Shiken: JALT Testing & Evaluation SIG Newsletter. 14(1) March 2010 (p. 14 - 18)

Insights in Language Testing: An Interview with Spiros Papageorgiou

by Mark Chapman

Spiros Papageorgiou is a language assessment specialist in the English Language Institute of the University of Michigan. He is involved in designing and conducting research projects related to testing programs and in overseeing and managing all stages of test development projects at the University of Michigan. Spiros has a PhD in language testing and an M.A. in language studies from Lancaster University, and a B.A. in English language and linguistics from the University of Athens. He has worked as a teacher of English in Greece and has taught undergraduate courses in Linguistics at Lancaster University. He has participated in a number of standard setting projects in Europe, consulting examination providers on the process of relating their examinations to the Common European Framework of Reference and has presented his work in various international conferences. His PhD dissertation won in 2009 the Jacqueline Ross TOEFL® Dissertation Award and has been published by Peter Lang. This interview was conducted in person in February 2009 in Ann Arbor, Michigan, where the University of Michigan main campus is located.

Can you give us a little background information on the University of Michigan's suite of English proficiency tests? What types of tests do you make and where are they taken?

The English Language Institute at the University of Michigan has a long tradition in developing English language tests since the 1950's. Robert Lado, the second director of the Institute, authored the famous 1961 book *Language Testing* and is considered, according to the website of the International Language Testing Association (ILTA), the "founder of modern language testing research and development". Nowadays, the English Language Institute develops English language tests for use at the University of Michigan, for use by other institutions for their own internal assessment purposes. We also offer the following four international, high-stakes testing programs:

- The **Michigan English Language Assessment Battery (MELAB)**, a multilevel examination principally for admissions screening to colleges and universities in which English is the primary language of instruction.
- The **Examination for the Certificate of Competency in English (ECCE)**, an upper-intermediate-level test that provides an assessment of examinees' language proficiency in all four language skills.
- The **Examination for the Certificate of Proficiency in English (ECPE),** an advanced-level test that provides an assessment of examinees' listening, reading, writing, and speaking proficiency.
- The **Michigan English Test (MET)**, another multilevel examination assessing examinees' proficiency in listening, reading, and language usage (grammar and vocabulary).

These four tests are taken in over 25 countries and are used by tens of thousands of examinees every year for professional and educational purposes.

And which test are you responsible for?

After joining the English Language Institute in November 2007 as a Language Assessment Specialist, my main responsibility has been the MET. In fact, the decision to design and launch the MET was taken just before my arrival, thus I was involved in the very initial design and validation stages of this program.

Thanks to the expertise of our in-house test development team and other testing researchers in the Institute, we managed to launch the program in January 2009 - a great accomplishment, given the very ambitious launch schedule. The MET has been very successful in Latin American countries and test taker numbers increase every time we administer a new form. We believe that this success is because of the following features:

- frequent monthly administrations
- excellent psychometric characteristics, as documented in our annual report that is available at http://www.lsa.umich.edu/eli/testing/met/admin
- short turn-around time for score reports, within four weeks from the test administration date
- testing of multiple proficiency levels, from A2 to C1 on the Common European Framework of Reference (CEFR)
- use of audio recordings and reading passages reflecting authentic American English in social, educational and workplace contexts
- transparent reporting of test results in relation to CEFR levels, based on extensive research conducted by the Institute
- very competitive pricing set locally for each country

You won the Jacqueline Ross TOEFL dissertation award in 2009 for your PhD thesis, which was supervised by Charles Alderson at Lancaster. Can you go over the main points of that paper?

My interest in pursuing a PhD study was triggered by the use of the Common European Framework of Reference which, since its publication in 2001 by the Council of Europe, has been the most frequently-cited performance standard in language testing inside and outside Europe. As test providers were interested in relating test content and scores to the CEFR levels, the Council of Europe published the *Manual for Relating Language Examinations to the CEFR* (Council of Europe, 2003, 2009) and as a result, there was a growing interest in setting cut scores, an area of educational measurement usually referred to as 'standard setting', which has been researched outside language testing systematically since the 1970's.

Similarly to standard setting, the process described in the Manual to relate test content and scores to the CEFR levels is primarily based on judgments by trained participants; nevertheless judgment-making in this context remained largely unexplored, even though it might affect score users of tests that report results in relation to CEFR. My PhD dissertation, which is now in revised book form (Papageorgiou, 2009), addressed this issue by employing quantitative and qualitative methods with a group of 12 trained judges involved in a CEFR standard setting project. Despite the judges' good understanding of how language ability progresses from lower to higher CEFR levels, it was found that describing test content and examinee performance in relation to the CEFR levels was not without problems and decision-making was affected by a number of factors that were irrelevant to the judgment task. Along with providing a better understanding of judgment-making during the CEFR standard setting, the dissertation had important implications for examination providers and users of CEFR-aligned test scores.

The Common European Framework of Reference (CEFR) is obviously very central to your work, but how important is the CEFR to the University of Michigan language tests?

Although the CEFR has been criticized for its technical language and its overwhelmingly detailed context, it is intended as a general reference book (as denoted by its name) to help researchers not only in language

testing but in a variety of second language areas, such as teaching, learning and curriculum design. The CEFR should not be misunderstood, as one of its authors has pointed out, as a cookbook for the development of language tests (North, 2004). In the English Language Institute we have used the CEFR as one of our main references to define the targeted proficiency levels for our tests, explore appropriate tasks for the intended proficiency levels and, overall, define the construct that our tests tap and help test users interpret test scores by conducting standard setting studies. There is no doubt that it has been a helpful tool in our on-going test development and validation procedures.

"The CEFR should not be misunderstood, as one of its authors has pointed out, as a cookbook for the development of language tests."

Can you go into a bit more detail about how the ELI makes these judgments in practice using the CEFR?

A recent example is the publication on the MET website of a CEFR equivalence table to help test takers interpret the scores they receive. There has been concern in the field of language testing that examination providers do not systematically explore and establish empirically how test scores relate to the CEFR. In the case of the MET we run a three-day standard setting meeting with a group of 13 judges in Latin America, training them in the use of the CEFR and then asking them to recommend cut scores for the MET scores. We then performed extensive analysis in Michigan to establish consistency of the recommended cut scores and related the judges' recommendations to data from the administration of a pilot test. With the use of Item Response Theory, we ensured comparability of the MET scores across different administrations and stability in the way the scores relate to the CEFR levels.

Moreover, the CEFR has been used not only to report scores, but also in relation to the content of our examinations, for example when designing the tasks that test takers respond to. At the B2 level, a learner is expected to be able to present a viewpoint and give the advantages and disadvantages of various options (CEFR, page 24). As can be seen in the online practice material of the ECCE, which tests language at this level, test takers write a letter to the editor, presenting their view on a new policy in a school and arguing about the advantages or disadvantages of this policy.

How widely is the CEFR understood in contexts outside of Europe?

Although some test users perceive it as a European-only document, the extended CEFR-related work by examination providers in Asia and North America has shown that this is not true. . . North America has a long history in standard setting and performance standards, primarily since the publication of the article *Scales, norms and equivalent scores* by Angoff (1971). His article introduced a standard setting method, the Angoff method, which, along with its variations, is the most frequently used and well-researched method for setting cut scores.

The book *Standards for Educational and Psychological Testing* published by three professional US organizations has been very influential since its first edition in 1985(AERA, APA, NCME, 1999) as is the case with the ACTEFL Guidelines (ACTEFL, 1986) ; the No Child Left Behind Act (US Department of Education, 2001) has resulted in growing interest in standardizing K-12 assessments in the US; and finally, the Canadian Language Benchmarks (CCLB, 2000) have been extensively used in Canada. With such a long history in standards-based assessments, it is not surprising that North American providers of international examinations, such as the English Language Institute, make extensive use of the CEFR. We report scores in relation to the CEFR because so many of our test takers and test score users are in parts of the world where the CEFR is very widely used, thus they can better understand what their scores mean.

What other standard setting projects have you been involved with?

I've been involved in standard setting projects for examinations that varied significantly in terms of purpose and use. For example, while working on my PhD at Lancaster, I participated in the standard setting panel for the TOEFL. I have consulted teams in two European countries developing matriculation examinations and a third team developing a test of academic English for local use in a university. I have run a standard setting project for an international examinations provider and finally for a provider of a non-English language examination. Apart from having the opportunity to work with colleagues around the world and learn from them, I am very happy to see that nowadays standard setting is an important item in the agenda of examination providers; and this is good news for test takers and other score users, as it has a direct influence on test scores and the decisions that are made on the basis of these scores.

Can you talk a little bit about the difficulties involved in setting such standards?

". . . no matter how laborious standard setting might be, it is as important as all other stages of test development and validation."

Standard setting, contrary to what some might believe, is not a difficult task, but it requires a very systematic approach and can be time-consuming and demanding in terms of resources, both human and technological, as well as logistics. A first important issue is to identify who the expert judges are, and who is qualified to recommend cut scores, as human judgments are the basis of standard setting. Apart from the logistics, there are other theoretical issues, such as the method or methods to be used when recommending cut scores and the quantitative and qualitative analyses to ensure that judgments are reliable and that the recommended cut scores are valid and make sense in the context they will be used. However, I would like to stress that no matter how laborious standard setting might be, it is as important as all other stages of test development and validation. This is because, as I mentioned earlier, standard setting has a direct influence on test scores. Thus, setting a cut score arbitrarily, for example by having a couple of people deciding whether a score of 60 or 70 is appropriate, without proper standard setting procedures, is in my opinion unacceptable and does not abide by the professional guidelines of organizations such as the International Language Testing Association (ILTA, 2007) and the European Association for Language Testing and Assessment (EALTA, 2006).

Works Cited

ACTFL-American Council on the Teaching of Foreign Languages (1986). *ACTFL proficiency guidelines.* Hastings-on Hudson, NY: ACTFL.

AERA, APA, NCME, (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education) (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington: American Council on Education.

CCLB-Centre for Canadian Language Benchmarks (2000). Canadian Language Benchmarks 2000. Available from http://www.language.ca/pdfs/clb_adults.pdf

Council of Europe (2003). *Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Preliminary pilot version.* Strasbourg: Author.

Council of Europe (2009). Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. A Manual. Retrieved 15/02/2009, from http://www.iltaonline.com/index.php?option=com_content&view=article&id=122&Itemid=133

EALTA (2006). EALTA Guidelines for Good Practice in language testing and assessment Retrieved 05/09/2008, from http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

ILTA (2007). Guidelines for Practice. Retrieved 05/02/2010, from http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

North, B. (2004). Europe's framework promotes language discussion, not directives. Retrieved 31/01/2007, from http://education.guardian.co.uk/tefl/story/0,,1191130,00.html

Papageorgiou, S. (2009). *Setting performance standards in Europe: The judges' contribution to relating language examinations to the Common European Framework of Reference*. Frankfurt: Peter Lang.

US Department of Education (2001). No Child Left Behind Act. Public Law 107-110. Retrieved 15/02/2010, from http://ed.gov/policy/elsec/leg/esea02/107-110.pdf

- Some University of Michigan Resources -

Michigan English Test CEFR Equivalence Table:
http://www.lsa.umich.edu/eli/testing/met/certificates

Examination for the Certificate of Competency in English Practice Material:
http://www.lsa.umich.edu/eli/testing/ecce/examinees