Optimizing language assessment – focus on test specification and piloting

Yolanda-Mirela Catelly*

POLITEHNICA University Bucharest, Bucharest 61327, Romania

Abstract

The main focal point of the paper is represented by a proposal meant to enhance the quality of the language proficiency assessment process in the non-philological higher education, by emphasizing the role of the test specification and that of piloting in designing effective forms of evaluation by language teachers who embark upon test designing. Context-specific features of the concrete educational situation are discussed in their essential characteristics insofar as they can be influential upon the decisions conducive to the generation of the final format and content of a large-scale English proficiency test for the engineering undergraduates of the POLITEHNICA University of Bucharest. A framework of main principles to be taken into account is thus attempted at, as a response prompted by reasons of practicality to the range of contextual constraints to be faced, with the observance of the validity and reliability requirements in establishing test design priorities. The paper is also intended as an open expertise-sharing exercise of a reflective teacher/test designer addressed to fellow trainers, inviting debate and encouraging transferability of good practices in the foreign language teaching and evaluation.

© 2014 The Authors. Published by Elsevier Ltd. Open access under CC BY-NC-ND license.

Keywords: foreign language assessment; test specification; test piloting; language proficiency test; good test criteria; non-philological higher education

1. Paper focus and rationale

In a knowledge society foreign language university teachers who wish to develop both professionally and personally should be also competent in test design for their students, as assessment is an important component of the instructional process, by means of which they contribute to preparing the qualitative efficient graduates of tomorrow. Over the last decade, both the macro and micro levels of assessment, policies and practices have been changing considerably, and their information on activities and outcomes should become public (Brindley, 1997).

* Yolanda-Mirela Catelly. Tel.: 004-0723690587
E-mail address: yolandamirelacatelly@yahoo.com
Therefore, evaluation as a continuous process goes beyond the mere focus on objectives, aiming to examine the conditions underlying learning (Hutchings, 1990). Moreover, a thorough analysis is necessary of the ways teaching can affect the trainees’ cumulative levels of understanding, if by means of evaluation aspects such as the critical thinking level, cultural understanding or empathy are covered (Astin, 1996). Among the final results of the new teaching/learning/evaluation models, the development of a superior form of independent thinking becomes essential. Consequently, implementing change in higher education is a complex process, requiring good planning, flexibility and maintaining control over the components in a holistic approach – which is even more valid when a new type of assessment practice is initiated.

Hence, the array of aims of this study, which describes the test specification design and piloting stages in a relatively large scale language proficiency project, which took place in the POLITEHNICA University of Bucharest: to analyze the contextual factors influential upon the decisions to be made, to identify ways of coping with the constraints of all types involved in such an enterprise, while ensuring and observing the essential qualitative requirements of validity and reliability of the final product and to emphasize the significance of the test specification and piloting, i.e. the dynamics of feedback versus optimization-oriented adjustment. The rationale of the main changes operated in the test specification and test content proper is given, with a twofold purpose, viz. improving assessment quality and sharing experience within a professional arena of debate. The paper should therefore be seen as a problem solving approach to test design, written with a view to identifying criteria of good practice and examples for similar contexts. It also aims to identify paths towards test standardization at intra- and inter-university levels, as test specifications can be, we believe, useful instruments in ensuring test validity, reliability and interactivity, as well as good impact on all the actors involved in the assessment process.

From a wider perspective, evaluation should be seen as having among its purposes supporting an enhancement of the students’ increased responsibility for their learning process and having a positive influence on designing an effective syllabus, able to meet the demands of both the present and the future as far as the learners’ communicative competence is concerned. Evaluation thus becomes a way of measuring progress in time and of reflecting the complexity of holistic learning – and this will induce an extension of the current model over to the pre- and post-program phases, involving its roles in collecting, analyzing and interpreting data on learning/teaching. In such a model, a particular interest goes to finding out exactly when and how development has taken place. The result of this view is, in practical terms, if we envisage the observance of the evaluation efficiency requirements as regards costs and practicality, a model of continuous evaluation that should provide a relatively accurate estimate of the size of the intervention effects, in order to determine not only if an educational intervention has had a statistically significant result, but also whether it has generated an impact that is substantial enough to be considered pedagogically relevant, justifying further investment. In the concrete educational context described in this paper, I was aware of the risks of the language proficiency test project. However, as we must learn to live with human fallibility while endeavoring to overcome it, I tried to observe certain principles in constructing it - although there were not few contradictions to be taken into consideration in a rather intricate picture of constraints, as well as to assume the best options and have them confirmed (or not) by the feedback generated by piloting the test.

2. Qualitative assessment principles revisited – essential theoretical background

Assessment should be understood as a feedback generating process meant to help making educational decisions and to provide the learner the necessary feedback about their strong or not so strong points in learning, but equally to inform educational policy and adequate curricular development (Sanders, 1990). This is confirmed in the literature (Rivers, 2001), who sees testing as supporting learning, encouraging that side of evaluation which can enable the students to demonstrate their competence, rather than the one connected with the strict exclusionary procedure of ranking, an obstacle erected in front of the learners.

The basic qualities of a good test have been long debated upon in the literature, with numerous test designers and researchers being concerned with fundamental questions regarding the necessary reliability level of a test, the validity of scores in relationship with the aim of the evaluation, the best ways of interpreting the test results a.s.o. Of interest for a test designer remains therefore the complexity of aspects to be taken into consideration when creating a form of assessment for a certain context, as there can be numerous such influential factors which may come in various proportions. To validity, reliability, relevance, new elements are added as being significant for the field of
foreign language evaluation, viz. authenticity, interactivity, impact and practicality (Bachman and Palmer, 1996). If we discuss the features of foreign language (English, in our case) proficiency tests, which assess the competence of a person to really apply in actual situations what they learnt, i.e. not directly related to a certain language course, we should mention that they have both washback and washforward effects – providing feedback on what has been taught/learned, but also guiding the (re)shaping of the tuition conducive to qualifying for them. As such, the role of a detailed test specification becomes even more important in designing proficiency test, as it should point out to what successful candidates have to demonstrate they can do. One should not prioritize the quality criteria to be met by a good proficiency test; however, we should be aware of some viewpoints that may help us, as test designers, in making our options and establishing our priorities in a well justified manner. Thus, practicality is a must (BBC Online (2002), as features such as time constraints, financial limitations and easiness in administration and scoring are paramount factors. This does not mean that everything can (or should) be sacrificed for the sake of practicality – therefore validity and reliability remain decisive theoretical requirements for a good test. Validity is enriched in meaning in the literature (Lacity and Jansen, 1994), as it should take into account, beside content, criterion and construct, its implications as a basis for further action, as well as the consequences of score use. We believe this is a justified view point, particularly as far as proficiency tests are concerned, with implications for all the main actors involved – teachers, students, university management and even curriculum designers. In the same vein, a new understanding of washback was proposed (Wall, 1998), that put forth the idea of test impact. Piloting then becomes an important component of the test design process, as it will not only provide feedback required in order to ensure test quality criteria observance, but it will offer pertinent information on the enhanced evaluation duties of teachers/test designers, documenting the duration required in order to create appropriate test forms.

The test specification is the most useful instrument required whenever tests or various test items/parts are designed/reviewed. It can provide detailed input for interested stakeholders at all levels. It can include – in function of context – learner profile data (age, level of proficiency, cultural background etc.), while skill based test specification should cover aspects such as time allowed, test focus, expected skills, source and length of input texts, item types and rubric formats (Alderson et al, 2001). For proficiency tests, the specification should show the test potential to check whether the candidates can use the language specific to certain areas, for instance in the case of an ESP test - in our situation an English for Science and Technology (EST) one - the specific linguistic competence in this particular field is evaluated. The specification should provide an image not only of the abilities measured but also on the manner in which they interrelate. Due to reasons of practicality, information must be included on apparently less important elements, such as length of test/texts, timing of sections, item exemplification. We opinate that, as research has not provided definite answers to the question on what variables do affect construct validity, it may be a good thing to have detailed specifications rather than too succinct ones.

Certain criteria should be remembered in creating a test specification (Rueckert, 1998, quoting Hughes, 2003):

- In terms of content, it should cover: operations, types/addresses/length/topics of texts, structural and vocabulary range,
- In terms of structure – number of items/passages, medium/channel, timing, techniques,
- As to criteria levels of performance: accuracy, appropriacy, range, flexibility and size,
- As regards scoring procedures – rating scale and raters.

It cannot be of low importance to take into account the basic, quite realistic, criterion of practicality, such as the local means of administering big size tests, the time constraints in designing the test, limitations of terms of logistics required to produce/administer and score them. Time and effort required to pilot such a test are worth attaching them importance, it is not pointless to emphasize that – not only at the initial level of having the test piloted with a fellow teacher aware of the context features and parameters, but equally at the level of micro groups that reflect the main elements of the candidates’ profile. Each stage in piloting is useful in getting rid of a range of potentially confusing/erroneous aspects. As pointed out in the literature (Barber, 2003), research does not provide the best criteria in selecting the test piloting context – they should mainly cover, beside practicality and availability of colleagues/students, in our opinion, the criterion of diversity (viz. testees’ technical profile, assumed level of proficiency, assumptions for their short and long-term needs of English).

To conclude at this point, piloting may increase the chances of a test to be clearer and, implicitly, the feedback thus obtained is useful to test designers who also wear the ‘hat’ of teachers in adjusting the structure and methodological approach of the course underlying the test, in order to turn it more effective.
3. Projecting assessment - research presentation and discussion of results

The proficiency test project represented a large scale evaluation activity, whose essential options will be briefly shown and justified in what follows as against the grid of constraints that had to be faced.

3.1. Background and phases

The university imposed to all students to sit for a foreign language proficiency test as a requirement for acceding to the last study year. Students were consequently informed about the test requirements, level of competence expected for a pass (viz. 70 points out of 100, as one aim was to increase the trainees’ motivation to become more competitive on the job market at first entry level), the basic four-term syllabus of the course, comprising General English, English for Professional Communication and EST, as well as a list of the competences/skills expected (reading, listening, writing and language in use). The expected level was, in the Common European Framework of Reference for Languages - CEFR terms, of B2/approaching the C1 level, with the same weighting of the point as the one used for the students’ course regular grades.

Although the first test session was to take place in less than three months, which was a severe constraint of time, the scheduler included a first, experimental, session, when the first test draft was designed, as a specification and test sample, and then piloted in five different faculties. Those were selected in such a manner as to ensure a wide variety among the sampled population as far as their technical profile, level of proficiency and needs of the English language assumptions were concerned. Feedback was then processed and the proficiency test was revised, to be administered in the first formal session to over 1,000 candidates (two testing days, comprising four test variants per day were finally created, with identical format/skill sampling and observing the test specification, but different in terms of input texts). It may be of interest to underline here the fact that, although the test designer of the first draft and the first formal test was the author, at all the subsequent stages the proficiency test meant a team effort in point of invigilating, administering, scoring and feedback providing, in which almost all the department members got involved, thus ensuring the test designing skill transferability.

3.2. Priorities versus constraints – the challenge

The scope of that venture was considerably broader than that of current end-of-term tests, which implied collecting feedback from all the factors involved in it, before passing to designing of a large scale template – hence the importance attached to the specification and the piloting phase as extremely useful tools in enhancing the evaluation quality. The most difficult thing was, in our opinion, to manage to observe the framework of theoretical principles underlying test design, without making any significant concession to the concrete limitations and constraints of the given context, by maintaining a reasonable balance between the criteria of feasibility/practicality and good validity and reliability. The major issues considered priorities in designing the test materials at the initial stage were an appropriate sampling of the content in relationship with the various technical profiles of the students, attaining a test format approaching the estimated average level per university of the students and ensuring a high similarity degree between the test versions created per each evaluation day and/or session. The constraining factors to be faced were of two types, namely those generally encountered in assessment activities and the elements that were context-specific, such as scoring speed requirements versus the large number of candidates, the lack of computer-based scoring and statistical processing of the data and of some logistic support. Moreover, we could not neglect the existence of certain social-human features of influence at all levels/stakeholders (in principal as to the sometimes big differences in terms of students’ proficiency from one faculty to another).

3.3. Specification design, test piloting, feedback and changes

The decisions made were meant to diminish the effect of the constraints and limitations envisaged. Thus, the tested areas were thoroughly covered, in order to avoid lucky guesses. It was therefore a large size test of 100 minutes, in order to increase the testees’ chances that a wrong answer did not affect a fair evaluation of their abilities globally. The one-solution questions were carefully worded so that to ensure that there was indeed only one possible
answer and the instructions and rubrics (in the target language) were clear. The facts revealed by the piloting stage feedback referred to the following non prioritized aspects. If the Writing section scores were taken into consideration, the average group scores were diminished by at least 10% - the reasons mainly included lack of teacher training in scoring subjective tests, differences of proficiency of the testees and a huge amount of time required to score subjective writing tests. Scores per faculties were significantly different, with over 83 point averages for the electric profile faculties, going down to around 73 points for the mechanical profile and to a rather low 61 points for the rest. As a result of feedback analysis, a series of changes were made, which was actually possible due to the flexibility/modularity of the initial test specification. Thus, the Reading section underwent a fairly large transformation, meant to preserve the whole-test consistency in terms of point weighting. Instead of the initial draft with five four-point worth questions and one, rather long, input text, a ten-question formula with two-point items, based on two smaller size texts was preferred, to increase the chances of assessing the candidates’ reading skill level more accurately and fairly. The most substantial change was operated to the Writing section, for which an objective scoring format was preferred to the initially subjective one. In the final test it had two parts, see Figure 1. Distractors were provided in order to increase the challenge while preserving the quite objective character of scoring.

<table>
<thead>
<tr>
<th>Pre-piloting – writing section specification</th>
<th>Post-piloting – writing section specification</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. WRITING</td>
<td>5. WRITING</td>
</tr>
<tr>
<td>Essay writing, the topic and introduction given.</td>
<td>Two parts.</td>
</tr>
<tr>
<td>Topics: University life, environmental issues, EST at general level, etc</td>
<td>Tasks:</td>
</tr>
<tr>
<td></td>
<td>1) filling in the gaps in five small contexts with linking words/phrases (5 items x 2 points = 10 points); three distractors in the list of choices.</td>
</tr>
<tr>
<td></td>
<td>2) completing a gapped letter with the appropriate missing parts, by selecting from a list containing three distractors – 10 points.</td>
</tr>
</tbody>
</table>

Fig. 1. Pre- and post-piloting writing section specification (abridged)

The Listening section was amended in terms of reformulation of the multiple-choice items, which were initially too long, of the type that is usually used for after-listening tests. They were replaced with while-listening items, based on news in the technical and scientific domains input, checking comprehension at sentence and word level mainly, as well as mental inferences from the context. The Language in use section was only subjected to editing and formatting changes, for the convenience of both the test writer and the candidate in understanding and solving the questions.

4. Open concluding remarks

The test specification proves to be a useful instrument in enabling test designers to keep trace of the changes made to the test. It also informs the next test writers about a range of aspects, thus ensuring flexibility in proficiency test designing. Moreover, it can also help test designers in their teaching activity, a good reason for them to continue to revise their products in a process approach to conceiving assessment forms, while preserving the most relevant theoretical principles and reducing the constraints impact.

Moreover, we maintain that teachers who embark upon designing valid reliable tests should use the test specification flexibly when they recurrently make choices and prioritize options. Thus, their endeavor has good chances to contribute to their development as reflective professionals.
References


