Using Relation Tables to Improve Validity in External Evaluation

MARCOS PEÑATE CABRERA
Universidad de Las Palmas de Gran Canaria
PLÁCIDO BAZO MARTÍNEZ
Universidad de La Laguna

Received: 12 March 2013 / Accepted: 28 April 2013
ISSN: 1697-7467

ABSTRACT: Implementing external tests implies that test developers will need to address questions about validity and reliability. In this paper we are going to explain the process we have followed to design the English as a Foreign Language test intended for the fourth grade of primary school children in the Canary Islands, since we had been previously selected by the local authorities to coordinate this project. At the same time, we intend to provide a means of being accountable when evaluating. By accountability, we mean demonstrating to teachers and schoolchildren that all the items of our tests are justified and have a specific purpose.

Keywords: relation tables, validity, external evaluation, English as a foreign language, Primary education

Las tablas de relación como instrumento para la mejora de la validez en la evaluación externa

RESUMEN: A la hora de diseñar una evaluación externa necesariamente debemos prestar especial atención a los conceptos de validez y fiabilidad. En este artículo queremos presentar el proceso que hemos seguido a la hora de diseñar el formato que hemos utilizado en Canarias para realizar la prueba de diagnóstico destinada a la evaluación del inglés en cuarto de primaria. Además intentamos demostrar al alumnado y al profesorado que los diferentes ítems de las pruebas de evaluación están justificados y tienen una incidencia directa en su proceso de aprendizaje y enseñanza.

Palabras clave: tablas de relación evaluación externa, inglés como lengua extranjera, educación primaria

1. INTRODUCTION

External evaluation was introduced into Spain by law in 2006 (Ley Orgánica 2/2006 de Educación) and is now carried out in primary and secondary education two years before the end of each stage (fourth and second grade respectively), although the new government has decided to implement it right at the end of these stages in the future.

The main purpose of the above mentioned law is the implementation of diagnostic tests to assess the attainment level of the key competences. The different autonomous communities have subsequently taken the decisions they deemed necessary to adjust the way the assessment should be carried out, by selecting which key competences should be evaluated and how the tests would be designed.

Obviously one of the first key competences to be selected was linguistic competence which, in most cases, has been divided into communication in the mother tongue and communication in foreign languages, as was stated by the Recommendation of the European Parliament (Official Journal of the European Union, 30.12.2006). However, some autonomous communities have not evaluated the foreign language, and, in other cases, the assessment of the foreign language has not included the skills of speaking and spoken interaction. In our community (the Canary Islands), it was decided to test the foreign language by including the five skills mentioned in our curriculum: listening, speaking, spoken interaction, reading and writing.

Incorporating oral tests implies that test developers will need to address questions about the costs and the availability of resources for collecting the required information, that is, considerations related to the exam organisation (practicality). Besides the organisational problems, we had to face perhaps the most important challenge - establishing how we could guarantee the validity of the resulting tests. Before explaining how we have confronted this challenge, we will concentrate on some theoretical considerations related to two concepts: validity and external evaluation.

1.1. Validity

Reliability has always been considered the main factor when carrying out a test, but it does not imply validity, meaning that a reliable measure which is measuring something consistently, may not be measuring what you want it to measure. An example of what we are trying to say can be found in some of the items of an external evaluation carried out a few years ago in compulsory secondary education by the Ministerio de Educación (2004: 58-60). Though the main aim of these items was to assess the writing skill, we can see that they were in fact only asking students to fill in a gap with the right verb or noun, that is, they were only evaluating their linguistic accuracy.

In the same way, a test that has a high level of validity but is not reliable cannot be considered suitable. We therefore decided to concentrate on the validity of the test when planning and designing it and then check its reliability in the piloting and implementation process.

Validity guarantees that we are really measuring what we intend to measure (Camacho and Sánchez, 1997:28). But in order to find out exactly what this means and the way in which this concept has been understood and defined, we will address it from the perspective of our research field. To fully understand the concept of validity, we have to consider different approaches or ways to define it. Hughes (1989: 22-27) distinguishes the following types of validity:

- **Construct validity**. When it can be demonstrated that the test measures just the ability which it is supposed to measure.
- **Content validity**. A test is said to have it, if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned.
- **Face validity**. If a test looks as if it measures what it is supposed to measure, that is, when it is accepted by candidates, teachers, etc.

- **Criterion-related validity**. A test is said to have it, when the results of the test agree with those provided by some independent and highly dependable assessment of the candidate's ability. Both can be administered at the same time (concurrent validity) or the test can predict candidates' future performance (predictive validity).

One year later, Bachman (1990: 235-6) points out that although it has been traditional to classify validity into different types, it should be considered a unitary concept, and we should gather information about the different factors that are part of the validity concept. Some authors have organised validity from a slightly different perspective, as is the case of McNamara (1996: 16-22), who points out the following types: content, construct, predictive and consequential validity. Consequential validity is understood as the impact of the test on the educational system, a concept that deepens the idea behind face validity.

Different studies recently carried out have paid special attention to most of these features. For instance, Küçük and Walters' (2009) study, in which the authors analyse the predictive validity and face validity of a test implemented at a university in Turkey, as well as the reliability level of the test they used.

1.2. External evaluation

Schools today face a number of challenges as they strive to fulfill their important role in society. At the same time, the educational authorities have to assess whether students are obtaining the expected results by means of an external evaluation. This institutional evaluation may be defined as a continuous and concerted process of analysis and appreciation of the carrying out of the educational assignment of the schools (Beaumier et al., 2000: 3). However, this external evaluation has to be closely linked to the internal evaluation and some conditions have to be met to establish a constructive dialogue between both types of evaluation (Peñate and Bazo, 2007:311).

The continuous evaluation carried out by teachers in their classrooms provides quite significant information that allows teachers to take decisions to improve the teaching – learning process. However, an external evaluation is also felt to be needed to find out if schools are fulfilling their duties, and there has, moreover, been a hope that such external evaluations would motivate teachers and school principals to work harder to improve their schools (Gimeno Sacristán, 1989:375), although this has also been highly criticised by innovative educators who encouraged internal evaluation as an alternative (Nevo, 2001: 96). At the same time, we can see for ourselves that the internal evaluation is not always trusted. Our challenge, therefore, is to find ways by which we can improve both types of evaluation. Firstly, we should avoid mistaking evaluation for control (Mateo, 2000: 42). We believe that school teachers should participate in the designing of the external evaluation process, especially the assessment rubrics, but, at the same time, the internal evaluation should be in accordance with the agreed assessment indicators. That is to say, we should all be aware that assessments and the interpretation of assessment results have consequences for individual teaching and learning, and society at large (Stoynoff, 2012: 527).

It is of the upmost importance to establish the reasons for the introduction of a test, as these will help us to specify its educational, social and political context. McNamara (1996: 92-3) suggests answering the following four questions to define this context:

1. Who wants to know?
2. What do they want to know about?
3. About whom is information required?
4. For what purpose is information about the language abilities of test takers being sought?

In our case, questions 1 and 3 are made explicit by the type of assessment we are dealing with. Question number 2, however, is not so straightforward and simple and we intend to answer it by means of the relation tables we will explain in the next section. As for the purpose of the information obtained, this will be discussed in the conclusions of our paper in which we will insist on the importance of the washback effect (Prodomou, 1995).

2. OUR PROJECT

Taking into account all the opinions we have so far outlined, we decided to design the following four-year plan:

Phase 1: Incorporating assessment indicators in relation tables as a way to implement construct validity. This phase also included the negotiation of indicators with teachers responsible for the level to be assessed. In short, we wanted to identify and define what we wanted to measure, since when we define what that is, we are, in effect, defining a construct (Bachman, 1990: 255).

Phase 2: Coordinating the writing of two tests by a selected small group of teachers following the indicators of the relation tables. The main purpose was to control content validity, by linking the items to the indicators by means of specification files.

Phase 3: Piloting both tests to choose the most reliable items and writing the final test, making sure there was a balanced distribution of items per skill and that each one of them was linked to only one indicator.

Phase 4: Implementation of the test in each school by the English teacher. This phase also included a general question asking for their opinion about the test in order to check its face validity.

Phase 5: After four years we will compare the results obtained by a representative sample of pupils when they are evaluated again, at the end of their second year in secondary schools, as a way to check the predictive validity (criterion-related) of the primary school test.

So far we have completed the first four phases and in this article we will pay special attention to phases 1 and 2.

2.1. Phase 1: relation tables and construct validity

The local authorities in the Canary Islands decided to carry out this external evaluation using cognitive relation tables. The same tables were used for different subjects: Spanish, English, Mathematics, etc. The first thing we had to do was to study the proposed table carefully. As can be seen below, it is divided into three main increasingly difficult cognitive processes: Reproducing, Connecting and Reflecting. In addition, each main cognitive process is in turn divided into two different categories. The level of difficulty increases as we go from left to right i.e. "Connecting" is more challenging than "Reproducing" and in the same way "Analysing and evaluating" has a higher level of difficulty than "Applying".

Table 1: Type of relation table used to incorporate the assessment indicators

REPRODUCING		**CONNECTING**		**REFLECTING**	
Accessing and identifying	Understanding	Applying	Analysing and evaluating	Synthesising and creating	Judging and regulating

Before writing the assessment indicators for each of the linguistic skills to be evaluated, we had to bear in mind the exact meaning of each of the cognitive features used. A brief description of each one of them is presented in the following table.

Table 2: Definitions of the cognitive categories

REPRODUCING	
Accessing and identifying	**Understanding**
This represents the actions of remembering and recognising terms, facts, elementary concepts in a given field and reproducing previously learnt formulae.	This consists of understanding the meaning and purpose of texts, of specific language and related codes and interpreting them in order to solve problems.
CONNECTING	
Applying	**Analysing and evaluating**
This involves the skill to select, transfer and apply information in order to solve problems that have a certain degree of abstraction, as well as the ability to intervene accurately in new situations.	It means the possibility to examine and break up information into parts, find causes and motives, deduce, and find evidence to support generalisations. It involves a certain level of compromising.
REFLECTING	
Synthesising and creating	**Judging and regulating**
It corresponds to the actions of compiling information and relating it in a different way, establishing new patterns, discovering alternative solutions. It can be associated with conflict-solving abilities.	It represents the capacity to form judgements of one's own, question clichés, express and sustain well supported opinions. It is also associated with actions related to complex planning, establishing regulations and negotiation.

At the same time, it was decided that the evaluation was going to be carried out using the linguistic skills of the Common European Framework of Reference for Languages (CEFRL) of the Council of Europe (2001): listening, reading and audiovisual reception (receptive skills), speaking and writing (productive skills), and spoken and written interaction (interactive skills). Although in the Spanish Primary Curriculum there are only five skills (listening, speaking, spoken interaction, reading and writing), we can verify that audiovisual reception has been integrated into the listening skill, and written interaction into the writing skill.

We then proceeded to complete the five tables with the relevant indicators according to our curriculum and also to the CEFRL. This means that, if we want to establish the validity and the reliability of our test, we must begin with a set of definitions of the abilities we want to measure (Bachman, 1990: 163). This process was carried out together with a selected group of primary school teachers, as we wanted to avoid the gap sometimes found between external evaluation and school teachers. More than thirty years ago, Woodford (1980:97) already insisted on the idea that we should not "consider foreign language testing apart from the teaching-learning process." The final result was also negotiated with the educational authorities since the resulting document was to be published as an official assessment relation table.

We will now proceed to present the indicators incorporated into each relation table. As can be seen, these indicators are specific to each skill and are not intended to be used in the different skills in a cross-sectional way, as was proposed in this journal by Guillén Díaz *et al.* (2009:15).

Relation table 1: Listening

Accessing and identifying: 1. Listening carefully to the interventions of others. 2. Recognising phonic elements of rhythm, stress and intonation based on subjects familiar to the students. 3. Recognising lexis and linguistic actions by identifying words and essential sentences related to students' daily lives.

Understanding: 1. Grasping the general meaning of short oral texts in a foreign language by listening and understanding. 2. Listening to and understanding a story told by the teacher when given the necessary help. 3. Acquiring a general understanding of short, simple situations while watching an audiovisual document several times. 4. Identifying specific information in short oral texts in context on subjects that are familiar or of personal interest.

Applying: 1. Extracting specific information from a simple text (with two clues and no distractor). 2. Listening to and understanding sequences of instructions or simple directions provided by the teacher or by technical means (up to two actions).

Analysing and evaluating: 1. Listening and carrying out tasks of the following types: relating, sequencing, contrasting with visual information etc.

Relation table 2: Speaking

Accessing and identifying: 1. Reproducing the pronunciation and stress of the foreign language. 2. Reproducing the intonation characteristic of the foreign language. 3. Reading or reproducing simple texts aloud with adequate intonation and pronunciation. 4. Singing a song making adequate use of paralinguistic elements (gestures, mime, tone of voice..) as well as of linguistic elements. 5. Reciting a rhyme, a tongue-twister or short poems… with adequate intonation and pronunciation. 6. Expressing themselves in the activities carried out in class and in situations drawn up for such purposes (speaking with the help of notes, a written text, visual elements or spontaneously).

Understanding: 1. Acting out simple dialogues.

Applying: 1. Saying or reading aloud to the rest of the class short texts of the students' own production. 2. Gradually using characteristic linguistic structures in controlled texts. 3. Delivering statements (of three or more words) on subjects that are familiar or of personal interest to the student.

Relation table 3: Spoken interaction

Accessing and identifying: 1. Recognising phonic elements of rhythm, stress and intonation based on subjects which are familiar or of interest to the students.

Understanding: 1. Showing understanding and taking part in basic oral exchanges containing information about themselves.

Applying: 1. Using the foreign language as a vehicle for communication in the classroom, employing the basic norms for interaction. 2. Interacting in the foreign language with the teacher and peers in normal classroom activities and in communicative situations set up for such ends. 3. Reacting linguistically to spoken language related to subjects familiar to the students or of interest to them. 4. Playing different games using language previously acquired or learnt. 5. Taking part in guided oral exchanges (in pairs) on familiar subjects in real or simulated communicative situations.

Analysing and evaluating: 1. Valuing the foreign language as a tool for communication with others.

Synthesising and creating: 1. Acting out in class oral exchanges that have been previously prepared individually or in groups

Relation table 4: Reading

Accessing and identifying: 1. Reading and understanding sentences without visual support. 2. Reading and understanding sentences and short texts accompanied by visual information. 3. Starting to learn how to use a bilingual dictionary.

Understanding: 1. Grasping the gist of simple texts. 2. Grasping specific information in simple texts. 3. Obtaining relevant information about previously specified facts or phenomena from internet or other sources.

Applying: 1. Extracting the specific information required from a text. 2. Reading books adapted to a beginners' level and recognising aspects of everyday life, comparing them with their own experience. 3. Reading and understanding a comic.

Analysing and evaluating: 1. Reading and carrying out tasks such as: relating, sequencing, contrasting with visual information etc. 2. Showing initiative and interest in reading certain texts. 3. Showing a positive attitude towards reading as an activity typical of everyday life.

Relation table 5: Writing

Accessing and identifying: 1. Presenting texts in a clear, neat and orderly fashion.

Understanding: 1. Completing sentences with words heard on a CD or other technical means. 2. Writing words and sentences using visual support. 3. Filling in an incomplete text using visual support.

Applying: 1. Applying the following basic norms for spelling and punctuation: the use of capital letters both at the beginning of a sentence and with proper names as well as the use of the full stop. 2. Putting words in the right order so that they are grammatically correct. 3. Writing in the foreign language sentences and short texts of interest, in situations typical of everyday life, using previously studied models. 4. Making progressive use of linguistic structures characteristic of the foreign language.

Analysing and evaluating: 1. Writing short texts using the most frequent connectors at this level (‘and’ and ‘but’). 2. Valuing the written foreign language as a tool for communication with others.

Synthesising and creating: 1. Composing texts used socially in everyday life. (short messages: notes, letters, postcards, emails).

2.2. Phase 2: the writing of the test and content validity

Once we had identified and defined what we wanted to measure, we proceeded to design the test. The same group of primary school teachers was again asked to start writing two tests to assess the indicators established in the relation tables. First it was established that the final test should have a balanced number of items per skill and that the listening, reading and writing skills were going to be evaluated in the whole group by means of a pencil and paper test, whereas the speaking skill was to be assessed with the one-to-one format (teacher and pupil) and the spoken interaction using the paired format. It was also considered necessary to have few activities but to have items evaluating different features.

Once the two tests were completed, we piloted them in four primary schools to collect feedback which would help us modify and revise the different tests. The results were then analysed from two different perspectives: discrimination power and difficult level of each item. This study allowed us to design the final test.

Once we had finished writing the test, we had to draw up a file for each item (see examples below) defining what it was actually evaluating, so that we could explain the main objective of each item to the school teachers who were to assess their school children. This is extremely important, since assessment should also be used as a tool for clarifying teaching aims and for evaluating the relevance of these aims and the activities based on them (Bachman and Palmer, 2010:12). We expect examination practices to reflect good classroom practice, and to encourage it. Examinations and tests must connect with the classroom, because they exist as part of a longer and more important process: the process of learning and teaching (Együd and Glover, 2001:75).

This information was also important so that each teacher could decide whether each item was really evaluating what it was intended to. This feedback was taken by means of a list, in which teachers were asked to say if each item really assessed the defined indicator.

In the following example we can see the file designed for item 6, in which we intended to assess the ability to extract specific information from a reading text.

Table 3: File for item 6

Item 6: reading an e-mail
Instruction: Read the e-mail and then choose the right answer.
6.1 Hannah's hobbies are … A) watching TV and listening to music. B) playing with her pets and reading comics. C) collecting stamps and coins. 6.2 Hannah likes … A) Maths. B) Science. C) Art. 6.3 Hannah doesn't like … A) watching TV. B) playing with her pets. C) collecting stamps.
Answers: 6.1 B, 6.2 A, 6.3 A.
Correction criterion: Right: no mistakes. Partly right: 1 mistake. Wrong: 2 or 3 mistakes.
Evaluation criterion Skill: reading Cognitive process: Applying Indicator: 1. Extracting the specific information required from a text.

The next example belongs to item 8 and is devoted to the listening skill. Pupils had to listen to a simple recorded text about the daily routines of an English child.

Table 4: File for item 8

Item 8: listening to a CD
Instruction: Listen and put the pictures in order (1 to 3)
Three labelled (a-c) pictures depicting three daily routines of a boy but in the wrong order as pupils are expected to number them correctly.
Answers: A: 2, B: 3, C: 1
Correction criterion: Right: no mistakes. Wrong: 1 or more mistakes.
Evaluation criterion Skill: listening Cognitive process: Analysing and evaluating Indicator: 1. Listening and carrying out tasks of the following types: relating, sequencing, contrasting with visual information etc.

The feedback given by teachers about each item at the end provided us with the necessary information to confirm the validity of most of the proposed items.

3. Conclusions

In the first part of this article, we tried to answer the four questions addressed by McNamara (1996: 92-3), but we decided to answer the last one in this section. The question is: For what purpose is information about the language abilities of test takers being sought?

Our main purpose has been to set standards and benchmarks as a means for school improvement. To reach that aim we have to prove the validity of our assessment approach, and we have done so by paying special attention to the construct validity. However, what we consider to be the most relevant feature of our project is providing teachers, students and examiners with a tool that can easily demonstrate if a test in fact measures the ability it is supposed to. We are here referring to the relation tables designed to list all the aspects under evaluation. Without this in-between tool, it would probably be quite difficult to guarantee the most important aim: the washback effect. This would mean that we were not looking for the perfect test, but rather developing an assessment whose intended uses we can justify to stakeholders (Bachman and Palmer, 2010: 10). Without the relation tables, it would not be easy to find out what is being assessed by means of multiple choice, gap-filling item types, etc. We could have the feeling that we are testing accuracy rather than language development and form rather than content. This was even more obvious when we asked the school teachers to evaluate content validity by establishing whether each test item was really linked to the defined indicator. Teachers found it quite a straightforward task and the results obtained point out that some of the spotted weaknesses of the test were due to the test items and not to the indicators. This means that we now not only have a tool to improve our external evaluation but also to clarify the aims of the teaching and learning process. We are obviously referring to the relation tables.

To sum up, we think that the external evaluation should be closely linked to the teaching and learning process that takes place in every classroom. It should also become a real diagnosis and thus improve the way it is perceived by teachers and pupils. However, this will only be feasible if teachers are involved in the diagnostic assessment, and this involvement is not merely limited to the role of implementing the provided tests. That is, the test should be written by a chosen group of primary school teachers following the relation tables, and, at the same time, these relation tables should be improved and redefined according to the feedback provided by all the school teachers. In the same way, the continuous assessment carried out in each classroom should use the indicators included in the relation tables as a tool to shed light on this challenging process.

4. References

Bachman, L.F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. and Palmer, A. (2010). *Language Assessment in Practice*. Oxford: Oxford University Press.

Beaumier, J.P., Marchand, C., Simoneau, R. and Savard, D. (2000). *The Institutional Evaluation.* Québec: Government of Québec.

Camacho Martínez, C. and Sánchez García, E.F. (1997). *Psicométrica.* Sevilla: Editorial Kronos.

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Együd, G. and Glover, P. (2001). "Oral testing in pairs – a secondary school perspective", in *ELT Journal,* 55, 1: 70-76.

Gimeno Sacristán, J. (1989). *El curriculum: una reflexión sobre la práctica.* Madrid: Morata.

Guillen Díaz, C., Santos Maldonado, M.J., Ibañez Quintana, J. and Sanz de la Cal, E. (2009). "Los criterios de evaluación del ELE. Análisis de sus índices y propiedades en los referentes curriculares actuales", in *Porta Linguarum,* 11: 9-32.

Hughes, A. (1989). *Testing for Language Teachers.* Cambridge: Cambridge University Press.

Küçük, F. and Walters, J. (2009). "How good is your test?", in *ELT Journal,* 63, 4: 332-341.

Mateo, J. (2000). *La evaluación educativa, su práctica y otras metáforas.* Barcelona: ICE-Horsori.

McNamara, T. (1996). *Measuring Second Language Performance*. London: Longman.

Mendoza, A., de Amo, J., Ruiz, M. and Galera, F. (eds.) (2007). *Investigación en Didáctica de la Lengua y la Literatura*. Barcelona: Universidad de Barcelona.

Ministerio de Educación, Cultura y Deporte. (2004). *Evaluación de la enseñanza y el aprendizaje de la lengua inglesa. Educación Secundaria Obligatoria 2001*. Madrid: INECSE.

Nevo, D. (2001). "School evaluation: internal or external?", in *Studies in Educational Evaluation*, 27: 95-106.

Peñate, M. and Bazo, P. (2007). "El inglés en primaria: un caso de evaluación institucional", in Mendoza *et al.* (eds.), *Investigación en Didáctica de la Lengua y la Literatura*. Barcelona: Universidad de Barcelona, 311-334

Prodomou, L. (1995). "The backwash effect: from testing to teaching", in *ELT Journal,* 49, 1: 13–25.

Soynoff, S. (2012). "Looking backward and forward at classroom-based language assessment", in *ELT Journal,* 66, 4: 523-532.

Woodford, P.E. (1980). "Foreign Language Testing", in *The Modern Language Journal*, 64: 97-102.