Compiling a test: how to solve calibration issues

Anaje Armendariz*a,*, Tomás A. Pérezb, Javier López-Cuadradoa

aSchool of Industrial Technical Engineering (University of the Basque Country), Plaza de la Casilla, 3, Bilbao 48012, Spain
bComputer Science Faculty (University of the Basque Country), Paseo Manuel de Lardizábal, 1, San Sebastián 20018, Spain

Abstract

When compiling a test, there are several questions that have to be solved before proceeding to include the items. For instance: do I have to evaluate a minimum required knowledge (i.e. pass or fail) or do I just have to classify the examinees? How much accuracy shall I need? Which set of items fits best my needs? How difficult should these items be? How do I ensure that all those items are correct? Do they assess only the contents I want them to? How can I verify that? These issues are well documented in Psychometrics and the answers to the questions posed are usually based on statistic procedures that have to be followed during the so-called item calibration process. CALLIE is a revolutionary system in the field of item calibration. Its main objective is to support the calibration process and therefore let teachers, pedagogues and other content developers without specific background on psychometrics easily calibrate their own item banks. CALLIE includes a variety of workflow processes to solve some of the questions previously posed. These processes are integrated into a bigger workflow process that contains the milestones and decisions of a complete calibration. This paper describes the system and its features.

© 2012 Published by Elsevier Ltd. Selection and/or peer review under responsibility of Prof. Dr. Hüseyin Uzunboylu
Open access under CC BY-NC-ND license.

Keywords: calibration process workflow, computer adaptive testing, decision taking supporting tool, item response theory;

1. Introduction

Evaluation is a key issue in computerized-aided learning because it is the usual way to identify the success or failure in the learning process. The number of systems that integrate their own assessment modules, instead of delegating evaluation activities to some other external entity, is growing every day.

Most classifications in the literature show two kinds of computerized testing (e.g., Drasgow & Olson-Buchanan, 1999; van der Linden & Glas, 2000; Mills, Potenza, Fremer, & Ward, 2002): computer-based tests (CBTs) and computerized adaptive tests (CATs). The former refer to the use of computers in test administration and scoring, and they are often used interchangeably with computerized fixed tests (CFTs) (Parshall, Spray, Kalohn, & Davey, 2002). However, the acronym CBT covers any computerized assessment, including adaptive tests, whereas CFT refers only to non-adaptive tests, i.e. those that are most similar to paper-and-pencil (p&p) trials.

CATs (Wainer, 2000) 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 the fact of being computerized, CATs offer many advantages over the traditional p&p tests, mainly regarding to the conditions of application, the control, and the processing of the responses. In addition, when the compilation of the tests is carried out adaptively, the
evaluation is more secure, the administration takes less time, and the estimations are more precise. The most used framework for computerized adaptive testing is the item response theory (IRT), an item-oriented background that offers models that associate the ability of the examinee with the probability of a correct response (Lord, 1980).

The underlying idea of a CAT is to dynamically select the most appropriate item depending on the examinee’s performance during the test. This means that after a correct answer the next question will be slightly more difficult, whereas incorrect answers will be followed by easier items. To identify the most appropriate item during a CAT application, the implemented algorithm makes use of some psychometric characteristics of the items. So, from the developers’ point of view, one of the most important issues to be taken into account is that the item bank must be calibrated, which means that the item parameters required by the IRT model must be estimated. The most common models for that purpose are the 1-parameter logistic model (1PL), which characterizes the items by their difficulty (Rasch, 1960), and the 3-parameter logistic model (3PL), which also handles their discriminative power and guessing probability (Birnbaum, 1968).

The calibration of an item bank is not an extremely difficult process, but it requires time and effort since it involves administering the whole item bank to a large population. So, if the sample size is big enough, statistics will provide accurate item parameter estimates. This fact by itself could be enough to impede the development of some CATs (Craig & Stocking, 1995). However, once the item bank is calibrated and available, the implementation and use of this kind of tests becomes a fairly straightforward task. Moreover, not only is a calibrated item bank essential to build adaptive tests, but it also allows to perform studies about the behaviour of any test even before it is administered.

This paper is devoted to describe CALLIE, a tool intended to help the user in most of the calibration tasks. CALLIE focuses on the main decisions to be taken during the process and provides support to the tasks involved. Its implementation is based on the calibration process described by López-Cuadrado, Perez, Vadillo & Gutiérrez (2010), which is illustrated by figure 1 and fully described in the next section.

2. CALLIE and the calibration experience

Regardless of the particular calibration process followed, the objective will always be to obtain estimations for the parameters of the items, which in this context are identified by an IRT model. CALLIE’s wizard allows pedagogues, science educators and CAT-generating system developers to calibrate their item banks easily without requiring a specific background. The system shows the existing relations among the variables of the process and explains how changes in one of their values affect the rest of variables. Once the user has set up a calibration, CALLIE will automatically outline its configuration, warn of some potential risks of doing some modifications (especially when some of the previous work could be lost) and advise of any useful information, such as that related to the calibration of an item bank by two or more different processes.

CALLIE’s users need 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 thanks to many explanations about anything needed whenever it is required.

The IRT has a theoretical basis that is widely accepted, but it is well-known that the judgment of one or more experts is also a valid method to calibrate the items. Arruabarrena (2010) demonstrated that both statistical and
expert-based calibrations provide equivalent results under several conditions. CALLIE is aware of that circumstance and covers it too. 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. Although the internal representation of items in CALLIE is based on the IMS Question & Test Interoperability standard, which is not trivial for most people, the system allows entering usual items in an easy way to automatically and transparently convert them to that format.

The first task of the process described in figure 1 is the preparation of the item bank; at the same time it is necessary to prepare the system for the item administration, so that the administration of subtests to a big sample of individuals is available; once the responses are gathered, the next step consists in conducting some preliminary analyses in order to finally obtain the definitive parameter estimates. Next sections will focus on each activity.

2.1. Prepare the item bank

This activity consists in preparing the item bank by ensuring that all items are correct. At this point features like spelling, grammatical errors, item readability and consistency have to be taken into account. The preparation of the bank becomes necessary for instance when different authors, with different criteria, have written the items. At this step all the items should be analyzed and if problems were detected the responsible for calibration should decide what to do (for example, correct them or simply delete them). CALLIE does not perform any task during this high level activity. It will simply give the user information and advice.

2.2. Prepare the system for item administration

The data needed to compute item parameter estimations can be collected either by means of a traditional p&p administration or by an electronic system. A computerized administration can be more appropriate because p&p forms require a subsequent gathering of the responses into, for instance, a database. However, experience shows that there are some situations (such as those in which the subjects are computer-illiterate) when there is no choice.

CALLIE considers both item administering methods, as well as a combination of them. To do the former, one can find websites (such www.kwiksurveys.com) that are dedicated to the presentation of questionnaires that allow the person in charge inviting other people to fill them in. However, these systems are usually questionnaire-oriented, rather than item-bank-oriented. This means that if one wanted to include the same item in different questionnaires, it would have to retype it several times. On the same way, the user has to set the individuals that are being administered every questionnaire. CALLIE is item-oriented, so it helps organizing questionnaires. In addition, it can create a tiny website based on the one described by López-Cuadrado, Armendariz, & Pérez (2005) to perform and validate the computerized administration of each item. For the p&p item administration, the subtests can be printed and given to the users. Then, CALLIE would request to provide the results of the administration back to the system.

The logistics to administer the items have to be prepared in advance, since a large sample of responses is required to statistically estimate the parameters of the IRT model. For instance, if the 3PL model is used, the user would need 500 to 1000 individuals to give an answer to each of the items (Bunderson, Inouye & Olsen, 1989). To perform such a dense task (many items, many individuals), and also because of security matters, the evaluation items may be distributed into subtests and applied separately. To avoid the situation in which the parameter estimates do not share a common scale, a common-item non-equivalent-groups anchor design (Kolen & Brennan, 1995) can be used. The underlying idea is that the estimates for the common items, which form the anchor item set, can be compared, providing the key to equate the different subtest scales into a common one, which will be shared by the whole item bank.

This is the most critical activity in CALLIE, because the plan of the subsequent activities is based on it. A small change in this task would have a big impact in the following ones, so the system has not only to be clear, but also show which the consequences of any decision taken will be. The variables involved in the generation of the set of subtests are the type of calibration, the number of parameters to be estimated, the number of items to be calibrated, the number of individuals, and the number of administrations needed. These variables are mutually dependent. For example, the type of calibration depends on the number of parameters to be estimated. CALLIE knows that 1PL estimates can be done using either IRT or expert judgments, whereas 3PL estimates are based only
on IRT. This decision determines the number of responses that have to be gathered for each item. Besides, the number of individuals is also important, because it will define the amount of items that will have to be answered in each subtest.

2.3. Administer subtests

In this step the items are administered to a sample of examinees representing the population that will later use the final item bank. As stated before, depending on the number of both individuals and items available, it is recommended to distribute the items into several test forms or subtests and apply them separately. 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 (Kolen & Brennan, 1995). Then, the estimates for the common items, which form the anchor item set, will be equated. 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.

2.4. Preliminary analyses

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 analyses, which are intended to detect and rectify existing anomalies. Actually, this step is probably the most delicate one because, after having revised and debugged the set of responses, it has to be determined 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 analyses concern the proportion and distribution of omitted responses, the identification of anomalous response protocols, reliability analysis, several studies based on CTT indicators such as the Spearman-Brown coefficient, Cronbach’s alpha, and item-subtest correlations, the verification of the unidimensionality of the items, and the study of differential item functioning.

For each of the analyses, CALLIE allows the user determine the threshold values that will be used and discusses their effect on the resulting data. The user could learn, for instance, that requiring a 100% of responses would invalidate 80% of the administrations, and thus opt for a less demanding percentage. And it is at this point where CALLIE can do the appropriate calculi in a transparent way to the user, so the responsible for the calibration will be able to make decisions based on the information provided by the system, but avoiding technical terms during the interaction. For example, to test the unidimensionality, CALLIE uses the exploratory factor analysis of tetrachoric correlations. Some of these calculi can actually be made by specific statistical software, so CALLIE will interact with it in order to get the results needed, in a way that is absolutely transparent for the user.

2.5. Estimate parameters

The final stage of the calibration process is obtaining item parameter estimates using as input the responses given to all the previously administered non-invalidated subtests. As a result, scale-unified estimations for the item parameters of difficulty, discriminative power and guessing factor are obtained.

CALLIE estimates the parameters by using XCALIBRE, which also measures the model-to-data fit to verify that the selected IRT model and the parameter estimates empirically fit. Whenever it is needed a scale equation, CALLIE will carry out the mean-sigma procedure (Marco, 1977).

3. Conclusions

CALLIE will allow pedagogues, science educators and CAT-generating system developers to easily calibrate their item banks 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 when doing some modifications.

CALLIE uses workflows to implement the tasks. It provides automation since it automatically does system to system talk, saving time; continuity, since it allows instant notification of events within the workflow; centralization, because the workflow itself collaborates with as many applications as needed, simplifying the interlocution with the user; efficiency, as there is one way of showing how the process is; adaptability, since change is inevitable by modifying an existing activity to suit its evolution, or providing alternatives to it. Furthermore, the activities may be enabled or disabled at user's will; extensibility, because it is easy to include new activities, and standardization, which ensures that all users will follow the activities in the same way.

Besides, since every answer is stored in a database, it will be very easy to obtain automatically-generated reports that include up to date, ordered, precise, and homogeneous information about the subtest administrations. Moreover, it is possible to control the response times at item-level, something that becomes very useful to reject invalid answers and also to perform some psychometric studies. However, there are also some disadvantages in administering tests through the web. The most important one is that allowing any person to connect to the server could be an important source of adulterated tests, basically because theoretically it is not possible to distinguish which of the subtest administrations are serious and which of them had been completed just to try the web-application. Since this potential problem could mean an absolute lack of reliability in the results, it is crucial to implement some kind of validation.

References


