Challenges of Test Compilation and Evaluation Encountered by 1st year Professional Master Degree Students in "Assessment, Testing and Standards of Foreign Language" Module

Lorena Robo

Fan S. Noli University, Korçë, Faculty of Education and Philology, Foreign Language Department, Albania

Abstract

The aim of this paper is to present the challenges students of master degree and teachers face when it comes to test compilation. Assessment is an important process of learning which intends to provide a systematic indication of the quality of students' learning for both teachers and students. It maintains standards in professional education and in higher education generally and motivates students throughout their studies. The study highlights the difficulties students encountered while designing a test during the "Testing, assessment and standards of foreign languages" module. The focus of the test compilation was the set of 12 lectures covered in this module. The study was conducted under a task driven research. After students designed the tests they were analyzed according to criteria of reliability and validity. The paper also explores the challenges of peer assessment through the analysis of tests. The paper brings insights to student potential in designing tests as well as peer assessment challenges. Recommendations are made about improving the design of tests and ways to reduce the difficulties are implemented.

Key words: peer assessment, challenge, reliability, validity, compilation

1. Theoretical background

According to McNamara, Tim (2014:3) testing is a universal feature of social life. Throughout history people have been put to test to prove their capabilities or to establish their credentials. Meanwhile testing is an important part of every teaching and learning experience. There are many reasons for developing a critical understanding of the principles and practice of language assessment. Language tests play a powerful role in many people's lives, acting as gateways at important transitional moments in education, in employment, and in moving from one country to another. As McNamara (2014:4) states language tests are devices for the institutional control of individuals, it is clearly important that they should be understood, and subjected to scrutiny. Secondly, working as a teacher or professional teaching to a test, administering tests, or relying on information from tests to make decisions on the placement of students on particular courses. Thus understanding language testing is relevant both for those actually involved in creating language tests and in a broad sense for those involved in using test or the information they provide, in practical or research context (McNamara, 2014:5).

Tests differ with respect to how they are designed, and what they are for, otherwise to test method (paper-and-pencil test and performance tests) and test purpose (achievement and proficiency tests). Paper-and-pencil tests are used for assessment of different or separate components of language knowledge (grammar, vocabulary, etc.) or of receptive understanding (listening and reading comprehension) constructed under the multiple choice format. Performance tests are commonly tests of speaking and writing in which language skills are assessed in an act of communication. On the other hand, achievement tests are associated with the process of instruction which accumulate evidence during, or at the end of, a course of study or the period to gather information on the degree of student progress. Proficiency tests refer to the future of language use without necessarily any reference to the previous process of learning.

2. The design of language tests

The view of language and language use is essential to the activities of designing tests and interpreting the meaning of test scores. Referring to terms of test construct focus is exerted on the knowledge or skill possessed by the candidate being assessed. Understanding what view the test takes of language use in the criterion is necessary or determining the link between test and criterion in performance testing. According to what view the test takes, the ‘look’ of the test will be different, reporting of scores will change, and test performance will be interpreted differently (McNamara 2014:13).

The most influential multi-componential model of Communicative Language Ability (CLA), that of Bachman (1990), provided test developers with a wide ranging account of CLA and useful theoretical questions to ask in design of language tests. However, a critical weakness of the model is that it proved to be extremely difficult if not possible to operationalize for its lack of clear prioritization as to what might constitute criterial parameters for language testing purposes (Kunnan, 1998; McNamara, 2003; Chalhoub-Deville and Deville, 2006). Bachman’s model (1990) has contributed less than might have been hoped to empirical test validation. McNamara (2003:468) emphasized this point and argued that “those who have used the test method facets approach have found it to be difficult to use, and it has in fact been implemented in relatively few test development projects…” He also criticizes the model for being essentially psychological, seeing communicative language ability as a mental ability, while the context of use is increasingly understood theoretically as a social arena, as in virtually all current work in discourse analysis. After many linguists’ viewpoint testing is seen primarily a “social phenomenon” - in which test scores, test score users and score use contexts are “inextricably linked”.

On the other hand, Deville and Deville (2006) is supportive of Bachman’s model to the extent that it addresses “issues related to language use”, but agrees with McNamara view as it represents an essentially psycholinguistic view of performance and is largely missing important interactional and sociolinguistic elements. Meanwhile, work has been carried out at the same time on the Common European Framework of Reference for Languages (CEFR), which aimed to be directly useful to both testers and teachers as a descriptive of framework of language ability over series of distinct levels (Council of Europe, 2001).

Furthermore, as test construct is considered, John Oller in 1970s offered a new view of language and language use underpinning tests focusing less on knowledge and more on the psycholinguistic processing involved in language use. Language was seen as involving two factors: (1) the on-line processing of language in real time (in speaking and listening activities), and (2) a ‘pragmatic mapping’ component which implies the way the formal knowledge of the systematic features of language was drawn on for the expression and understanding of meaning in context. Oller thus proposed the Unitary Competence Hypothesis, the performance on a whole range of tests, termed pragmatic tests, able to integrate grammatical, lexical, contextual, and pragmatic knowledge in test performance. Cloze tests measured the same kinds of skills as pragmatic tests. Such tests resulted in the same information of readers’ abilities, they were easy to construct, easy to score and more attractive than elaborate and expensive tests. Despite the drawbacks cloze tests are widely used nowadays since 1970s.

3. The study

The study aims at exploring the significance of test design and challenges of students in test compilation. The following research questions guide the study.

1. What difficulties did the students encounter while the test compilation?
2. Did the tests meet the criteria of reliability and validity?
3. How was peer assessment done related to fairness?

3.1 Method

3.2 Participants

The participants of the study were 32 students of Professional Master of “Teacher of English Language” study program, in the department of Foreign Languages, Faculty of Education and Philology, “Fan S.Noli” University. Participants outnumbered 24 females and 8 males.

3.3 Research tools and materials

The research questions were examined by conducting a literature review of relevant studies of criteria and design of tests through the past years. The study was carried out through a qualitative approach. The instrument of the study were 12 lectures and the test compiled by the first year students of Professional Master Program in “Testing, assessment and standards of foreign language learning” module. Students have to construct the test based on criteria of validity and reliability. Each student was at the same time a test compiler and a student as well. After test designation students underwent the process of peer evaluation. Data were driven out of the research and the results were empirically interpreted. Focus groups and observation was used to analyze and interpret the task given to the students and to generate results according to the research questions raised in the study.

3.4 Discussion

Assessment is a very important element in learner’s education. Besides measuring the amount of knowledge an individual has gained in a specific subject, it provides the teacher with information regarding the progress of the student and sets the goals of the teacher what and where to continue with the preparatory and explanation stage of the lesson. The aim of this study was to identify the main obstacles and difficulties students positioned in the role of the teacher encounter during the test compilation. The focus of the study has been the parameters under which students designed the test, to what extent did the test meet the criteria of reliability and validity, transparence and equality as well. Tests were designed first and then were peer-evaluated. Every stage of the process was monitored and controlled by the teacher.

Suggestions and recommendations were given according to the accuracy and test relatedness to the overall aim of the task-based approach.

To begin with, students were given a task as homework to design a test on the group of 12 lectures they have taken in the module of “Testing, assessment and standards of foreign language teaching”. Since students have not yet done active or passive practice in public primary and secondary schools of the town, it was better thought to analyze and design the test on the group of lectures they have taken, as they knew well the content and information provided, facilitating the task of the students. In the group of twelve lectures they have studied, they knew well the concepts of reliability, validity and transparence. They were already known with the types of tests and the characteristics of each one of them, so they had a clear view of the aim of the task.

According to Wanner, Thomas and Palmer, Edward (2018:1032), self and peer-evaluation are becoming central aspects of student-cantered assessment processes in higher education, increasing evidence that both forms of assessment are helpful for developing key capabilities in students’ capabilities of taking more responsibility for their learning, developing critical reflection skills, developing a better understanding of subject material, assessment criteria and their own values and judgments. According to them, many researchers have pointed out that self and peer-assessment requires careful design and implementation for it to be an effective tool for formative assessment processes and the development of students’ capacities for giving feedback, and the continuous and timely involvement of the teacher.

3.5 Data analysis of the test construct

The task of the students as previously described was to construct a test based on a variety of lectures (12) studied in the module of “*Testing, assessment and standards of foreign language*”. After the tests were designed by the students, in total 32 of them, they were interpreted according to the following elements:

1. Types of exercises used by the students to construct the test
2. The analysis of the content of the questions and lecture information coverage
3. The amount of time specified to do the test
4. The validity and reliability fulfillment criteria
5. The overload of the test (the number of questions included in proportion with the time available)
6. Peer-evaluation consequences and results

The module was done during online teaching the previous academic year, the pandemic year (2020-2021). The task was given as assignment in Teams Platform. Students attached their files and then they presented the questions of the test during the online seminar. The purpose of the test was not only to design it but also to undergo peer-evaluation. Students were monitored and asked to explain the purpose of doing what and how.

The discussion of the test was carried out using the task driven approach. After they have fulfilled the task, students were organized in 8 focus groups each with 4 members. They have first to exchange the test compiled and after a peer evaluation conducted, the tests were evaluated under the criteria of reliability, validity, the content of the questions, and their level of importance under the teacher surveillance. Observation was used as a technique to monitor student’s work. As students have to compile the test within a 2 hour class guided observation was a method used to comprehend their work in progress. In the first class students were given the appropriate time to focus and highlight the most important parts of the lectures and the second class they have to compile the test according to the criteria predetermined. Direct observation was conducted even in the process of peer evaluation. Students were asked to justify the reasons of the specific assessment done to their peer's test as well. The observation was done through the platform where each one of the students had their cameras opened and could be guided and controlled online.

Analyzing the focus group work, out of 32 students' participant in the group, only 25 of them handed in on time the assignment, and 7 of them did not respond to the task at all.

Based on the overall process observed through focus groups tests were analyzed due to the following criteria:

1. According to the types of exercises as far as the tests analysis considered, the most common questions were: multiple choice exercises, open or direct questions, open cloze exercises, give the definitions of the terms, match the definitions with the meanings, true or false exercise, complete the blank spaces, given the cluster and complete with the characteristics (rarely), compare and contrast exercise presented in form of a table and the essay.

- Out of 25 tests handed in, 7 students included one multiple choice exercise. Open questions were present in all the tests handed in. Some students included 1 some others ranging from 2 to 3, and 4 of them only open questions ranging from 3 to 4 questions per test.
- 12 students have compiled one True/False exercise in their test. 5 of them have included the exercise of defining the terms; they have either given the terms and asked their peers to give the definition or have given the terms and provided the term/notion explanation where students have only to put the right term.
- One student has included the essay to be written with 1000 words, beside the two other exercises (one True/False and another one multiple choice) which in fact did not match the time provided with the number of the exercises given (40 minutes). 4 of the students compiled compare and contrast exercise, one student has designed a chart to be completed with the missing information (Canale and Swarn model of communicative competence). 4 students have included 'fill in the blanks' exercise and one of them 'match the definition with the terms' exercise.
- Most of the tests contained from 4 to 7 or 8 questions. In most of the tests, in 17 of them, time was not specified in the test paper, only 8 of them have determined the time ranging from 30 to 40 and 45 minutes. In some cases, time did not respond to the total time provided.
- Only 2 of the students have given the table of points at the end of the test and the corresponding marks.
- 2 students were out of topic at all. They have designed an English test in general including grammar, vocabulary, listening and pronunciation (the four skills).
- What was noticed in some of the tests was the variety of exercises ranging from multiple choice exercises up to True/False, term definition, open questions, compare and contrast exercise, open cloze, etc. Out of 25 students' test compiled papers in only 3 of them did we see a variety of exercises. In most of the other papers there were mainly one to two or three types of exercises.

- 7 out of 32 students in total of the group who did not respond to the task given were negatively assigned in Teams Online Platform (the platform we used to teach online during the two years of the pandemic in Korça University, Fan S. Noli beside the Classroom Platform, Zoom and Google Meet as well during lectures presentation).
- One student has included a feedback of the activity:

Feedback: *After I shared this test with the group, that in my case I had the teacher's role and two other persons were students, I found them very motivated to complete their test and to send me back in the time. The time allowed was one hour and in the test were 6 exercises, each of them with different points.*
Two members of my group were willing to complete the exam, they were focused on what I prepared for them and even in cases they faced with any struggle, they asked me to explain better; and I really appreciated that fact that they asked me politely.
I enjoyed my experience as a teacher for an hour and the feedback and grades for my test and exercises I created were high and this shows better the fact that they reading and comprehension skills were well-organized and well- developed about lectures. ***(K. Rrukaj)***

2. As far as the lecture content and the inclusion of the information in the test is considered, in almost all the papers students have marked out considerable questions valued from the range of importance. Most of them have included differentiation of the types of tests, definition of terms, open or direct questions, compare and contrast exercise between linguists' viewpoints, types of assessment, etc. In tests where there were a variety number of exercises the information included covered a wider range of material.
3. Time available to do the test was in most of the cases not specified for the sake of truth. In these cases, students were asked during the seminar for the time limit of their peer and if it corresponded to the time they have planned. Time should be included in the test paper. This was an advice given to them so that the student that underwent the test (their peer) could manage the time while answering the questions.
4. Since students have clear concepts of the criteria of validity and reliability explained in the lecture, their tests were analyzed under these criteria as well. What was noticed in the majority number of tests was the meeting demand of the criteria aimed. Students were made known the cases when their tests lack these criteria. To measure these criteria, and to conclude with correct outcomes the same test was given to two different students and it was seen whether the result (answers) were the same or not.
5. In some cases it was seen a discrepancy or inconsistence between the number of questions and the time provided. In some tests there was enough time available, few questions for the limit amount of time. Another important element noticed was good students compiled difficult tests, included a variety of questions, comprehended more information from the lectures, and their peers struggled for time when answering the test questions. While the others, usually compiled easy questions, included less exercises and had more time to solve them.
6. As far as peer-evaluation was an important intention of this study, the task-based approach resulted to be a motivating and entertaining activity for the students. They felt at the same time a teacher and a student, and the task resulted to be an effective practical activity, where students learned to compile and evaluate at the same time.

3.6 Overall discussion and conclusions

This study aimed at identifying the main challenges and difficulties students encountered at the same time in the role of the teacher and the student, during the task-based approach of the test compilation and the qualitative data gathered from the tests handed in on the online Teams platform.

The study revealed positive and negative consequences while designing the test. As students were for the first time put in the role of the teacher, responsibility upon compiling a 'good' test that should measure student learning capacities arose. As far as the results of the tests were seen students showed professionalism and readiness to construct a test according to criteria of reliability, validity and in concordance with the lectures taken in "Testing, assessment and standards of foreign language" module.

Regarding the test content variety of the questions and types of exercises, students revealed a high competence in selecting carefully and with attention exercises. This tendency was seen in most of the students' tests. A few of them had a lack of exercise variety mainly focused on open or direct questions.

Time was an important factor related to test compilation. As far as the students' work is concerned the majority of the tests 68% did not include the amount of time needed to do the test. This should be an important element of the test. However, time might have been specified orally to the peers but students were recommended to involve it for some reasons. First, it helps the students to better organize the time required for each exercise. Second, it sets clear objectives of the teacher to measure student ability to answer in relatedness with the time. Third, it accomplishes the layout of the test format.

Referring to the criteria of reliability and validity, students showed a better understanding of the test components. The tests were answered by the peers so the results were easier identified under these important components.

Peer-evaluation was a crucial phenomenon for increasing student's self-confidence and raising awareness of the learning results in developing a better understanding of the subject material, developing critical reflection skills, assessment criteria and their own values and judgments. However, self and peer-assessment requires careful design and implementation for it to be an effective tool for formative assessment processes and the development of students' capacities for giving feedback, and the continuous and timely involvement of the teacher.

The qualitative data drawn from the study showed satisfying outcomes of the student potential in designing tests that surpassed the expected compilation capabilities. Designing a test is not a simple and careless activity and it should be carefully treated by the teachers, since it sets the learning outcomes of the students and is an important factor in determining the point of regress or student's progress in the learning environment.

The aim of the study was to identify the main difficulties and challenges students encountered during the process of test compilation through a qualitative and quantitative research. Since it was a new and unexplored activity of their learning time during their university studies the following advantages and disadvantages were outlined as positive and negative outcomes of the study.

- The positive attitude of the students in their role
- Felt motivated, they took the role of the teacher and the student at the same time

- Were happy and energized of the good practice
- Wanted more activities of the kind
- Enjoyed peer group work
- Could be able to compile exercises on the group of lectures available, to sort out the most important information and to select the appropriate exercises for the test
- It was a challenge of themselves to design tests as future English teachers and to understand the weak and strong points of the test compilation
- Seminar discussion was an advantage- learnt what could be done better and where it had place for improvement
- Peer evaluation was a challenge in itself
- The task provided student interaction with one another
- It was concluded with high requirements for more activities of the kind during the module extension

Drawbacks

- Time not in concordance with the amount of questions
- Not well constructed tests, did not include the most important information
- Peer evaluation was very subjective; it was not fair in some cases. They might have had the pressure of the grade from the professor
- The lack of all the elements of a test format name, surname, date, time, table of points and the corresponding marks
- Students took more time than the limited amount of time provided in some cases

Assessment is a very important element in setting clear and specific goals of the learning environment for the teacher. It is a crucial aspect of measuring student's needs and interests of the subject matter and it provides a great opportunity in progress determination, achievement and success ranking of the learner.

As it is commonly recognized that assessment literacy, the understanding and know-how about assessment principles and practices, is a valuable and essential skill for both actors in the process of learning. Good practice in language assessment, finding the most appropriate tools for each purpose, requires commitment and expertise.

Taking assessments to demonstrate proficiency and using assessment results to make decisions generate challenges for both students and teachers. These challenges are twofold: one is based on the teacher's or administrator's role, while the other is focused on the student's perception and response. This study has aimed at focusing on the main challenges students encounter while test compilation. As a result, testing is a very important element in the process of student learning and education. Self-assessment, peer and educator feedback, and one's ability to self-monitor and reflect on personal learning styles are fundamental factors that influence achievement. These methods of assessment encourage students to judge and evaluate their own performance as well as their peers. It encourages a more active role in one's learning and empowers students to be lifelong learners. (Carless, 2015)

Further studies in the field are recommended to be conducted in order to have better and clearer results of the assessment process in the learning environment in the continuous and ongoing process of education.

References

Bachman, L. (1990). Fundamental considerations in language testing. New York: Oxford University Press.

Carless, D. (2015). *Excellence in University Assessment: Learning from Award-Winning Practice*. London: Routledge.

Chalhoub-Deville, M., & Deville, C. (2006). Old, borrowed, and new thoughts in second language testing. In R. L. Brennan (Ed.), Educational measurement (4th ed.) (pp. 517–530). Westport, CT: American Council on Education/Praeger.

Gilbert, R. (1992). Text and context in qualitative educational research: Discourse analysis and the problem of contextual explanation. Linguistics and Education, 4(1), 37–57

Madson, Harold S. (1983). *Techniques in Testing*, Oxford University Press.

McConlogue, T. (2015). “Making Judgements: Investigating the Process of Composing and Receiving Peer Feedback.” *Studies in Higher Education* (9): 1495–1506.10.1080/03075079.2013.868878.

McNamara, Tim (2014). *Language Testing*, Oxford Introductions to Language Study Series Editor H.G. Widdowson, Oxford University Press.

Oller, John W “Transformational theory and pragmatics”,(504-507), Volume 54, Issue 7, in *The Modern Language Journal* November 1970, available in https://doi.org/10.1111/j.1540-4781.1970.tb03585.x.

O’Sullivan, Barry (2011), *Language Testing: Theories and practices*, Palgrave Advances in Linguistics, Palgrave Macmillan.

Yucel, R., F. Bird, J. Young, and T. Blanksby. (2014). “The Road to Self-Assessment: Exemplar Marking before Peer Review Develops First-Year Students’ Capacity to Judge the Quality of a Scientific Report.” *Assessment & Evaluation in Higher Education* 39 (8): 971–986.10.1080/02602938.2014.880400.

Warner, Thomas & Palmer, Edward “Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback”, (1032-1047), Volume 43, Issue 7 in *Assessment & Evaluation in Higher Education Journal*, Published online on 18 January 2018, available in https://doi.org/10.1080/02602938.2018.1427698.