javier López-Cuadrado, Anaje Armendariz, Tomás A. Pérez,
Rosa Arruabarrena and José Á. Vadillo
University of the Basque Country (UPV-EHU)
Spain

1. Introduction

Many educational systems are provided with a mechanism that assesses students' progress while acquiring knowledge. This is something critical to identify the success or the failure during the learning process. Moreover, the developers of e-learning systems in general, and online language learning tools in particular, have to be aware of the fact that every new learner has their own 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 level can discourage the student and cause them to lose interest.

Usually, tests formed by multiple-choice items are administered to estimate the initial skill of the students as well as to supervise, later, their learning improvement. This chapter will focus on Computerized Adaptive Tests (CATs) (Wainer, 2000), which emulate the intelligent behaviour of human evaluators. Actually, they dynamically select and administer the most appropriate items depending on the previous responses given by the examinees (i.e. those that really provide useful information about their ability). Just for being computerized, CATs offer many advantages over the traditional paper- and pencil (p&p) tests, primarily regarding to the conditions of application, control and processing of the responses. In addition, when tests' compilation is adaptive, the evaluation is more secure, the administration takes less time, the estimations are more precise and anxiety and frustration rates among the examinees are lower.

The concrete operation of different CAT generators might vary, but they will surely follow the steps described by the following algorithm (Thissen & Mislevy, 2000). Figure 1 schematizes the minimal functionalities that every CAT system must provide and shows how a general CAT works: after presenting the instructions to be followed during the assessment (1), the system takes an initial ability estimate (θ^*) and selects the first item to be shown (2); then, the examinee answers the item (3) and the ability estimate is updated by some maximum likelihood or Bayesian method (4); if the stop criterion is satisfied (5), then the CAT finishes and the final score is computed; otherwise, the algorithm selects a new item (1) (providing both maximum information for the provisional θ^* and fulfilling any defined constraint), and presents it to the examinee (3); after collecting the answer, the

ability estimate is updated (4), and so on. The cycle continues until the stop criterion is satisfied, something that, depending on the particular implementation of the algorithm, can happen, for instance, when a predefined number of items has been administered, if the error of the new estimate is smaller than a certain value, or a time limit has been reached.

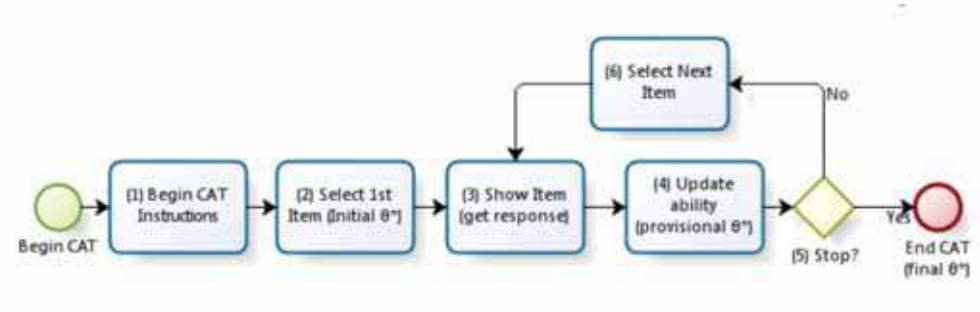

Fig. 1. Algorithm for the administration of a CAT.

To identify the proper item from the item bank during a CAT application, the implemented algorithm makes use of some psychometric characteristics of the items, i.e. the parameters of an underlying model that in most cases is stated by the Item Response Theory (IRT) (Lord, 1952). In spite of the fact that this theory provides powerful techniques to carry out the evaluation, particularly when implementing CATs, it imposes significant constraints. The most important constraint is that the item bank must be calibrated. This means that some parameters of the items must be settled to fit the requirements of the chosen IRT model. The most common models are the 1-parameter logistic model (1PL) (Rasch, 1960) and the 3-parameter logistic model (3PL) (Birnbaum, 1968). While the former characterizes every item by its difficulty, the latter also handles their discriminative power and guessing probability.
The calibration of an item bank consists in determining the values for the item parameters. Although the tasks involved in the process are not complicated, they might be time- and resource- consuming, especially if one wants to generate CATs but has limited means. Actually, obtaining the necessary entry data sample for the item parameter estimation is a long process. Moreover, the calibration might require the definition of an anchoring design to divide the bank into subtests, as well as the collaboration of several experts and/or hundreds of individuals. A minimum knowledge on psychometrics and statistics is also needed to manage and equate the scores of the administered subtests, to estimate values for the item parameters and to check their goodness-of-fit to the selected IRT model. Next two sections will explain the calibration processes done to obtain parameter estimates for the item bank that will feed a CAT generator for the assessment of the Basque language; section 2 is dedicated to the description of the process based on a set of experts that was conducted to find the difficulty of the items, whereas section 3 discusses the statistical calibration carried out to model the item bank in terms of the 3PL model.
The earliest CAT systems were developed for high-stakes testing within standardized assessment programs, such as the pioneering Armed Service Vocational Aptitude Battery (ASVAB) (Segall & Moreno, 1999), the Test Of English As a Foreign Language (TOEFL) (Wainer & Wang, 2000) or the Graduate Record Examination (GRE) (Mills & Steffen, 2000). However, the number of CAT-based low-stakes test generators is growing day by day. In

fact, the authors have recently completed the calibration of an item bank that will be used within the admission CAT generator included within ELSA, a Basque language e-learning system formerly known as Hezinet. As previously said, the calibration process has been implemented in two different ways: first, the 1PL difficulty parameter has been estimated by a set of experts, and next a statistical calibration process has been done to obtain estimates for the three parameters of the 3PL model.

This experience has been the starting point for the authors to formalize the general calibration process and to implement it in CALLIE. This computerized tool will allow pedagogues, science educators and CAT-generating system developers to calibrate their item banks easily, using the wizard incorporated in CALLIE and without requiring a specific background. CALLIE shows the existing relations among the variables of the process and how changes in one of them affect the others. Once the user has set up a calibration, the system automatically outlines its configuration, warns of some potential risks of doing some modifications (especially when some of the previous work could be lost) and advises of any useful information, such as that related to the calibration of an item bank by two or more different processes. Section 4 will present CALLIE, the help tool that assists during the calibration process, and section 5 will focus on the decision making that the responsible for calibration has to deal with. Next, section 6 will introduce the modules that form CALLIE's architecture. And finally, section 7 will summarize conclusions and future work.

2. Expert-based calibration

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.

While using the experience of professionals it is recommendable to estimate only the difficulty of the item, since the other parameters (discrimination and guessing probability) can be problematical to measure. In this case, one should use the 1PL model, which handles only the difficulty of the items and can be considered as a variation of the 3PL in which the discrimination power and the guessing factor are both constants.

One way to obtain estimation for the difficulty of the items is to prepare several questionnaires that experts will complete. Either if they are computerized or based on p&p (Arruabarrena & Pérez, 2005), and even if the items are distributed among different questionnaires, a number of judgements per item will have to be gathered.

Not only the number of judgements per item but also other decisions have to be made. 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 complete 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; being voluntary or not is significant as it determines the length of the questionnaire, to be precise, the number of items that each questionnaire will contain and, therefore, the time that each expert will spend answering it. Additionally, the developers will have to

decide how and when capture the experts and the period of time by which the latter ones will have to return the completed questionnaires.
Apart from the decision-making, while elaborating the questionnaires it is advisable to carry out two quality controls: one to assure de validity of the items and the other to verify the validity of the questionnaires. The former should be carried out before distributing the items among questionnaires. For this purpose, qualified personnel will review the items and will correct mistakes. The latter control should be performed once the questionnaires have been compiled and before delivering them. At this point, some pilot tests can be done to identify misleading material and to verify that the required fulfilment time adjusts to the previsions.
During the expert-based calibration of ELSA's item bank, some experts were recruited to assess the difficulty of 252 multiple-choice items. The experts were philologists and Basque Language teachers as well. At the end of the process 17 assessments per item had been gathered. However, some assessments were discarded because they did not adjust to the questionnaire requirements (the experts had chosen various options, or they had chosen an option out of the given choices). An ad-hoc estimator, similar to a bounded mean, was defined to fix the difficulty level of each item. This estimator only uses the most frequent estimations of difficulty given by the experts and, therefore, it avoids the influence of extreme judgements and favours the consensus.

3. Statistical calibration

The psychometric calibration allows obtaining 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, one needs to collect the responses given to the items by a large group of examinees that has to be representative of the population that will later use the final item bank. 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 (called 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 (Kolen & Brennan, 1995). The most typical approach consists in using different (not necessarily equivalent) groups of individuals, with the intention that each of them answers a different subtest, but having some items in common with the rest of the 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.
The next stage of the calibration process consists in carrying out some reliability analyses, 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, to be precise, 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 to remember that a CAT applies fewer items than a traditional test, so the effects of defective items can be critic. So, as a result of the model-fit assessment, it is very common to remove some items from the bank because their characteristics, specifically, 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 (Kolen & Brennan, 1995).

During the IRT-based calibration of the ELSA 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 540 volunteers, thanks to a web-based tool (Figure 2) that was developed for this purpose (López-Cuadrado, Armendariz et al., 2005). 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 presented in the following section.

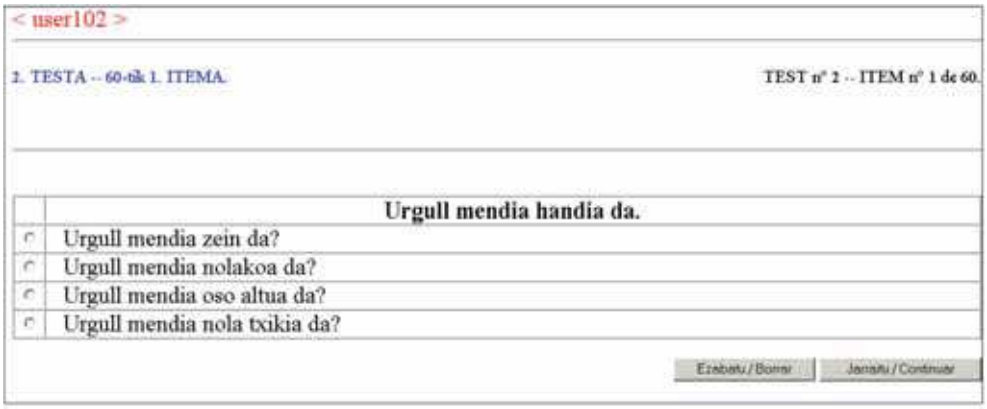

Fig. 2. Administration of a subtest by the web-based supporting tool.

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 the authors 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, the identification codes were validated by telephone or e-mail, and then the responsible for calibration decided whether admit them or not. At the end, a total of 3976 subtests were completed, 2268 of which corresponded to supervised sessions, 976 to non supervised but validated administrations, and 732 to test forms that had been refused. Besides rejecting those non supervised administrations that could not be confirmed, the authors 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 longer than 200 seconds to be answered.

4. The help tool CALLIE and the general calibration process

As stated above, the calibration of an item bank is a process that requires some knowledge on pedagogy, statistics, psychometrics and computer science, which is a constraint particularly difficult to fulfil for most of the people potentially interested in conducting such a process. That is why, after having calibrated the item bank for the admission CAT that will be used in ELSA, the authors have detailed and automated the whole procedure. The result is CALLIE, a help tool that is intended to guide the responsible for calibration during the process, from supporting the making of decisions needed to set up a calibration procedure to estimate the item parameters. The system focuses on the idea of receiving requests, so it will act as a response to those requests. The user will be able to supervise the tasks done by CALLIE, as well as to make some decisions and modify any requested data. Whenever the user has to make a decision, the system will notify the consequences of any choice beforehand. The system is conveniently ready to help somebody who has never conducted an item bank calibration before and who is neither psychometrician nor a computer scientist. Actually, the main target audience is high school, science or language-school teachers.

There may be several ways of performing the calibration, but at the present moment CALLIE implements only the two followed during the calibration of the item bank for the assessment of the Basque language, leaving for future releases any other existing alternatives. So CALLIE supports both expert-based and statistical calibrations. Regardless of the particular calibration process followed, the objective is always to obtain estimations for the parameters that characterize the items of the bank, which in this context are identified by an IRT model. Thus, if the 3PL model is selected, the system will recommend a psychometric procedure in order to obtain reliable item parameter estimates. On the other hand, if the user wants to calibrate an item bank in terms of the 1PL model, then both expert-based and statistic procedures will be available, and the responsible for calibration will have to choose one of them.

Due to the fact that the whole calibration process consumes many resources, it is crucial to plan in detail all involved tasks in advance. In both statistical and expert-based calibrations, the whole process development can be divided into two consecutive stages: the first one will be devoted to the data-gathering and the next stage to the elaboration of the calibration itself using the previously filtered data sample. No matter which type of calibration is selected, CALLIE follows a similar procedure, which is shown in Figure 3: first, the item bank must be prepared; at the same time it is necessary to arrange the system for the item

administration to facilitate the application of items to a set of experts or to a big sample of individuals; once the responses are gathered, the next step consists in conducting some preliminary analyses, to finally get the definitive item parameter estimates.

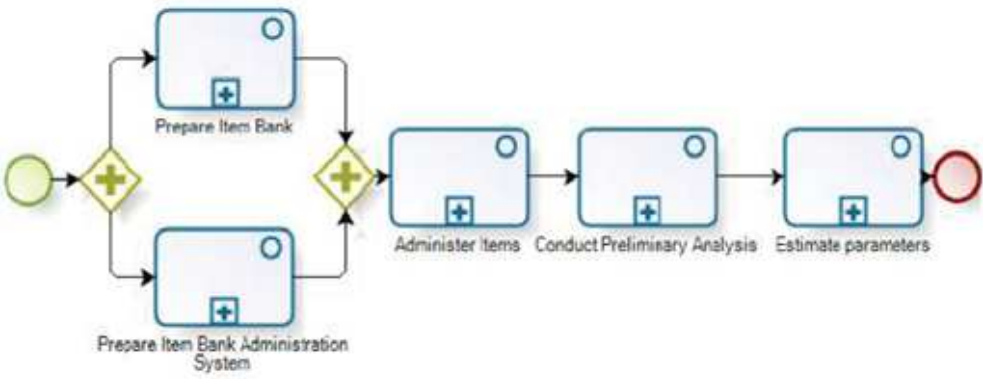

Fig. 3. The item bank calibration process.

The first step is identical for both expert-based and statistical calibrations. It consists in preparing the item bank, for instance, ensuring that all items are correct. At this point features like spelling, grammatical errors, item readability and consistency will be taken into account. The preparation of the bank becomes necessary when the item sources are diverse, that is, when different authors, with different criteria, have written the items. All the items should be analyzed and if any problem is detected, then the responsible for calibration should decide what to do (for example, edit or delete the item). This task, which can be trivial when the user is the author of the item bank, is done without any software support. For this reason, even though the process includes this task, CALLIE will not perform it, the system will simply remind the user to prepare the item bank instead.

The rest of the general process is fairly similar for both types of calibration, but there are some differences when comparing them task-by-task.

In the case of an expert-based calibration, the responsible must be extremely careful while preparing the administrations. It is advisable that a minimum of 7 experts assess each item (Dalkey, Brown et al., 1970), being crucial to provide very precise instructions about what they are supposed to do. In addition, the user will be able to describe which information wants to obtain from the set of experts: the difficulty of the items should always be asked, but it is also recommended to require their correct response, for instance, if the responsible for calibration is interested in identifying problematic items (i.e. those that are not answered correctly by all the experts).

The item administration can be p&p-based or computerized. To conduct the former, CALLIE is prepared to print the forms to be administered later. On the other hand, the latter

not only automatically gathers the opinions of the experts, but also controls the performance of the procedure, for instance, not allowing an expert to classify an item into more than one difficulty category, or warning them if some required information is missing.

Once all data have been recorded, it is usual to examine the opinions so that erroneous or problematic items would be edited or even removed from the bank. Some of the clearest examples are those cases in which some experts agree that there is a problem with certain item, or an expert answers an item incorrectly. It is also advisable to study the difficulty levels proposed by the experts, just in case there is much disagreement among them for some items. In these cases, CALLIE will recommend the removal from the bank of those items not concentrating certain amount of opinions (by default, 85%) in a continuous range of 35% of the numeric difficulty scale. So, for instance, if the used scale manages 12 levels of difficulty, then the system will recommend the elimination of those items not having at least 85% of the opinions given by the experts assembled in 4 consecutive difficulty levels. In no case will CALLIE remove an item without explicit agreement of the user, who can enable or disable this filter whenever they want.

Finally, the system will proceed to estimate the difficulty of the items that have passed the previous filtering. To do it, CALLIE computes, for each item, an arithmetic mean using the maximum number of judgments that are concentrated in a third (35%) of a continuous range of the difficulty scale. The calculated means are the difficulty levels to which each item belongs. Note that this procedure does not consider the outliers, i.e. the most extreme judgments, in favor of determining means with maximum consensus. The estimation process ends with a conversion to a (-3, 3) scale, made by means of a lineal equation that performs both a change of the centre of the original difficulty scale and a compression of its range of values to the new one. This conversion will allow the item bank to be homogenized and ready to be used in the generation of CATs.

For a statistical calibration the user has to prepare the logistics to administer the items of the bank, since a large sample of responses is required to estimate the parameters of the IRT model. For instance, in the case of wanting to calibrate an item bank using the 3PL model, the user would need at least 500 individuals to give an answer to each of the items (Bunderson, Inouye et al., 1989). The necessary data can be collected either by means of a traditional p&p administration of the items or by an electronic system. Although it is advisable to use a computerized administration, because employing p&p forms requires a posterior transcription of the responses into, for instance, a database, experience shows that there are some situations (such as those in which the subjects are computer-illiterate) when there is no choice.

The following step is to administer the items to a sample of examinees representing the population that will later use the final item bank. Depending on the number of both individuals and items available, it is recommended to distribute the items into several test forms (called subtests) and apply them separately. The problem in partitioning the sets of items and individuals 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 (Kolen & Brennan, 1995). 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. Consequently, it will be possible to get a common scale for the parameter estimates of the whole item bank, which will be the same that states the ability estimates given by any CAT created from it.

Once the answers to the items have been gathered, and before obtaining statistical estimates for the item parameters, it is strongly recommended to carry out some preliminary analyses, which are intended to detect and rectify existing anomalies. Actually, this step is probably the most delicate because, after revising and debugging the set of responses, the user has to determine not only which subtest administrations should be invalidated, but also which items should be removed from the bank for not satisfying a particular IRT constraint or having poor psychometrical properties. The first prior study concerns the analysis of the proportion and distribution of omitted responses. If there is a small number of omissions, they can be treated as incorrect responses, but otherwise the user should decide whether those subtests having more than a certain percentage of omissions should be invalidated or not. Moreover, items that have been generally omitted can be detected, so they should probably be discarded. Note that reducing either the sample of subtest administrations or the number of items in the bank affects seriously the calibration process; for instance, if the user invalidates too many subtest applications, it could be necessary to go back to the previous step of the process in order to restore the sample size.

The next analysis covers the identification of anomalous response protocols, such as those involving individuals that have selected the same multiple-choice option continuosly. Then, a conventional reliability analysis is recommended, that is, a study based on Classical Test Theory indicators like the Spearman-Brown coefficient, Cronbach's alpha, and item-subtest correlations. 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. There are many ways to perform the study of unidimensionality, which is essential for IRT one-dimensional models, but the most widely used technique is the exploratory factor analysis of tetrachoric correlations.

Finally, the last analysis considered among the preliminary ones is the study of differential item functioning, which occurs when examinees from different groups (commonly gender or ethnicity) with the same ability have different probability of giving a certain response to an item.

Since the average user that could want to calibrate an item bank will probably not have the background on statistics and psychometrics needed to perform these studies, in most cases this stage is not only the most delicate but also the most complicated to execute. And it is at this point where CALLIE can do appropriate calculi in a transparent way to the user. Thus, the responsible for calibration will be able to make decisions based on the information provided by the system, but no technical terms will be used during the interaction.

After having completed the preliminary studies and, maybe, having removed some items from the bank (for example, because they do not satisfy the one-dimensionality constraint), the user has to obtain item parameter estimates using the responses given to all the previously administered and non-invalidated subtests as input. It is important to measure the model-to-data fit, since it is the way of confirming 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 some IRT properties are lost:

item information will be unreliable, and, consequently, the ability estimates provided by some CATs that are generated from the item bank will be untrustworthy. As a result of the model-fit assessment, it is very common to remove some items from the bank because their characteristics (i.e. parameter estimates) do not match the IRT model.
The last step of the calibration process is the equating of the scales that measure the item parameters for each subtest, which will surely be different. However, since an anchor design has been used, 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. At this point, CALLIE will execute the equating stage if it is needed, so users will have the calibrated item bank at their disposal.

5. Decision making with the help of CALLIE

CALLIE needs the user to take a number of decisions that might be difficult to understand, especially for those people who have never conducted a calibration before. This is the reason why the system has a simple user-interface, which allows not only the skilled user to enter data in an easy way but also the inexperienced one to properly make decisions by explaining anything needed to them whenever it is required.
No matter which type of calibration is being performed, the first piece of information that the user must supply is the set of items to be calibrated. The items must be following the IMS Question & Test Interoperability standard (QTI) (IMS, 2002), which is not trivial for most people, so the system allows the user to enter usual items in an easy way to automatically and transparently convert them. This way, if the responsible for calibration already has the items in IMS QTI format, they only will need to import them or copy their code in the corresponding text box, whereas, otherwise, they will have access to a screen in which they will be asked for the statement and the response choices, as well as for the correct option, as shown in Figure 4.
The first decision to be made is to establish the kind of calibration process to conduct. As said before, CALLIE will recommend a statistical calibration if the followed IRT schema is the 3PL model, whereas both expert-based and statistical procedures will be suitable for the 1PL model.
When choosing an expert-based calibration, the user must provide some information regarding (1) the preparation of questionnaires, (2) the set of experts, and (3) the decisions about the filters to be applied.
To carry out a proper preparation of questionnaires it is essential to previously establish the number of difficulty levels that are being managed to classify the items. Then, the system will give some information about the number of items that every expert should assess, the advantages of forcing the experts not only to evaluate the items but also to give their correct answers, and some decisions like allowing response omissions (blank answers) and letting the expert write comments and suggestions.
Figure 5 presents an example of screen display in which CALLIE helps the responsible for calibration during the decision-making. The number of items to be calibrated (110 in the example) is shown as a reminder, while the value for the amount of required assessments per item (7, as previously said) is proposed by the system, and the total of experts in the set (14 in the example), and the number of items that each questionnaire will include (55 in the

example) will be asked to the responsible for calibration. Whenever one of these latter three values is entered or modified, the other two quantities will be automatically recalculated; so, for instance, if the user considers that 55 items are too many to be included in a questionnaire and therefore decides to reduce that quantity to 30, then the system will automatically change the number of needed experts from 14 to 21. Note that the number of required assessments for each item does not change, unless the user explicitly modifies its value.

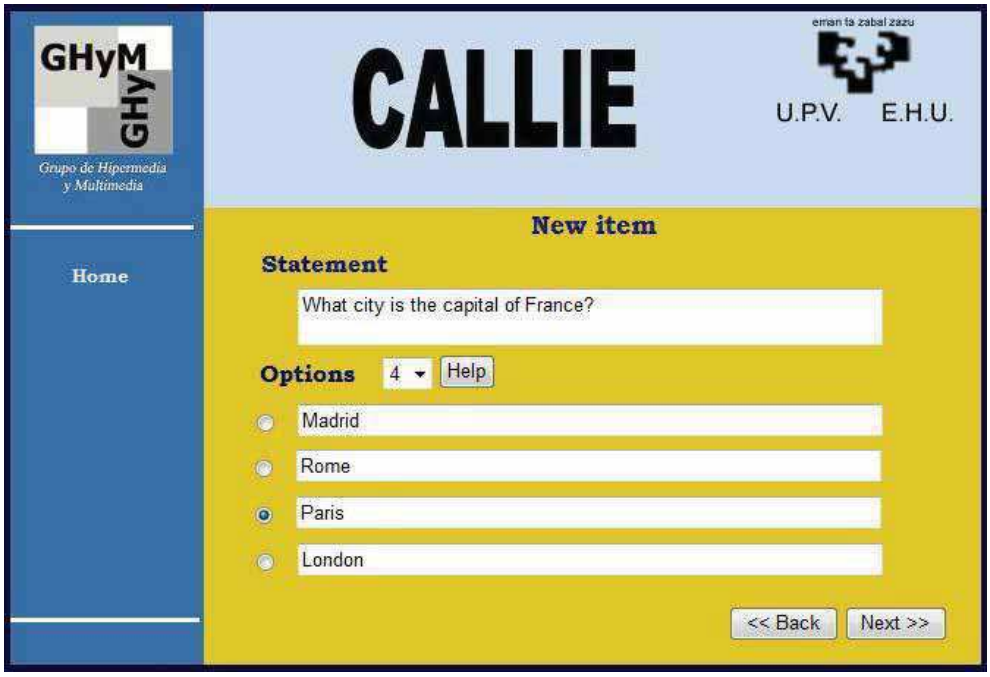

Fig. 4. Addition of a new item into CALLIE.

Additionally, CALLIE asks the user for the number of experts that will provide their assessments through the Web (12 in the example). Although all of them are expected, by default, to complete their questionnaires online, in some cases it is desirable to do it offline, i.e. by using p&p forms. Finally, next to each text box there is a help button that gives the user information about the meaning of each asked input value, as well as some recommendations in each case.

With regard to the information about the set of experts, the responsible for calibration must provide the name and electronic address of those experts who are going to give their judgements online. The email address will be used later to inform the corresponding set of experts about the web address to which they will have to connect to give their assessments.

Finally, the decisions about the filters that the responsible for calibration has to make are related to the deletion of administrations and the removal of items. The former involves applying more questionnaires to reach the number of assessments needed, whereas the latter implies a decrease in the size of the item bank. If the exclusion of questionnaire

administrations is allowed, then the responsible for calibration will have to specify the criteria to be used to filter them. For instance, the system can automatically discard any questionnaire that has not been answered at least at a certain percentage of items. On the other hand, if the deletion of items is allowed, then the user will be asked for the reasons under which an item should be removed from the bank, something that in most cases has to do with how well a particular expert guesses the correct choice of the items.

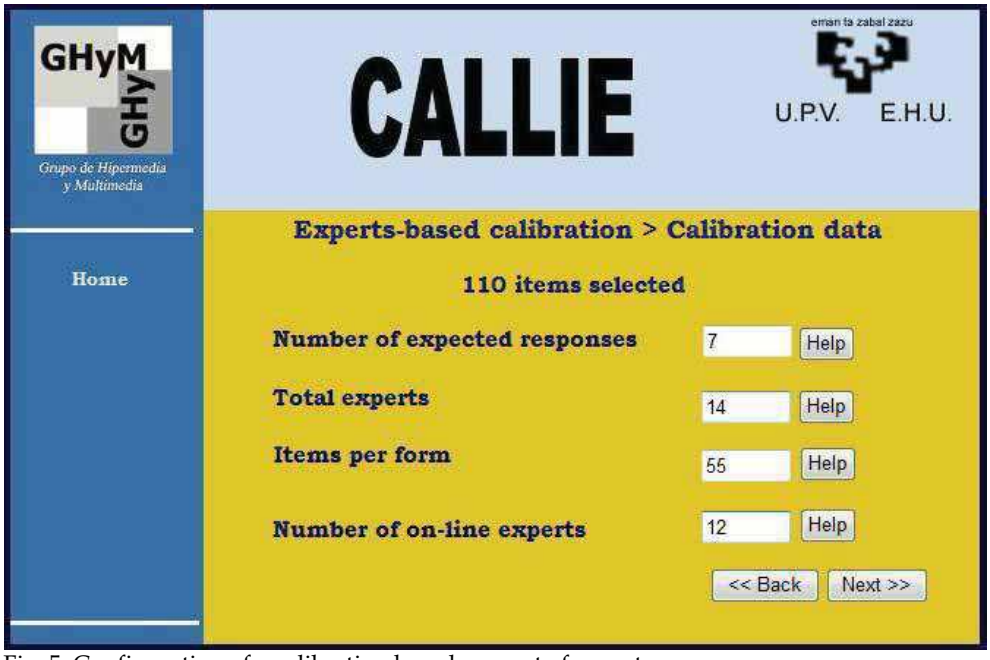

Fig. 5. Configuration of a calibration based on a set of experts.

The result of this decision making process is an XML document that gathers all the data given by the responsible for calibration. Figure 6 shows an example of an expert-based calibration (tag type, line 34) request in which only one parameter is being estimated (tag numparameters, line 35) for a set of 110 items (tag items, line 4). This sample-calibration is intended to classify the items into 12 levels of difficulty (tag levelnum, line 37), and to achieve this goal the system will require 7 judgements for each item (tag nresp, line 38). A set of 14 experts (tag nexperts, line 39) will be used, so each one will be asked to assess a 55 items-long questionnaire (tag nipq, line 40). They will be requested to give the correct answer (tag correctanswer, line 41) for each item, although they can omit them (tag blankresponses, line 43). They can also give a comment for each item (tag comments, line 42) if they want to, but in no case will they be allowed to omit their assessment of difficulty (tag blanklevel, line 44). With regard to the posterior automatic revision, it is possible to remove items from the bank (tag deleteitems, line 47). Any item being answered correctly by more than 50% of the experts (tag percentcorrect, line 50) and having at least 85% of its difficulty estimations (tag percentrequest, line 49) laying over a continuous range that contains no

more than 35% of the defined difficulty levels (tag percentlevel, line 48) will remain in the bank.

```
1<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
2<request>
3     < user> the.administrator@ehu.es </user >
4     <items n = “110”>
5          <item cod="1">
6               <questestinterop>
7                    <item ident = “56” title = “c1m3”>
8                         <qticomment></qticomment>
9                         <presentation>
10                              …
11              </questestinterop>
12         </item>
13         …
14         <item cod = “110”>
15              <questestinterop>
16                   <item ident = “603” title = “c5m4”>
17                        <qticomment></qticomment>
18                        <presentation>
19                              …
20              </questestinterop>
21         </item>
22    </items>
23    <stakeholders n= “12”>
24         <stakeholder cod = “1”>
25              <name> John Doe </name>
26              <mail> John.Doe@ehu.es </mail>
27         </stakeholder>
28         …
29         <stakeholder cod = “12”>
30              <name> Jane Doe </name>
31              <mail> jane.doe@ehu.es </mail>
32         </stakeholder>
33    </stakeholders>
34    <type> experts</type>
35    <numparameters> 1 </numparameters>
36    <detailsform>
37         <levelnum> 12 </levelnum>
38         <nresp> 7 </ nresp>
39         <nexperts> 14 </nexperts>
40         <nipq> 55 </nipq>
41         <correctanswer> yes </correctanswer>
42         <comments> yes </comments>
```

```
43          <blankresponses> yes </blankresponses>
44          <blanklevel> no <blanklevel>
45  </detailsform>
46  <filtering>
47          <deleteitems> yes </deleteitems>
48          <percentlevel> 35 </percentlevel>
49          <percentrequest> 85 </percentrequest>
50          <percentcorrect> 50 </percentcorrect>
51  </filtering>
52  </request>
```

Fig. 6. Example of an XML request for a calibration based on a set of experts.

In the same way as the previous case, if the user chooses a psychometrical calibration, then they will have to provide some information concerning (1) the preparation and administration of subtests, and (2) the decisions about the filters to be used.

The most important part of the preparation of questionnaires has to do with the definition of the anchor design. If this task is necessary, then the responsible for calibration can make some decisions, like selecting which items will form the anchor set, determining their position in each subtest, or deciding whether allowing the examinees to omit their responses. However, the responsible for calibration can let the system decide, especially if they are not experts in the field of calibration.

Figure 7 shows how CALLIE assists the user during the preparation of questionnaires. Concretely, it presents the number of items to be calibrated (482 in this example) as a reminder, as well as three more pieces of information, whose default values can be modified by the user: the number of parameters to be estimated (3 by default), whether an anchor design is needed (in the example it is, because of the quantity of items to be calibrated), and the minimum sample size required (500 subtest administrations, which is the default value for a calibration that follows the 3PL model). As happens during the whole interaction with the system, CALLIE shows, next to each text box, a help button that gives the user information about the meaning of each asked input value, as well as some recommendations in each case.

Finally, as happened with the calibration based on a set of experts, the decisions about the filters that the responsible for calibration has to make in this case are also related to the deletion of administrations and the removal of items. If the former are allowed, then the responsible for calibration will have to decide what to do with incomplete administrations and how to limit response times. In this context, CALLIE will recommend discarding any subtest administration that has not been completed in more than 5 and less than 50 minutes or that includes at least one item that has taken longer than 200 seconds to be answered. But, of course, these decisions also depend on the domain and the type of items that are going to be calibrated.

In the case of a statistical calibration too, the result of the decision making process is an XML document that gathers all the data given by the user. Figure 8 shows an example in which the responsible for calibration (tag user, line 3) wants to obtain three parameter estimates (tag numparameters, line 8) for a set of 482 items (tag items, line 4) by means of a statistical process (tag type, line 7). To achieve this goal, the system is ready to manage an anchor design that is formed by 11 subtests (tag nform, line 25), having 62 items each (tag nipq, line

26). The anchor set in this case is formed by 20 items (tag nanchoritems, line 11) whose identifiers (tag itemcod, lines 13, 18) and relative positions in the subtests (tag pos, lines 14, 19) have been chosen by the responsible for calibration. Each test form will be applied to a sample of 500 individuals (tag numadm, line 23), who will not be able to omit their responses (tag allowomissions, line 24). Both item removal and administration deletion are allowed (tags deleteitems and deleteadmin, lines 29 and 30 respectively) during the filtering stage. Finally, any data regarding a non-finished test form administration will be automatically deleted (tag incomplete, line 31), as well as the data related to subtests that, although they have been completed, have taken less than 5 minutes (tag minutesminadmin, line 32) or more than 50 minutes (tag minutesmaxadmin, line 33) to be finished, or include at least one item whose response has been given after more than 200 seconds (tag secmaxitem, line 34).

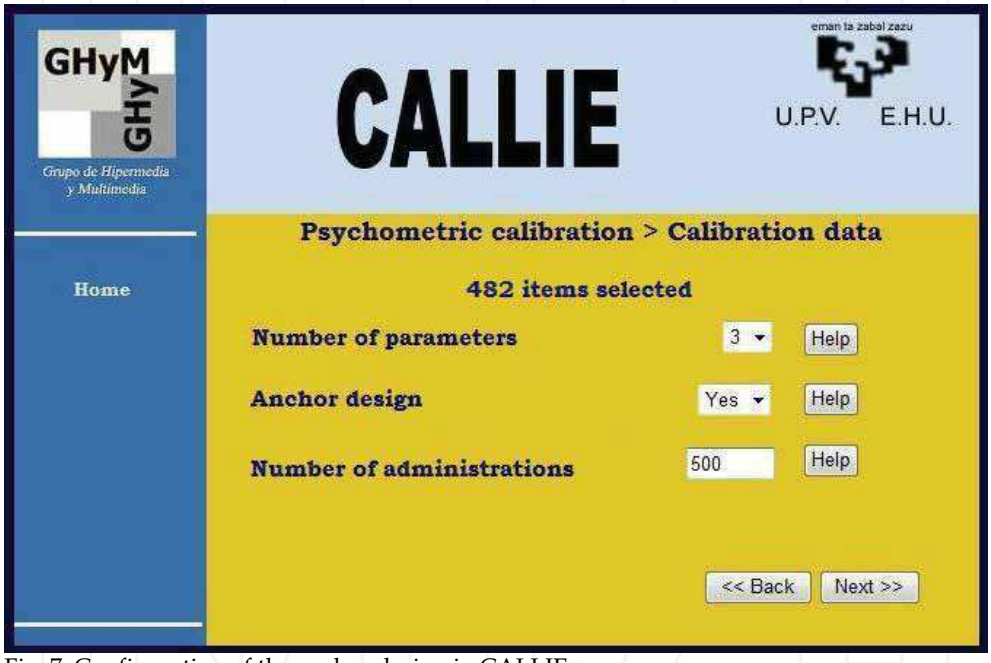

Fig. 7. Configuration of the anchor design in CALLIE.

Not only can users make their requests in such an easy way, but they will also be able to monitor and modify the progress of their calibration processes. Some of the elements that the responsible for calibration can control are related to the number of subtests that have been completed at the present time, the identity of the experts that have finished their questionnaire online, and the responses given to each administered item. In addition, CALLIE lets the responsible for calibration change their mind about any decision they took. Note that there are modifications that will not affect the defined calibration process, such as those concerning the data filtering, especially if they are made while the questionnaires are still being administered. In contrast, some other modifications could mean quitting the calibration process and setting up a new one, as happens, for instance, if the user decides to

change from a procedure based on a set of experts to a statistical calibration, in which case, any work done would be lost.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes"?>
<request>
        <user> the.administrator @ehu.es </user >
        <items n= “482”>
                …..
        </items>
        <type> psychometric</type>
        <numparameters>3</numparameters>
        <detailsform>
                <anchor r= “yes”>
                        <nanchoritems n = “20”>
                                <anchoritem>
                                        <itemcod> 3 </itemcod>
                                        <pos> 4 </pos>
                                </anchoritem>
                                …
                                <anchoritem>
                                        <itemcod> 407 </itemcod>
                                        <pos> 43 </pos>
                                </anchoritem>
                        </nanchoritems>
                </anchor>
                <numadm>500</numadm>
                <allowomissions>no</allowomissions>
                <nform> 11 </nform>
                <nipq> 62 </nipq>
        </detailsform>
        <filtering>
                <deleteitems>yes</deleteitems>
                <deleteadmin>yes</deleteadmin>
                <incomplete>reject</incomplete>
                < minutesminadmin> 5 </minutesminadmin>
                < minutesmaxadmin> 50 </minutesmaxadmin>
                <secmaxitem> 200 </secmaxitem>
        </filtering>
</request>
```

Fig. 8. Example of an XML request for a statistical calibration.

To make the supervision of an ongoing process easier, CALLIE lets the responsible check the state in which their calibrations are at the moment (Figure 9). The initial state for a process, which will change as the process goes through different stages, is allocated under the label “Received” when the user sends the corresponding request to the system. As soon as

CALLIE validates the request, i.e. verifies that it includes every needed piece of information, the state of the calibration will change to "Accepted"; otherwise, it will be set to the value "Pending" and the system will automatically send a warning to the responsible reporting any error detected. In that moment the responsible will be able either to fix the errors or cancel the calibration process, in which case its state will change to "Aborted". Once a request has been accepted, CALLIE will prepare the questionnaires or subtests to be administered and let the experts or individuals involved know that they are ready. When the first test form is filled in, the system will set the state of the calibration process to the value "In progress", which will remain until the last questionnaire or subtest administration is completed. In that moment the state will change to "Finished" and the responsible for calibration will be able to accept or reject the administration stage. As a result, CALLIE will carry out another change in the calibration state to the value "Calibrated" (if the process is accepted) or "In progress" again (if some administrations are discarded) or "Aborted" (if the calibration process is abandoned). Note that the user can "Abort" the process at any time.

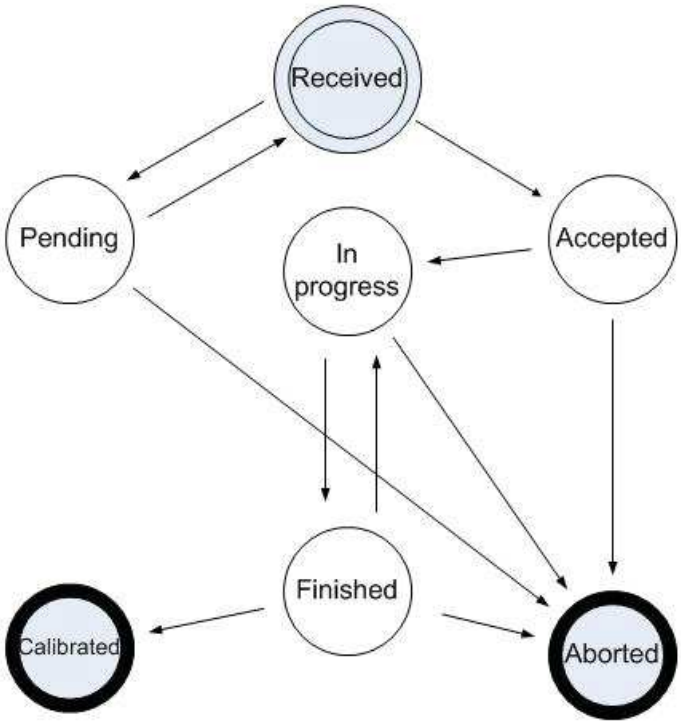

Fig. 9. State chart for a calibration process in CALLIE.

6. CALLIE's architecture

CALLIE's architecture is divided into three layers (Figure 10): a user interface (left on the image), which assists the responsible for calibration during the decision-making; a business logic module (right side module) that gauges the item bank as the decisions taken suggest; and a data storing module, which interacts with the business logic module. The communication between the different layers is carried out by means of XML strings and items are represented following the IMS QTI standard.

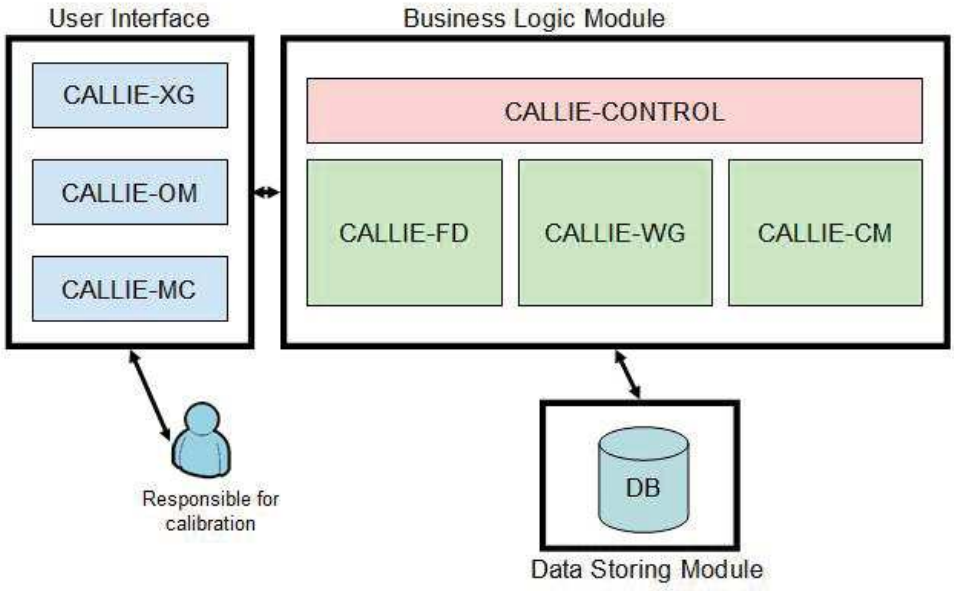

Fig. 10. CALLIE's architecture.

The data storing main module is composed of a database that stores, after having validated them, any request that is sent to the system. Whenever a new request is received, it is assigned an identification number and stored as an XML document. Figure 11 presents the schema of the database that gathers information about, among other topics, the sets of items to be calibrated, the questionnaires and subtests defined, the assessments provided by each expert, the responses given by the individuals, and the status of each calibration process.

The user interface main module is responsible for communication between the system and the user, who can perform any calibration request, manage the modification of options during the execution or supervision of an ongoing task. The interface is divided into three more specific modules, which are CALLIE–XML Generator, CALLIE–Option Manager and CALLIE–Monitor of Calibration.

CALLIE-XML Generator (CALLIE-XG) allows the user to make a request for a calibration. As previously said, the system will automatically generate an XML document, including the data concerning every decision taken by the responsible for calibration, which will be sent to the business logic module to formalize the calibration request.

The module CALLIE-Option Manager (CALLIE-OM) copes with option changes. For that purpose, the system evaluates the consequences of a modification in the calibration. For example, if the user wanted to change from a 3-parameter calibration to a 1-parameter one, the number of administrations required would decrease and it could possibly occur that the new sample size had been already reached. On the contrary, if the user wanted to change the anchor design, the administration stage would have to start from scratch.
The third module included in the interface is called CALLIE-Monitor of Calibration (CALLIE–MC) and it checks how the process is going. This is especially relevant when there are many administrations to be done. The user could check how it is going. This functionality is also included in an administration module created at the same time as the system automatically prepares the administration website for a calibration.

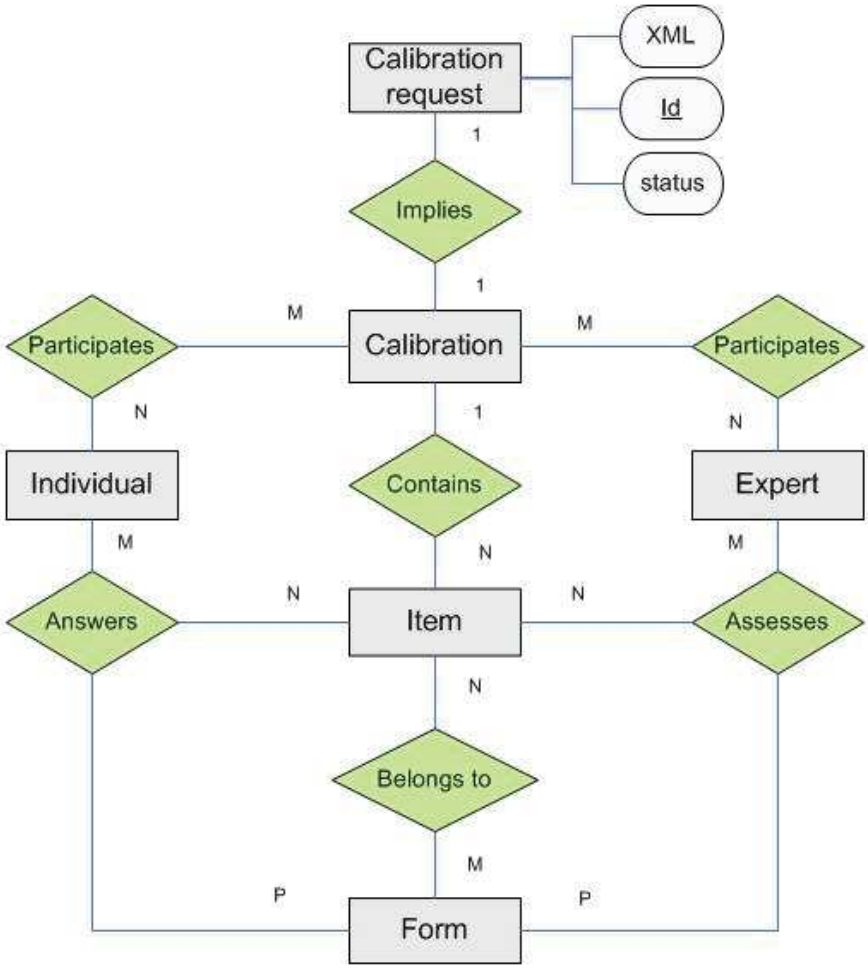

Fig. 11. Entity-Relationship diagram of CALLIE's database.

Although CALLIE has its own interface, it works as an independent module too, offering its services to multiple systems. To do it so, it offers an Application Programming Interface (API) to add these services to other learning management systems. The only condition required is that external systems have to send their requests following the previously discussed XML schema.

The third main module corresponds to the business logic layer, which interacts with both the interface and the data layer, and is responsible for conducting the calibration process. It includes four modules: CALLIE–Control, CALLIE–Form Designer, CALLIE–Web Generator and CALLIE–Calibration Module.

CALLIE–Control receives requests for calibration and processes them. In other words, it decides when a request for calibration is accepted or rejected, and, if it is accepted, it determines when its state should change. It is also responsible for conducting communication with the user interface, receiving requests for modification and supervision of the individuals responsible for calibration. CALLIE–Form Designer (CALLIE–FD) designs the subtests. It proposes an initial way of arranging the items in the questionnaires, implements any anchor design requested, and generates print-ready copies of the subtests for the case of p&p administrations. CALLIE–Web Generator (CALLIE–WG) administers the subtests designed by CALLIE –FD through a self-generated website that is prepared for that purpose. And CALLIE–Calibration Module (CALLIE–CM) takes the data gathered during the administrations of the items as input, and performs the calculations needed to conduct the preliminary analysis and estimate the parameters of the items, as previously discussed.

7. Conclusion

The need of a mechanism to assess the students for e-learning adaptive systems has been justified, particularly when placing every new student into their corresponding starting level. The use of a CAT generator in such situations is very appropriate, since this kind of tests provide accurate ability estimates by administering only about half of the items required by ordinary p&p assessments. The point is that, to work properly, the CAT algorithm needs the values of some psychometric characteristics (i.e. the parameters of the IRT model) to be estimated by means of a calibration process.

The work presented here focuses precisely on this issue; the calibration process has been studied, starting from the most used methods (expert-based and statistical) and reaching their unification into a general process that provides the steps to be followed when appropriately calibrating an item bank. The result of this work is CALLIE, a help tool that guides the responsible for calibration during the decisions they need to make in the course of their processes, whether they are expert-based or statistical calibrations.

CALLIE offers a number of improvements over traditional calibration processes. First, teachers, pedagogues and CAT-generating system developers will be able to calibrate their item banks, even if they lack of technical and psychometrical knowledge. Actually, the help tool is expected to make a step further to shorten the distance between IRT-based CATs and their application within e-learning systems. Besides, CALLIE will ease the preparation, administration and filtering of questionnaires and subtests, as well as the management of experts and individuals to be enquired. In addition, the help tool will allow the experienced responsible for calibration to take advantage from the resource management that it

provides. For instance, several concurrent calibration processes will be able to administer their subtests to a common set of individuals.

CALLIE's architecture has also been presented. Although the help tool has its own interface, it works as an independent module too, offering its services to multiple systems by a particular interface. The only condition required is that external systems have to send their requests following the XML schema defined for that purpose. Actually, it is intended to make it work together with ELSA, so that the e-learning architecture will provide not only adaptive assessments but also a framework useful to calibrate the item banks that will feed the CAT-generator.

Another future working line that is open is related to the addition of some other branches for the tasks of the calibration process that have not been taken into account yet. This will surely give more flexibility to the system. Besides, as CALLIE is an ongoing work, it is also intended to do some tests to assess its efficiency and effectiveness. First, it is projected to calibrate an item bank that measures musical dictation with 11-to-14-years-old students. The teachers that are building this bank at the present will be the testers of the final system. Besides, other educators that use e-learning systems have recently contacted the authors and are willing to be beta testers of CALLIE.

8. References

Arruabarrena, R. & T. A. Pérez (2005). Una experiencia arbitrando incidencias producidas en pruebas de campo, Proceedings of VI congreso nacional de Informática Educativa / I Simposio Nacional de Tecnologías de la Información y las Comunicaciones en la Educacion: SINTICE-CEDI'05), pp. 161-166, Granada (Spain), September 2005, Thomson Paraninfo, Granada (Spain).

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability, In: Statistical theories of mental test scores, F. M. Lord and M. R. Novick (Eds.), chapters 17-20, Addison-Wesley, ISBN 1593119348, Reading (USA).

Bunderson, C. V., D. K. Inouye & J. B. Olsen (1989). The four generations of computerized educational measurement, In: Educational Measurement (3rd Edition), R. L. Linn (Ed.), pp. 367-407, Collier Macmillan Publishers, London (UK).

Dalkey, N. C., B. Brown & S. Cochran (1970). The Delphi method, III. Use of self rating to improve group estimates. Technological Forecasting and Social Change, Vol. 1, No. 3, (March 1970) pp. 283-291.

IMS (2002). IMS question & test interoperability: an overview, version 1.2, IMS Global Learning Consortium, Inc. Available at http://www.imsproject.org/

Kolen, M. J. & R. L. Brennan (1995). Test equating: methods and practices, Springer-Verlag, ISBN 0-387-94486-9, New York (USA).

López-Cuadrado, J., A. J. Armendariz & T. A. Pérez (2005). A supporting tool for the adaptive assessment of an e-learning system, In: Recent research developments in learning technologies, A. Méndez Vilas, B. Gonzalez Pereira, J. Mesa González and J. A. Mesa González (Eds.) pp. 295-299, Formatex Research Center, ISBN: 609-5995-3, Cáceres (Spain).

Lord, F. M. (1952). A theory of test scores, Psychometric Monograph, Vol. 7 (September 1952) 357 pages.

Mills, C. N. & M. Steffen (2000). The GRE computer adaptive test: operational issues, In: Computerized adaptive testing: theory and practice, W. J. van der Linden and C. A. W. Glas (Eds.), pp. 75-100, Kluwer Academic Publishers ISBN 0-7923-6425-2, Dordrecht (The Netherlands).

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests, Danish Institute for Educational Research, ISBN 0-941938-05-0, Copenhagen, (Denmark).

Segall, D. O. & K. E. Moreno (1999). Development of the computerized adaptive testing version of the Armed Services Vocational Aptitude Battery, In: Innovations in computerized assessment, F. Drasgow and J. B. Olson-Buchanan (Eds.), pp. 35-65, Lawrence Erlbaum Associates, ISBN 0-8058-2876-1, Mahwah, New Jersey (USA).

Thissen, D. M. & R. J. Mislevy (2000). Testing algorithms, In: Computerized adaptive testing: a primer (second edition), H. Wainer (Ed.), pp: 101-132, Lawrence Erlbaum Associates, ISBN 0-8058-3511-3, Mahwah, New Jersey (USA).

Wainer, H. (2000). Computerized adaptive testing: a primer (second edition), Lawrence Erlbaum Associates, ISBN 0-8058-3511-3, Mahwah, New Jersey (USA).

Wainer, H. & X. Wang (2000). Using a new statistical model for testlets to score TOEFL. Journal of Educational Measurement, Vol. 37, No. 3, (September 2000) pp. 203-220.

Technology Education and Development

Edited by Aleksandar Lazinica and Carlos Calafate

ISBN 978-953-307-007-0
Hard cover, 528 pages
Publisher InTech
Published online 01, October, 2009
Published in print edition October, 2009

The widespread deployment and use of Information Technologies (IT) has paved the way for change in many fields of our societies. The Internet, mobile computing, social networks and many other advances in human communications have become essential to promote and boost education, technology and industry. On the education side, the new challenges related with the integration of IT technologies into all aspects of learning require revising the traditional educational paradigms that have prevailed for the last centuries. Additionally, the globalization of education and student mobility requirements are favoring a fluid interchange of tools, methodologies and evaluation strategies, which promote innovation at an accelerated pace. Curricular revisions are also taking place to achieved a more specialized education that is able to responds to the society's requirements in terms of professional training. In this process, guaranteeing quality has also become a critical issue. On the industrial and technological side, the focus on ecological developments is essential to achieve a sustainable degree of prosperity, and all efforts to promote greener societies are welcome. In this book we gather knowledge and experiences of different authors on all these topics, hoping to offer the reader a wider view of the revolution taking place within and without our educational centers. In summary, we believe that this book makes an important contribution to the fields of education and technology in these times of great change, offering a mean for experts in the different areas to share valuable experiences and points of view that we hope are enriching to the reader. Enjoy the book!

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Javier Lopez-Cuadrado, Anaje Armendariz, Tomas A. Perez, Rosa Arruabarrena and Jose A. Vadillo (2009). Computerized Adaptive Testing, the Item Bank Calibration and a Tool for Easing the Process, Technology Education and Development, Aleksandar Lazinica and Carlos Calafate (Ed.), ISBN: 978-953-307-007-0, InTech, Available from: http://www.intechopen.com/books/technology-education-and-development/computerized-adaptive-testing-the-item-bank-calibration-and-a-tool-for-easing-the-process

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820

Fax: +385 (51) 686 166
www.intechopen.com

Fax: +86-21-62489821

INTECH