Małgorzata Gos

Crossing the Barriers in English Language Testing: a Historical Overview

Abstract

In her article the author presents the historical overview of language testing. She stages of the development of testing methods and techniques as well as three main approaches to language testing proposed by Spolsky in 1975 and following them, the communicative approach which continues to this day:

- Pre-scientific period (traditional) – in the second half of the 19th century, closely associated with the traditional approach to teaching – grammar-*translation* method;
- psychometric-structuralist period – beginning in the 1920s – until 1960s, in which testing of individual language elements was introduced – so-called atomic testing;
- psycholinguistic – sociolinguistic period – mainly in the 70's of the last century, when *integrative / global testing* was introduced;
- communicative period – in which the concept of communicative competence appeared;

In the article the author also describes basic types of language tests and then presents the characteristics of a good test – that is the concepts of accuracy, reliability, impact and practicality.

Key words: *testing, testing methods, testing technique, test type*

Abstrakt

W swoim artykule, autorka prezentuje historyczny rys testowania językowego, opisując chronologicznie poszczególne etapy rozwoju technik i metod testowania. Zaprezentowano tu trzy główne okresy w układzie zaproponowanym przez Spolsky'ego w 1975r. oraz następujący po nich i trwający do dziś czwarty okres – komunikacyjny:

- okres przednaukowy (tradycyjny) – przypadający na drugą połowę XIX wieku, ściśle związany z tradycyjnym podejściem do nauczania języka, tzn. *metodą gramatyczno-tłumaczeniową*;
- okres psychometryczno-strukturalny – przypadający na lata 1920-1960, w którym wprowadzono testowanie poszczególnych elementów języka, tzw. *testowanie atomistyczne;*
- okres psycholingwistyczno-socjolingwistyczny – przypadający głównie na lata 70 ubiegłego stulecia, kiedy to wprowadzono *testowanie integracyjne*;
- okres komunikacyjny, w którym wprowadzono pojęcie *kompetencji komunikacyjnej.*

Autorka opisuje także podstawowe rodzaje testów językowych, a następnie przedstawia cechy *dobrego* testu – to jest trafność, rzetelność, wpływ oraz praktyczność.

1. Approaches to language testing: historical overview

Formal testing of the English language dates back to the mid-18th century. That is when, because of British colonial expansion, Latin lost the position of the language of education and the necessity of English examinations for foreigners appeared. Since then, English has become an international language and its role has expanded to such an extent that it has been introduced to school curricula as a compulsory subject, and the demand for courses of language for specific purposes arose. As Davies notices, language testing development comes "(...) from a long and honourable tradition of practical teaching and learning need" (Davies: 1990, 9). Thus, language testing methods and techniques, have been changing over the years and different approaches to it have been presented.

During the Congress of Applied Linguistics, Spolsky suggested three approaches to modern language testing: pre-scientific (later called traditional), psychometric-structuralist (to which he refers as

modern) and psycholinguistic-sociolinguistic (or post-modern) (Spolsky in: Rivera: 1984, 4).

1.1 Pre-scientific / traditional approach

The approaches are summarised in Skutnabb-Kangas, who claims that:

> The pre-scientific attitude would be the one characterised by an uncritical reliance upon an intuitive view of what are the important variables to be measured, with no attempt made to systematise them or give them a basis in theory about language as a system, language use, or bilingualism (1981, 219).

Spolsky explains that "(...) the traditional approach was, and still is, a method of examining rather than testing" (in Rivera: 1984:4). At this stage, language testing was based mainly on two test types: essays and translations. Krumwiede puts it this way:

> In this phase, test items included only translation, composition and sentence completion type exercises. Language courses followed a grammar translation approach, and students were supposed to "know" the language if they could translate properly. Oral skills were not taught. The testing went along these beliefs of learning the language (2008, 11-12).

Many objections to and attacks on traditional testing were due to the fact that the examinations relied on the judgement of one examiner. Spolsky states that "(...) as with many other things in traditional life, it is fundamentally elitist, being based on the assumption that certain people have the authority to make judgements about others. The system worked (and in some parts of the world still works) well, as long as the authority of the judges or examiners is not questioned" (1984, 4), which is unlikely in the process of democratisation and modernisation. He sees, however, a strong,

positive argument in favour of the pre-scientific approach:

> (...) it involved a direct encounter, face-to-face or on paper, between two people, the examiner and the candidate, which made the examiner aware of the candidate as a human being, with personality displayed either in person or writing style. In this way, the examiner was reminded of the potential implications of a decision to pass or fail. There was authority (and the lack of questioning that is implied), but was also a situation that constantly kept the examiner in mind of his or her responsibility (1984,4).

However, following Spolsky's views (1984), Smith states that "the major criticism of this form of testing is its lack of reliability" (1994, 6).

1.2 Psychometric-structuralist / modern approach

During the **psychometric-structuralist** era, beginning in the 1950's, "(...) tests were related to existing theories about language of the time. Problems with the language were thought to be due to transfer from the first language. It was the time of the drill patterns, concentrating on single elements and repeating them over and over in order to learn them. Contrastive analysis of both languages in bilinguals was conducted" (Krumweide: 2008, 10). As language learning was believed to be a linear process of completing one task after another, language tests were supposed to measure learners mastery of particular items. That kind of tests were called **discrete point tests** and they reflected the view that language could be broken down into linguistic components.

> Each of the(se) elements of language constitutes a variable that we will want to test. They are pronunciation, grammatical structure, the lexicon, and cultural meanings. The first of these,

> pronunciation, is itself made up of three separate elements, namely sound segments, intonation and its borders, and stress and its sequences which constitute the rhythm of the language. Within grammatical structure there are two main subdivisions, namely morphology and syntax. Syntax will be given priority in testing. Morphology will be treated as much as possible in connection with syntax (Lado: 1961, 25).

To avoid an individual's personal judgement of the learners' performance (Smith: 1994) and because of the connection of structural linguistics with psychometrically based testing, there was a great attempt to find objective methods to measure language. That is why "(...) the test items focused on isolated and discrete elements, decontextualised phonemes, grammar and lexicon and used multiple choice, true – false, and other types of objective items' (Shohamy: 1999, 141).

Smith also points out that reliability of the tests increased 'at the cost of decline in validity' (1994, 7). The same problem was also indicated by Weir, who claims that although the tests are efficient and reliable, they "suffer from the defects of the construct they seek to measure" (Weir: 1990, 2). Oller outlines the deficiencies of this approach in terms of construct validity:

> Discrete point analysis necessarily breaks the elements of language apart and tries to teach them (or test them) separately with little or no attention to the way those elements interact in a larger context of communication. **What makes it ineffective as a basis for teaching or testing languages is that crucial properties of language are lost when its elements are separated**. The fact is that in any system where the parts interact to produce properties and qualities that do not exist in the part separately, *the whole is greater than the sum* of its parts (...)

> organisational constraints themselves become crucial properties of the system which simply **cannot be found in the parts separately** (1979, 212).

During the research, the findings above were supported by the fact that students who performed well on discrete point tests were frequently not able to communicate in real-life situations using the target language, and the opposite, the ones who studied the language in the actual country, scored poorly on the tests (Krumweide: 2008,10). It is not surprising, considering the fact that (...) knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language (Morrow: 1979, 145 in: Weir: 1990,3).

1.3 Psycholinguistic-sociolinguistic / postmodern approach

The main goal of testing at **psycholinguistic-sociolinguistic** stage in the 1970's is to test functional language competence. Here all language skills – listening, speaking, reading and writing are tested and many language items are supposed to be assessed simultaneously. This kind of testing is referred to as **integrative or global testing** and as Oller (1979) notices, it enables to measure the actual process of the use of target language. He compares types of tests used in **modern** and **postmodern** eras:

> The concept of an **integrative test** was born in contrast with the definition of a **discrete point test**. If discrete items take language skill apart, integrative tests put it back together. Whereas discrete items attempt to test knowledge of language one bit at a time, **integrative tests** attempt to assess learner's capacity to use many bits all at the same time, and possibly while exercising several presumed components of

> a grammatical system, and perhaps more than one of the traditionally recognised skills or aspects of skills (37).

Spolsky believes that the **postmodern approach** caused a major change in language testing and explains the very beginning of the idea:

> The (...) approach, which I called **psycholinguistic-sociolinguistic**, and which includes the ethnographic approaches, grew out of questions raised by the various disciplines about what it means to know and to use a language. Whereas the **modern approach** tried to break down knowledge and behaviour into discrete components that lent themselves to statistical handling, the new approaches are struggling to deal with larger, natural chunks (1994, 5).

At the **postmodern** stage, language theorists and practitioners realised how difficult language testing was. They also tried to develop "more realistic methods of measurement" (Skutnabb-Kangas: 1981, 219) and began to take purposes of the tests under consideration.

Instead of the terms **postmodern** or **psycholinguistic-sociolinguistic**, Shohamy (1996: 141) calls the period *the communicative era*. She argues that although the *communicative approach* to teaching dominated the time, *non-communicative* tests were used to check language knowledge. Soon, language testers realised that a language test should reflect real-life language in real-life situations. Thus, *integrative* and ***communicative testing*** became widely used. Although both types of testing are performance-based, they are two different phenomena. Whereas *integrative testing* rejects discrete point techniques, *communicative approach* combines both discrete point and global testing techniques and, as Komorowska (2002) notices,

it deals with the development of both language skills and elements.

1.4 Communicative approach

For some researchers *communicative testing* falls into postmodern or psycholinguistic-sociolinguistic period (Spolsky: 1975; Valette: 1977), however some other researchers (Weir: 1990; Komorowska: 2002) view it as a separate approach.

According to Shohamy (1996: 142), the idea of *communicative competence* resulted indirectly from Chomsky's theory (1959), who as a response to a mechanical concept of language acquisition, claims that individual personal features of a learner have a great impact on language acquisition, and thus language acquisition is an active individual process. He also suggests that people are born with some universal linguistic knowledge, which he calls *Universal Grammar*. What is more, he believes that human beings have an internal tool, which enables language acquisition – *Language Acquisition Device (LAD).* During the process of acquiring first language, the tool helps children make hypothesis about the language and its structure, which consequently leads to establishing the rules that govern the language and eventually, to modify and improve them. These rules are called *language competence.*

In opposition to Chomsky's *language competence*, Hymes (1972) introduced a broader concept of *communicative competence***,** which apart from the knowledge of grammar rules includes the ability to use them effectively and in the way appropriate to the certain social situations. As Spolsky points out:

> The *communicative teaching* approach postulated that the second language learner must acquire not just control of the basic grammar of the sentences but all the communicative skills of a native speaker; it

> seemed easy to call these skills *communicative competence* (1989, 139 in: Shohamy: 1996, 142).

Although Spolsky talks about communicative teaching and not testing, Shohamy explains that "the connection between **communicative teaching** and **communicative testing** was direct and straightforward as language testing reacted to developments in language teaching" (1996, 142).

Canale and Swain (1980) support the concept of **communicative competence** since, in their opinion, grammar rules are meaningless without rules of use. They also claim that there are **four** components in communicative competence**: grammatical** (knowledge of the grammar rules), **sociolinguistic** (knowledge of the rules of use and discourse rules)**, strategic** (knowledge of communication strategies, both verbal and non-verbal) **and discourse competence** (cohesion and coherence). Shohamy explains what the particular competences refer to:

> **Grammatical competence** included knowledge of lexical items and of rules of morphology, syntax, sentence-grammar, semantics and phonology**. Sociolinguistic competence** included knowledge of sociocultural rules of use. **Discourse competence** was related to mastery of how to combine grammatical forms and meanings so as to achieve a unified spoken or written text in different genres, and **strategic competence** referred to the possession of 'coping strategies' in actual performance in the case of inadequacies in any of the other areas of competence (1996, 143).

The Bachmanian framework (1989; 1990) for testing **communicative competence** is consistent with earlier findings and definitions, and is referred to as **Communicative Language Ability (CLA).** He adds, however, one more component to the four suggested by

Canale and Swain – psychophysiological mechanism, and he describes **CLA** as follows:

> **Communicative language ability** consists of **language competence, strategic competence, and psychophysiological mechanism**. **Language competence** includes organisational competence, which consists of grammatical and textual competence, and pragmatic competence, which consists of illocutionary and sociolinguistic competence. **Strategic competence** is seen as performing assessment, planning and execution functions in determining the most effective means of achieving a communicative goal. **Psychophysiological mechanism** involved in language use characterise the channel (auditory, visual) and mode (receptive, productive) in which competence is implemented (Bachman: 1989 in Weir: 1990: 9).

Consequently, Bachman himself and Palmer (1996) elaborated on Bachman's model of **communicative competence** further to include both affective and metacognitive factors. As Byram (2004, 48) suggests their model of CLA is used as the theoretical basis for many international tests (e.g. the International English Language Testing System IELTS) and current research projects.

It is also worth mentioning that Stryker and Leaver agree with Spolsky that testing for **communicative competence** has a great impact on **proficiency** tests. They state that if "proficiency means the ability to communicate with native speakers in real-life situations, then a *proficiency test* must involve such spontaneous interactions" (1997, 23).

The connections between **communicative competence** and **proficiency** are also pointed out by Omaggio (1986 in: Chun: 2002), who believes that if the term **communicative competence** refers to *knowledge* about language and to the *use* of the

knowledge, it is similar to the notion of **proficiency**. In contrast to the components of **communicative competence** presented above, he gives three interrelated criteria to describe **proficiency**. These are: **context, function and accuracy**.

> The term *proficiency* includes specifications about the *levels* of competence attained in terms of the *functions* performed, the *contexts* in which the language user can function and the *accuracy* with which the language is used. Thus the notion of *proficiency* enables us to broaden our understanding of *communicative competence* to include more than the *threshold level* needed to simply get one's message across (2002,115).

Finally, it must be mentioned, as Shohamy (1999) points out, that although communicative language testing still dominates the field, performance and alternative assessments are getting more and more common. She adds that "(...) performance assessment is based on the interaction between language knowledge and specific content, usually of the workplace or of professional preparation. Test takers perform realistic tasks which call for the application of skills to actual or simulated settings in an attempt to replicate the language needed in these contexts. Thus, performance tests are task-based, direct, functional and authentic" (143).

2.Test types

According to McNamara, language tests differ with respect to their design and aim or as he puts it, "in respect to **test *method*** and **test *purpose***" (McNamara: 2002, 5). As far as the **method** is concerned, he distinguishes **paper-and-pencil tests** from **performance tests**. He explains that **paper-and-pencil tests** are traditional examination question papers and are typical while assessing either separate language

components, like grammar or vocabulary, or receptive understanding, which includes listening and reading comprehension. In standardised tests, test items are often in **fixed response format** with a number of possible responses presented to the candidate, who is supposed to choose the best alternative. McNamara argues that although that kind of tests are efficient to score and administer, "(...) they are not much used in testing the productive skills of speaking and writing, except indirectly" (2002, 6).

Performance tests, in which "language skills are assessed in an act of communication" (McNamara: 2002, 6), are common tests of productive skills, such as speaking and writing. Here, the samples of speaking and writing elicited in the context of simulation of real-life tasks and situations are assessed by one or a group of trained **raters** using an established **rating procedure**.

As for the **purpose**, McNamara distinguishes between two types – **achievement** and **proficiency tests** and explains the difference as follows:

> Whereas *achievement tests* relate to the past in that they measure what language the students have learned as a result of teaching, *proficiency tests* look to the future situation of language use without necessarily any reference to the previous process of teaching (2002, 7).

Sharma (2002) claims that **achievement tests** "(...) are aimed at finding out the quantum of language skills acquired by a learner during the course of instructions" (180). In other words, they are supposed to assess the learner's knowledge that has been learned during the language course and how much of the syllabus the learner has adopted.

McNamara (2002) adds that because **achievement tests** "(...) accumulate evidence during, or at the end of, a course of study in order to see

whether and where progress has been made in terms of the goal of learning" (6), they ought to support the process of teaching to which they relate. He notices that **achievement tests** tend to be innovative and also reflect progressive aspects of the syllabus and, thus, they are associated with "(…) some of the most interesting new developments in language assessment in the movement known as **alternative assessment"** (6). He also explains that this new approach "(…) stresses the need for assessment to be integrated with the goals of the curriculum and to have a constructive relationship with teaching and learning" (7). He argues that learners can share the responsibility for assessment. That is why they could be trained "(…) to evaluate their own capacities in performance in a range of settings" (7). This process is known as **self-assessment**.

Proficiency tests, on the other hand, are used to discover a learner's knowledge that is already existing, learned from a known or unknown curriculum (Sharma, 2002). In recent years, McNamara has observed the increase of performance features in proficiency tests design, where the *criterion setting* or real-life language use is represented. To make it clearer, he gives an example of a communicative abilities test for health professionals in work settings which should be based on representations of such workplace tasks, for instance, communicating with patients or other professionals. He also predicts the growth and further development of courses of study preparing candidates for that kind of proficiency tests which will have a *gate-keeping function* in case of the admission to overseas universities or jobs that need practical language skills.

2. Qualities of a *good* language test

Without any doubts, one of the most important considerations in a language test designing and

developing is its usefulness. For Bachman and Palmer (1996) "(...) test usefulness can be described as a function of several different qualities, all of which contribute in unique but interrelated ways to the overall usefulness of a given test" (18). They present their concept of usefulness as a formula and suggest that all these qualities are a kind of metric by which a test and its development process can be evaluated:

> **Usefulness** = Reliability + Construct Validity + Authenticity +**Interactiveness + Impact + Practicality**

However, they claim that "(...) in order to be useful, any given language test must be developed with a specific purpose, a particular group of test takers and a specific language use domain in mind' (Bachman and Palmer: 1996, 18). This view is repeated by other researchers such as Douglas (2000), or Purpura (2005).

Bachman and Palmer's theory is echoed by Douglas who also claims that "(...) the good qualities of testing practice include validity, reliability, situational and interactional authenticity, impact and practicality" (2000: 112). Although Douglas's discussion on the subject is drawn on Bachman and Palmer's concept of test usefulness, his ideas differ a bit. While Bachman and Palmer divide authenticity and interactiveness into two qualities of usefulness, Douglas suggests authenticity to be a single quality with two different aspects – situational and interactional. He states that all the qualities mentioned above

> (...) are ones that are common to a well-designed and well-executed tests, not just LSP tests, and they amount to a set of principles for ensuring that the tests we produce are as good as we can make them in terms of

(1) the interpretations we make of the test performance (validity),
(2) the consistency and accuracy of the measurements (reliability),
(3) the relationship between target situation and the tasks (situational authenticity),
(4) the engagement of the test taker's communicative language ability (interactional authenticity),
(5) the influence the test has on learners, teachers and educational systems (impact), and
(6) the constrains imposed by such factors as money, time, personnel and educational policies (practicality) (2000: 112).

In his later work, Bachman again repeats the concept of test usefulness and emphasises that language tests are evaluated "(...) in terms of several qualities, such as reliability, validity and practicality, **with these qualities considered to be more or less independent of, or even conflicting with each other"** (2001, 110).

Saville (2003) in Hawkey (2006) puts his view in different words. Saville claims that the most important features of a good test or exam (also mentioned by Bachman and Palmer) include:

- appropriacy to the purposes for which it is used
- the ability "(...) to produce very similar results in repeated uses" (Jones: 2001)
- positive influence " (...) on general educational process on the individuals who are affected by the test results' (2003, 73)
- practicality in terms of development, production and administration (2003, 65-78; 2006, 18).

Since, again, the four features are labelled *validity*, *reliability*, *impact* and *practicality*, Cambridge ESOL uses the short term VRIP to refer to them all. Contrary to Bachman, Hawkey agrees with Saville that the four components to VRIP "(...) **are by no means independent"** (2006: 19). He quotes Saville stating that "(...) **individual examination qualities cannot be**

evaluated independently. Rather the relative importance of the qualities must be determined in order to maximise the overall usefulness of the examination" (2003, 61; 2006, 19).

In conclusion, let me quote Weigle who maintains that

> (...) it is important to consider the six qualities of usefulness proposed by Bachman and Palmer – construct validity, reliability, interactiveness, authenticity, impact and practicality. **While it may not be possible to maximise each test quality, test developers should strive to maximise overall usefulness of a test by giving careful consideration to the qualities of usefulness and determining for each testing situation an appropriate balance among them** (2002, 56).

Bibliography

Bachman, L.F. *Fundamental Consideration in Language Testing*. Oxford University Press, 1990.

Bachman, L.F and A.S., Palmer. *Language Testing in Practice*. Oxford University Press, 1996.

Bachman, L.F. *Statistical Analyses for Language Assessment*. Cambridge University Press, 2004.

Byram, M. *Routledge Encyclopedia of Language Teaching and Learning*. Routledge, 2004.

Canale, M. and Swain, M. *Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing.* Ontario: Ministry of Education, 1980.

Chun, D.M. *Discourse Intonation in L2: From Theory to Practice*. John Benjamins Publishing Company, 2002.

Davis, A. *Principles of Language Testing*. Basil Blackwell, 1990.

Douglas, D. *Assessing Languages for Specific Purposes*. Cambridge University Press, 2000.

Hawkey, R. *Impact Theory and Practice: Studies of the IELTS Test and Progretto Lingue 2000.* Cambridge University Press, 2006.

Hymes, D. *On Communicative Competence*. Penguin, 1972.

Johnson, K. *An Introduction to Foreign Language Learning and Teaching*. Pearson Education, 2001, 2008.

Komorowska, H. *Sprawdzanie umiejętności w nauce języka obcego. Kontrola – Ocena – Testowanie.* Fraszka Edukacyjna, 2002a.

Krumweide, A., *Language Assessment: Testing Bilingualism?* GRIN Verlag, 2008.

Lado, R. *Language Testing*. Longman, 1961.

McNamara, F. *Language Testing*. Oxford University Press, 2000.

Oller, J. *Language Tests at School: A Pragmatic Approach*. Longman, 2000.

Purpura, J.E., *Assessing Grammar*. Cambridge University Press, 2004.

Rivera. C. *Communicative Competence Approaches to Language Proficiency Assessment: Researh and Application.* Clevedon: Multilingual Matters, 1984.

Sharma, T.C. *Modern Methods of Language Teaching*. Sarup & Sons, 2002.

Shohamy, E. *Competence and Performance in Language Testing* in: Brown, G. Malmakjaer, 1996.

Williams, J. *Performance and Competence in Second Language Acquisition.* Cambridge University Press, 1996.

Shohamy, E. *Second Language Assessment.* Wodak, R., Tucker, G.R., Corson, D. Edwards, V., Davies, B., Cummins, J., Lier, L. van, Clapham, C., Hornberger, N.H., 1999.

Encyclopedia of Language and Education: Volume 4: Second Language Education. Springer.

Skutnabb-Kangas, T., *Bilingualism or Not: The Education of Minorities*. Multilingual Matters, 1981.

Smith, V. *Thinking in a Foreign Language: An Investigation into Essay Writing and Translation by L2 Learners.* Gunter Narr Verlag, 1994.

Spolsky, B. *The Uses of Language Tests: An Ethical Envoi.* Rivera, C. *Placement Procedures in Bilingual Education: Education and Policy Issues*. Multilingual Matters, 1984.

Weigle, S.C. *Assessing Writing*. Bachman, L.F., Alderson, J.C., *The Cambridge Language Assessment Series*, Cambridge University Press, 2002.

Weir, C.J. *Communicative Language Testing*. Prentice Hall International (UK) Ltd., 1990.

Małgorzata Gos
Uczelnia Lingwistyczno-Techniczna w Świeciu
Wydział Zamiejscowy w Przasnyszu
ul.Szosa Ciechanowska6, 06-300 Przasnysz
E-mail:steelhorse@wp.pl