Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia - Social and Behavioral Sciences 29 (2011) 1189 – 1198

Procedia
Social and Behavioral Sciences

International Conference on Education & Educational Psychology (ICEEPSY 2011)

Prepare items for Large Scale Computer Based Assessment: Case study for Teachers' Certification on Basic Computer Skills

Christos Christakoudis[a*], George S. Androulakis[a,b], Charalambos Zagouras[a]

[a]*Research Academic Computer Technology Institute, Nikou Kazantzaki street , Univercity Campus of Patras-Rion, 26500, Greece*
[b]*Bussiness Administration Department, University of Patras - 26500,Greece*

Abstract

One of the main issues facing a Computer Based Assessment (CBA) system for large-scale examinations is the items' exposure due to repeated e-exams that take place in different times or places. Administrators could reduce the side effects of item's exposure *(possible leaks, copying in the examination process, etc)* by enriching the item bank with items that assess the same knowledge, skills and attitudes in a different way. This paper presents the life-cycle of items used for teachers' certification in basic computer skills in the frame of a national program that is carried out in Greece since 2003 (2003-2010). A methodology for preparing items for large scale assessment is introduced. The structure of items used and the roles of people involved are described. A process for preparing equivalent items (families) is introduced. The teachers' behavior concerning equivalent groups of items (families) was explored based on candidates' responses. Based on candidates' responses that have been recorded since 2003, we refined a set of rules that must be taken into account by items' authors or evaluators during the preparation of item bank for a large scale CBA system concerning basic IT skills.

Keywords: Teachers' certification ; Computer Based Assessment;Basic Computer Skills ;

1. Introduction

Many recent studies are focused on methods and tools for skills' assessment related to our modern society (Scheuermann & Björnsson, 2009; Scheuermann & Guimarães Pereira, 2008). Different groups of people may be interested in skills' assessment (policy-makers, companies, schools, universities etc).

* Christos Christakoudis. Tel.: +30-2610-960456; fax: +030-2610-960399.
E-mail address: christak@cti.gr.

doi:10.1016/j.sbspro.2011.11.353

Nowadays there is a shift on Computer Based Assessment (CBA) or Computer Adaptive Testing (CAT). These systems are actually software tools for administering e-tests in order to assess candidates through the responses that have been recorded electronically. Some of the most obvious advantages of a CBA could be the unbiased test administration and scoring, the ability to apply testing methodologies, the suitability for large-scale assessment etc (Asuni, 2008; Vrabel, 2004).

Since 2003 a national project has been carried out in Greece concerning the certification of teachers (elementary and secondary education) in basic computer skills. During this project (2003-2010) over 100.000 teachers participated in examinations that aimed to certify teachers' computer skills. These examinations are based on a Computer Based Assessment system (CBA) that has been developed by the Research Academic Computer Technology Institute (RACTI) (Androulakis, Zagouras, & Skiniotis, 2006). Most of the examinees followed a dedicated training program (48 hours) on basic computer skills (Theory-MS Windows, Word Processing, Spreadsheets, Internet & email, Presentations) (Papadakis & Chatziperis, 2005)

One of the main issues facing a CBA system for large-scale examinations is the items' exposure due to repeated e-exams that take place in different time or place. Administrators could reduce the side effects of item's exposure (possible leaks, copying in the examination process, etc) by enriching the item bank with items that assess the same knowledge, skills and attitudes in a different way. Moreover, a number of factors are involved and influence the whole process (eg. roles of people, infrastructure, geographic dispersion, security issues etc.).

In this paper we describe the model adopted for authoring and reviewing the items used for teachers' certification in basic computer skills by RACTI. The process of creating equivalent items (families) is presented. The results of items' analysis are given and a set of rules that refine the whole process is mentioned.

2. Theoretical frame

In psychometrics, Item Response Theory (IRT) is used for the design, analysis, and scoring of tests that measure abilities, attitudes or other variables (Lord, 1980; Linden & Hambleton, 1997; Baker & Kim, 2004). The IRT, also known as latent trait theory, is based on mathematical models that aims to describe the relationship of examinees' responses to a set of latent variables based on the assumption that an examinees' performance on a test can be interpreted as the observation of one or more latent traits (eg. ability).

The theory focus on the item by modeling the response of an examinee of given ability through a monotonically increasing function that is called Item Characteristic Curve (ICC). The basic idea in IRT is that the probability of a correct response to an item is a mathematical function of parameters that address to candidates or to items. Usually, the person parameter is called latent trait or ability while the items may be characterized by difficulty, discrimination and guessing parameters. A general model that links the ability of an examinee to the probability of correct response in a dichotomous dataset is given in the equation 1.

$$P_i(\theta) = c_i + (1 - ci)g\{a_i(\theta - b_i)\} \quad (1)$$

where θ is the ability level of a candidate, $P_i(\theta)$ is the probability of correct response, a_i is the discrimination parameter, b_i is the difficulty parameter and c_i is the guessing parameter.

In order to apply IRT it is assumed that (a) the observed performance of an examinee can be explained by a single trait or ability (unidimensionality) (b) the examinees' responses to any pair of items are statistically independent when the abilities that influence examinees' performance are held constant (local independent) (c) the probability of a correct response Pi(θ) will not depend on the group of examinees

used to estimate the item parameters and (d) a random examinee fails to answer an item correctly because of his/her limited ability, not because he/she is not given enough time to answer this item (speededness) (Hambleton, 1991).

Several models and estimation procedures have been proposed but it is most common to deal with three basic types of models: (a) one parameter model where the items in a test can vary only in difficulty. This model is also known as Rasch model (Rasch, 1981) (b) two parameter model where each item can have a different difficulty level and a different discrimination parameter while the guessing parameter is set to ci=0 and (c) three parameter model where each item can be characterized by difficulty, discrimination and guessing parameters.

In practice, item's ICC that follows the three-parameter IRT model can be defined by equation 2.

$$Pi(\theta) = ci + (1 - ci)\frac{e^{Dai(\theta - bi)}}{1 + e^{Dai(\theta - bi)}} \quad \text{(i=1, 2, 3, ...n items)} \qquad (2)$$

where D is a constant scaling factor that normally is chosen to be D=1.7 (Lord & Novick, 1968).

3. Teachers' certification for basic computer skills

Teachers' certification on basic IT skills requires (a) a Syllabus *(a well defined structure that clearly identifies the knowledge, the skills and the attitudes of candidates)* (b) an Item Bank (*a set of questions/items that shares quantitative, categorical and logical attributes)* and (c) a Certification System *(a set of processes that aims on "proving" that a specific candidate masters the particular subject into a predefined degree).*

The data gathered *(2003-2009)* are considered valuable because (a) different type of participants are involved *(elementary teachers, mathematicians, philologists, literature teachers, physicists, etc.)* (b) certifications are carried out in a national level *(large scale)* (c) five cognitive objects are involved *(theory-windows, word processing, spreadsheets, internet-email, presentation)* (d) time *(date and time, items' response time, etc)* and spatial *(prefecture, city, etc)* data have been recorded.

During this work a snapshot of 67,983 certification attempts were analyzed. Each certification attempt lasts maximum 2.30 hours and involves 60 items. It's a pass or fail test where an examinee is considered successful if he/she answers correctly 36 out of 60 questions (~60%). In total, 1.086 items were used and spread over five subjects (MS Windows, MS Word, MS Excel etc). A test generator module may utilize these characteristics in order to produce the e-exams based on certain criteria (Androulakis κ.ά., 2006).

The set of items (item pool or item bank) used for assessment is not a simple set of questions but a well organized collection of entities where each element is characterized by a well-defined set of quantitative, categorical and logical attributes (Linden, 2005).

3.1. Item bank

Traditional closed items (eg. multiple choice, true/false, fill in the gap etc) are not suitable for computer's skills assessment. Items should facilitate the demonstration of skills using a computer application. In this project, the items used are divided into the following categories: (a) **closed items** where the examinee has to choose the correct answer through a set of possible answers (destructors) (b) **inApplication items** where the examinee has to perform a set of actions using a specific computer application and one or more working files. Candidates' responses are recorded and evaluated automatically by a grading script without human interaction. The automatic evaluation is based on the final outcome produced by the examinee (usually a computer file). Any application that provides an

Application Programming Interface (API) can be used. (c) **pseudo-inApplication items** where the examinee has to choose the correct answer after he/she has performed a set of actions based on the attached working files. Usually it is not clear what the answer is and the teacher has to look for the proper answer using the available computer application and the corresponding working file.

Items' structure for computer's skills assessment can be described as a set of different modules that defines an item both educationally and technically. All these metadata varies by item's type and is presented in Figure 1.

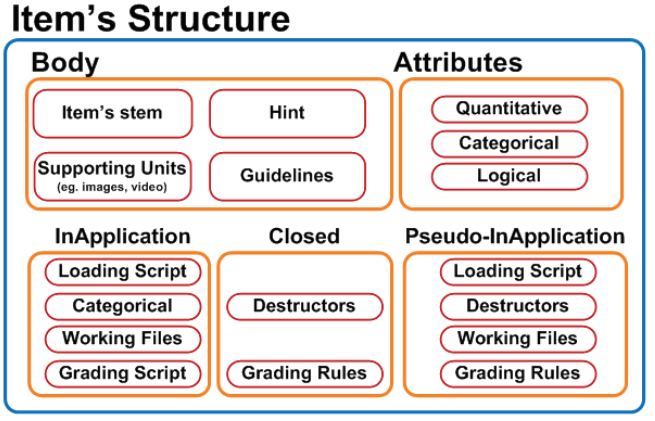

Fig. 1: Items' structure for basic IT skills

The item's body involves a short text (item's stem) that can be enhanced with multimedia elements such as text, video and images. It describes how the candidate will demonstrate that he/she possesses knowledge, a specific skill or an attitude. Since item's stem is stored in HTML format following appropriate cascading styles, the multimedia components cannot be integrated into the text and must be stored as attached files (supporting units). A short help that clarifies the item's stem (hint) or brief instructions in the form of predefined texts depending on the type of the item (guidelines) can be used. One or more files that can be edited by the application of interest and contain data relative to candidates' background (eg data coming from school life) were used (working files). The candidates were asked to perform actions (inApplication items) or to find the correct answer (pseudo-inApplication items) using these working files. A set of possible answers (destructors) in multimedia format (text, video and image) could be used on closed or pseudo-inApplication items. An item is coming with a set of rules about scoring (grading rules). Different grading policies could be implemented taking into account item's attributes (eg. type, degree of difficulty, expected response time). On teachers' certification project each item could be entirely right or entirely wrong (dichotomous data) and can be characterized by categorical, quantitative and logical parameters (attributes). These characteristics assigned by the item's author and shall be reviewed by others users involved in developing the item bank (quality control of two phases). Depending on the test generator, some other features could be utilized (eg. discrimination factor, exposing factor, etc.)

3.2. Roles of people involved in preparing the Item Bank

The life-cycle of an item in teachers' certification project includes: (a) authoring (b) programming a grading script (if it is necessary) and (c) review of item in an educational and a technical point of view. In order to manage the item's life-cycle a web based authoring tool was developed that enables the

communication and the collaboration among people with different roles. The roles of people involved are described:

- **iCon** (item constructor): An individual who has an extended knowledge of the related syllabus nodes (power users). Since the certification were targeting on teachers, the group iCon was based on active teachers from secondary education.
- **ScrCon** (script constructor): An experienced programmer. ScrCon must be able to control the relative application using a given API (eg OpenOffice, MS Word, MSExcel, etc.).
- **iEval** (item's evaluator): An experienced teacher not only having good knowledge of the relative application (power user) but also empirical knowledge on how a typical teacher acts using a computer. iEval plays a crucial role in this model having high pedagogical and technical skills, being able to communicate with other users involved (eg iTesters, ScrCons etc.) and being able to identify the problems in an item at all levels (educational and technical).
- **iTesters** (item's tester): iTester has to review an item from an educational point of view (reader's clarity, suitability of destructors, difficulty level, etc.) and check technical aspects as well (item's appearance, working files, responses to inappropriate actions etc.). During this project there was no need to be power users or active teachers. They act as a second and objective look of an item.

3.3. Life-cycle of items in large scale CBA

The communication among the groups of people involved in items' preparation (iCons, iTesters, ScrCons, iEvals etc) was supported by an authoring tool (Auto) that enables the management of item's life-cycle. During this life-cycle the users have different privileges on editing item's modules. For security reasons, a set of states defined for each item:

- **iPend** (item is not ready yet): the item has been uploaded by an iCon but it is not ready yet to be moved in the next stage.
- **Edupend** (pending for educational evaluation): the item has been completed by an iCon and is pending for educational review. The corresponding iEval must review this item (reading clarity, suitability of possible answers, difficulty level, etc.) by implementing the first phase of a quality control.
- **EduAcc** (educational accepted): the item has been tested by an iEval and can move on. If it's an InApplication or pseudo-InApplication item then it can be forwarded to the corresponding ScrCon in order the loading and the grading script to be prepared (TechPend). If it's a closed item then it can move directly to the second review level (TestPend)
- **EduDeny** (educational deny): item is deficient (eg. ambiguities in item's stem, wrong use of destructors, formatting problems, mismatch degree of difficulty etc.). The item must return to the relative iCon and corrections must be done based on iEval's comments. The iEval does not have the privileges to correct the item by himself and negotiate for this purpose with iCon using online tool.
- **TechPend** (pending for technical implementation): the item can be used by an educational point of view and the loading or scoring script should be developed. The ScrCon can read the debate's history between iCon and iEval and can develop the appropriate scripts.
- **TechAcc** (technical accepted): The loading and grading script have been developed. The iEval may continue the first evaluation phase by reviewing the item from a technical point of view (eg are there wrong actions that may lead to correct assessment or vice versa?).
- **TechDeny** (technical deny): technical failures were identified. The item must return to ScrCon for corrections. Comments that explain the problems found must be recorded in the Authoring Tool.
- **TestPend** (pending for testing): the first level of review has been done by iEval. The item can be moved to the second evaluation phase (pedagogical and technical) and can be examined by an iTester.

- **TestDeny** (test deny): problems or technical errors have been identified by iTester. The item returns to iEval's queue having a detailed review form filled. If iEval agrees that there are educational points that need to be corrected change item's status to EduDeny. If it is considered that there are only technical problems then moves the item in TechPend. If iEval considers that the comments made by an iTester are due to some misunderstanding may mark item as completed (Cmpl stage).
- **Cmpl** (completed): the item has been completed after two reviewing phases and is ready to be used by the test generator. Nobody can now change anything in this item (item's downgrading to lower stages must be rare and can be done only by a superuser)

The above process is presented in Figure 2 and tends to produce items free of errors and omissions in educational and technical level by applying two review levels.

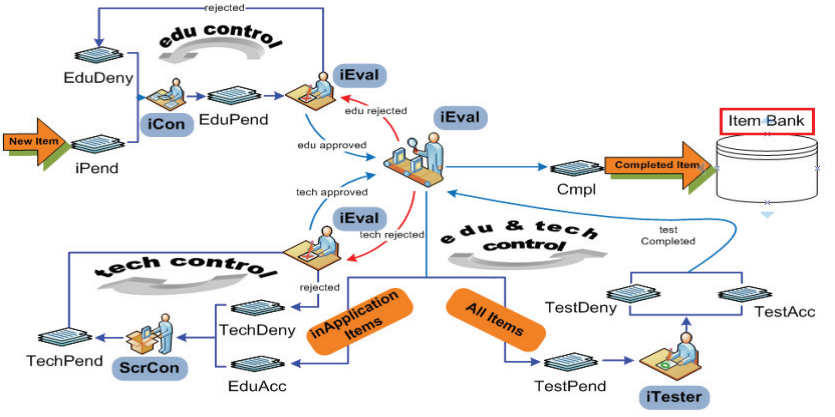

Fig. 2: Item's life-cycle

One of the main issues facing a CBA system for large-scale is the exposition of items due to repeated surveys that take place in different times or places. To reduce the side effects of item's exposure (possible leaks, copying in the examination process, etc) it is necessary to enrich the item bank with items that are equivalent (in terms of discrimination power, difficulty etc.) but also differ from each other (item's stem, possible answers, working files, etc.).

These items can be organized into groups (families). Items in a family must be exclusive (when an item is selected for an e-exam, no other item can be selected from the same family). A CBA system for large scale can be benefited from features that assist the management of such group of items.

A new process is/was added in item's life-cycle that is called *cloning*. The process of cloning aimed at creating a set of items (family) that share the characteristic of exclusive choice and are considered equivalent. The item bank was prepared by a team of experienced teachers (iCons) that designed the prototypes (fathers) for each family. Each father was exhaustively reviewed by an educational and a technical point of view and after that could be cloned. During the cloning process some rules had to be followed:

- an item must have the Cmpl status in order to be cloned
- the iCon must be able to copy and paste a completed item several times (cloning)
- the new items produced by cloning the family's father are assigned to the same syllabus's node and are moved to iPend queue.
- the iCon should modify the copies (clones) so that items in the family differ at least in item's stem, destructors or fetch files

- the items in the same family must have the same difficulty level

The iCons followed the above guidelines and designed the families. The basic assumption we made is that the families produced by this process would be treated the same way by the examinees (respond time, success rate, difficulty, discrimination power etc).

This research explores the following inquiry questions:

- how did teachers react on families' items?
- how the item's module influence the families' deviation?
- what rules must be followed by items' authors in order to improve the equivalence of family members?

4. Methodology

All families in the item bank were studied by following a process of three stages: (a) **Stage A**: prepare the statistical profile of each family based on IRT (b) **Stage B:** look for statistical differences in IRT parameters for each family's member (c) **Stage C:** review family items side by side in pairs trying to relate statistical deviances with item's modules (item's body, attributes, working files, grading script etc). In our analysis we used R (R Development Core Team, 2009), a powerful statistical application which is available for free to the scientific community.

4.1. Stage A

The IRT profile of each family was prepared based on three steps: (a) define a dataset of interest (b) calculate IRT parameters based on the current dataset (c) estimate item's IRT coefficients based on all dataset used.

4.1.1. Step 1: Define datasets of interest

For each e-test used in the examination process, the responses given by teachers to the items of the same subject were isolated and grouped according to the type of items into two categories: (a) closed and pseudo-inApplication items and (b) inApplication items. So in each data set dichotomous items of the same type were involved that assess the same ability concerning the basic computer skills (unidimensionality). We know from previous research that examinees complete the test in 2 hours approximately (speedeness) having enough time to answer to items (Christos Christakoudis, S. George, & Zagouras, 2010).

Step 2: Calculate IRT coefficients for a given dataset

An unconstraint 2PL *(each item in the dataset can have different $a \neq 1$ and $c = 0$)* and a 3PL *(each item in the dataset can have different $a \neq 1$ and $c \neq 0$)* IRT models were used in order to fit inApplication and closed/pseudo datasets respectively. Pearson's chi square (Drasgow, Levine, & Williams, 1985; Meijer & Sijtsma, 2001) used in order to test if the observed frequency distribution differs from the theoretical and dataset was refined based on the results of goodness of fit. IRT coefficients for the items in the current dataset calculated using the ltm package on R (Rizopoulos, 2006).

4.1.2. Step 3: Estimate the "true" IRT parameters for each item

Each item participated in many different datasets. Based on step2, for each item a list of IRT parameters are derived from its participation in several e-tests. In order to estimate the a, b, c IRT coefficients, a 95% confidence interval for the median (Lehmann & D'Abrera, 1975) of each parameter was calculated.

4.2. Stage B

Each family was exposed to non-parametric statistical tests looking for differences in IRT parameters. The index $\mathbf{F}_{idx}$ calculated for each family as showed in equation 3.

$$Fidx = \begin{cases} pvalue_a + pvalue_b + pvalue_c, & 3PL\,model(closed\ and\ pseudo\ items) \\ pvalue_a + pvalue_b, & 2PL\,model(inApplicationitems) \end{cases} \quad (3)$$

where $pvalue_a$, $pvalue_b$ and $pvalue_c$ are the pvalues of Kruskal-Wallis rank sum test *(Hollander & Wolfe, 1973)* over values for **a** (discrimination), **b** (difficulty) and **c** (guessing) IRT parameters for all family members. Since $\mathbf{F}_{idx}$ is defined as the sum of pvalues, the families with high $\mathbf{F}_{idx}$ are more likely to behave better than families with lower values (stronger evidences for differences in IRT parameters). So $\mathbf{F}_{idx}$ could be used as an ordering factor for families that are going to be processed in stage C.

4.3. Stage C

All families were reviewed in ascending order based on F_{idx} by comparing side by side the father with the rest items in the family (children) trying to relate item's modules (body, destructors, working files, grading scripts) with IRT profile of each item in the family. To better understand family's behavior a plot with ICCs for each item in the family was created as it is presented in figure 3.

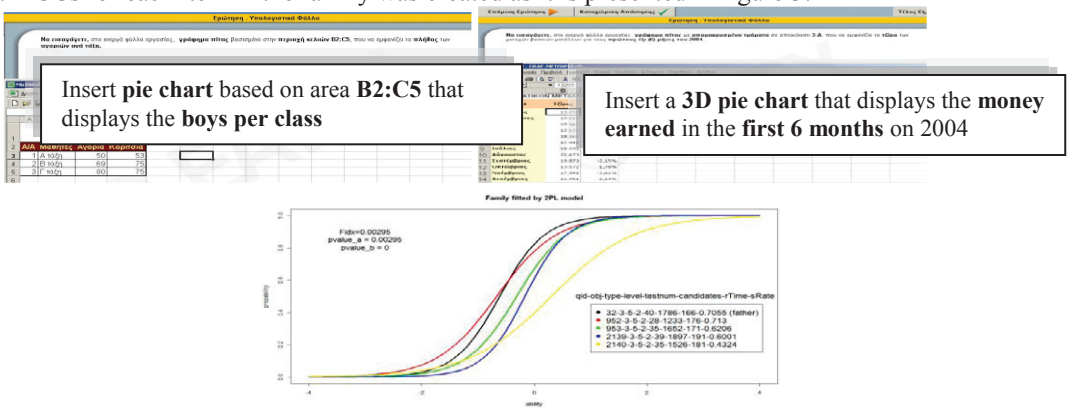

Fig. 3: Side by side review of family members based on family's IRT statistical profile

5. Findings

The differences found per IRT coefficient are presented in table 1.

Table 1. Results per IRT models explored

IRT model	total e-test	total responses	total items	total families	families that differ in IRT parameters					families with no differences
					a	b	c	a + b	a + b + c	
2PL	540	762.305	370	77	26 (34%)	43 (56%)	-	21 (27%)	-	29 (38%)
3PL		735.303	324	47	10 (21%)	25 (53%)	9 (19%)	9 (19%)	3 (6 %)	21 (45%)

For each family the $pvalue_a$, $pvalue_b$ and $pvalue_c$ of Kruskal-Wallis rank sum test was extracted and the $\mathbf{F}_{idx}$ was calculated. Families with small F_{idx} are considered more likely to include items that behave different than the rest of the family. The plot of $\mathbf{F}_{idx}$ for 2PL and 3PL IRT models is presented in figure 4.

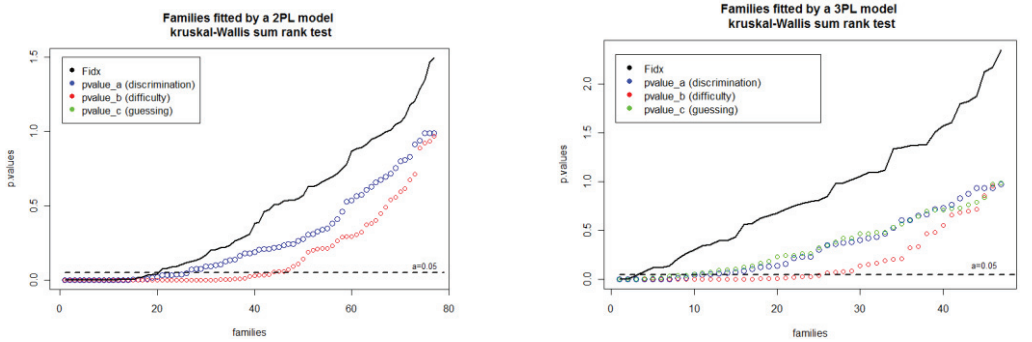

Fig. 4: Families' index (Fidx) for 2PL (inApplication items) and 3PL (closed and pseudo items) models

Enough statistical evidence found that almost 50% of families include items that differ in difficulty (56% for 2PL and 53% for 3PL) while differences in guessing or discrimination parameters are less. The inApplication families used in teachers' examination seems to behave more different than close or pseudo-InApplication families.

Families organized in two groups (2PL and 3PL) and families with small F_{idx} were checked first. In most cases the deviations recorded in a family that fitted by a 3PL model was due to changes in the item's type, the number and the quality of destructors. Usually, items' authors (iCon) during the cloning of a multiple choice item turned it into a True/False item with less destructors so the item appears to have higher success rate. However, in some cases although the destructors decreased the item's difficulty increased. After close consideration of these items (side by side check, ICC) we concluded that differences can be interpreted by the poor pronunciation in order to mislead the examinees.

In families that fitted by a 2PL model, the deviations come from differences in the wording of an item's stem. The authors attempt to differentiate the pronunciation of the item during cloning and often changed the meaning or the prerequisite knowledge needed to have the examinee to respond. In very few cases the variation observed in the family was attributed to the attached working file. It is worth to note that the working files used are very simple in general and coming from daily school life. Some observed differences in InApplication families were explained due to unexpected actions by the examinees (eg, using English versus Greek characters). So items that were constructed previously in time (fathers) behaved worse than items who prepared later in time having "better" grading script.

A comparative study of family items gave us examples of families with "good" or "bad" behavior. The main factors that affect family's behavior are (a) the item's type (true-false, multiple choice, fill gap, inApplication) (b) the number and the quality of destructors (c) the pronunciation used in item's stem (evaluate different knowledge / skill / attitude, wording) (d) the grading script (eg, using English versus Greek characters) (e) the working files (eg complexity of the data that the candidate had to work on).

6. Conclusions

In this paper, the model adopted for preparing item bank for a large scale assessment concerning teachers' certification on basic IT skills was described by exploring (a) the roles of people involved (b) the item's structure used (c) the item's life cycle (d) the process for educational and technical review of items (e) the process for creating families of items that behaves equally (cloning).

Based on IRT, a process for visualizing the statistical profile of each family was introduced (stage B). A process for ordering families based on the degree of deviance was introduced. Examples of families with "good" or "bad" behavior were isolated.

The findings enrich the set of rules followed during the preparation of item bank. The cloning process was refined by the following rules: (a) item's stem must assess the same knowledge, skill or attitude as father does. The rewording should be very careful in order not to alter the meaning of the cloning item (b) item's format adopted during father's preparation (guidelines, pronunciation, character formatting, etc.) should not be changed during cloning (c) during the cloning process the number of destructors must be kept the same (d) family's members must be of the same type (inApplication) having the same number of destructors (pseudo-inApplication or closed items) (e) the grading script must work in the same way among all family members. The use of the same script in order to grade all family members equivalent would be desirable (cloning of script, if it is possible) (f) cloning inApplication items cost more due to the grading script that have to be programmed (g) cloning multiple choice items (closed) by altering the destructors is not an easy job and is hard to create large families. On the other hand, cloning multiple choice items cost less than inApplication items.

Preparing an item bank for large scale assessment involves a lot of people that have to collaborate by following a well defined work flow. During this process knowledge and experience is producing from educational and technical point of view. This type of knowledge must be recorded by an authoring tool and can be used in order to prepare families.

Bibliography

Androulakis, G., Zagouras, C., & Skiniotis, T. (2006). TeCert: A management system for administering certification of knowledge, skills and attitudes concerning IT. Pedagogical Institute of Greece, 9, 3-27.

Asuni, N. (2008). Quality Features of TCExam,an Open-Source Computer-Based Assessment Software. Luxembourg: Office for Official Publications of the European Communities, Towards a Research Agend on CBA (Challenges and needs for European Educational Measurement), 58.

Baker, F. B., & Kim, S.-H. (2004). Item response theory: parameter estimation techniques. Marcel Dekker.

Christos Christakoudis, S. George, A., & Zagouras, C. (2010). Teachers' Certification on Basic Computer Skills: preliminary research. Proccedings of 5th Panellenic Conference on Teaching IT.

Drasgow, F., Levine, M., & Williams, E. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38, 67-86.

Hambleton, R. K. (Συγγρ.). (1991). Fundamentals of Item Response Theory (1st edition). Sage Publications, Inc.

Hollander, M., & Wolfe, D. A. (1973). Nonparametric statistical methods. New York: Wiley.

Lehmann, E. L., & D'Abrera, H. J. M. (1975). Nonparametrics: statistical methods based on ranks. Holden-Day.

Linden, W. J. van der. (2005). Linear Models for Optimal Test Design (1st edition). Springer.

Linden, W. J. van der, & Hambleton, R. K. (1997). Handbook of modern item response theory. Springer.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Addison-Wesley Publishing Company, Inc.

Meijer, R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. Applied Psychological Measurement, 25, 107-135.

Papadakis, S., & Chatziperis, N. (2005). Basic Skills on Information and Computer Technology. Ministry of Education, Lifelong Learning and Religion affairs- Pedagogical Institute of Greece.

R Development Core Team. (2009). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rasch, G. (1981). Probabilistic Models for Some Intelligence and Attainment Tests. Univ of Chicago Pr (Tx).

Rizopoulos, D. (2006). ltm: An R Package for Latent Variable Modeling and Item Response Analysis. Journal Of Statistical Software, 17(5), 1-25.

Scheuermann, F., & Björnsson, J. (2009). The Transition to Computer-Based Assessment. Citeseer.

Scheuermann, F., & Guimarães Pereira, A. (Editors). (2008). TOWARDS A RESEARCH AGENDA ONCOMPUTER-BASED ASSESSMENTChallenges and needs for European Educational Measurement. Institute for Prospective Technological Studies (IPTS), European Commission, Joint Research Centre. Technical Note: JRC, 48708.

Vrabel, M. (2004). Computerized versus paper-and-pencil testing methods for a nursing certification examination: a review of the literature. Computers, Informatics, Nursing: CIN, 22(2), 94-98; quiz 99-100.