Institution: University of Leicester

Unit of Assessment: UoA25 Education

Title of case study: Improving the Fairness and Accuracy of Language Testing and Assessment

1. Summary of the impact

Border agencies, employers and universities use language tests to make decisions about immigration requests, job applications and university admissions. The two largest tests in the world have 4.5 million test takers per year. Good test design is crucial in determining the fairness, relevance and accuracy of the results. Our research has enabled us to create new tools that have been used to enhance quality control and develop assessment skills. We have created new scoring methods to make performance assessment more reliable, and developed theoretical frameworks to improve test development. Our research has impacted upon professional practice and training within examination boards.

2. Underpinning research

Language assessment research at Leicester is carried out by Professor Glenn Fulcher, Dr Pam Rogerson-Revell, Dr Julie Norton, Dr Agneta Svalberg, and their research students, under the umbrella of the Testing, Assessment and Measurement Research Group. Our research falls into five broad categories:

- defining what should be tested for particular decision contexts;
- the analysis of test performance and its impact on test design;
- the design of scoring procedures;
- social and policy aspects of test use;
- assessment literacy for teachers and test developers.

Our research strengthens the usefulness of test scores taken by millions of candidates around the world for immigration, employment or for further education through the development of new scoring techniques and quality assurance processes. The research has a strong fairness and social policy agenda, and over the previous 7 years has proved relevant to examination boards and international training bodies such as [text removed for publication].

Our research in performance analysis informs what examination boards decide to test in relation to the decisions made by score users. This requires careful analysis of the skills, knowledge or abilities required to undertake real-world tasks (**4**). We have been particularly successful in defining second language speaking competence and language awareness (**6**), as Leicester researchers have been asked to serve as consultants to examination boards to inform future test development. Closely related is performance analysis, which describes the tasks undertaken in the real-world, and seeks to replicate these in test-tasks. The two aspects of this research are describing the task itself according to task features, and describing the discourse used to accomplish these tasks (**5**).

Scoring Procedures are the rules that guide how the test user assigns numbers (test scores) to performances. Our research has led to the creation of new data-driven rating procedures that make it possible for raters (or judges) to make decisions about the quality of test-taker performance during live performance, or from recordings. These are easier to use than previous versions, and offer sounder inference from the score back to the skills that contribute to successful performance (**4**).

Social and Policy Aspects are important because the outcomes of testing impact on the lives of individuals and the institutions that use the scores for high-stakes decision making. Research at Leicester focuses on the use of existing tests to make decisions for purposes they were not originally designed for. This primarily concerns using tests of English for academic purposes to make decisions about employment (e.g. health care), and international mobility (the language

component of immigration policy). We have developed test retrofit theory to inform practice, and developed a retrofit toolkit and test revision procedures to guide ethical test reuse by examination boards (**3**).

Assessment Literacy is important for teachers who are asked to produce tests or engage in testing. Our research, funded by the Leverhulme Trust, has produced an assessment literacy model and sets of text- and web-based tools that have been widely used in training (**1,2**).

3. References to the research

1. Fulcher, G. (2010). *Practical Language Testing*. London: Hodder Education.

2. Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly* 9(2), 113 - 132.

3. Fulcher, G. and Davidson, F. (2009). Test Architecture. Test Retrofit. *Language Testing* 26(1), 123 - 144.

4. Fulcher, G., Davidson, F., and Kemp, J. (2011). Effective rating scale development for speaking tests: Performance Decision Trees. *Language Testing* 28(1), 5 - 29.

5. Norton, J.E. (2013), Performing Identities in Speaking Tests: Co-construction Revisited, *Language Assessment Quarterly*, 10(2), 309 - 330.

6. Svalberg, A. (2009). Engagement with Language: Developing a Construct. *Language Awareness*, 18(3-4), 242-258.

4. Details of the impact

Research at Leicester has made an important contribution to assessment policy and practice in the UK and beyond. Researchers have worked with the largest examination boards and teacher trainers, such as [text removed for publication], to develop new assessment products, create new assessment tools, use new quality assurance procedures, and improve assessment training.

In the field of skills and performance analysis, Fulcher has acted as co-chair of the [text removed for publication] project since 2010, to design the next generation of language tests for [text removed for publication]. Research into speaking assessment (**A, B, D, E**) impacts on the current design process. The research will effect university admissions testing for the next 20 years. Fulcher has also been an advisor to [text removed for publication] (2011 – 2013) on their new university admissions English tests.

Leicester research on scoring procedures has had significant impact upon the practices of [text removed for publication]. Norton has worked with [text removed for publication] test data to investigate the impact on individual test performance of other speakers. Her research informs rating practice in speaking assessments (2000 – 2013; **E**). Fulcher's research on rating scales has directly impacted upon [text removed for publication] current scoring procedures for all its English language examinations. His research on fluency assessment also impacted upon the European Common Framework of Reference for Languages, and the Association of Language Testers in Europe scale project.

Our research also affects quality control in examination boards. Fulcher's research on retrofit theory was internationally acknowledged in 2011 by the International Language Testing Association. The 2009 publication *Test Architecture, Test Retrofit* received a "Best Paper Award" for its impact on the practice of examination boards. The citation (**A**) reads:

> "The award committee found the paper by Fulcher and Davidson an excellent conceptual paper that emphasizes the centrality of test purpose in test design decisions, and proposes

a systematic approach to evaluating test revisions and test retrofit. The paper is very well-written and the authors guide the reader step by step through the processes of careful decision-making that language testers should undergo when changing tests or test purposes. The use of the architecture metaphor is well-chosen and makes the argument compelling and accessible to a broad audience including practitioners. Test retrofit has been heatedly debated in public forums but has never seen such a systematic treatment in the scientific literature as in this paper. The authors make a strong and timely contribution to the field of language testing in a period where tests are being used for purposes they were not originally intended for, or misused entirely, but the consequences of test change or change of test purpose for the validation process can hardly be found in the literature. Fulcher and Davidson provide the language testing field with the appropriate terminology and guidance for this important and timely topic."

In 2011 this led to an invitation for Fulcher to conduct staff development training on retrofit theory and practice in quality control processes for [text removed for publication] staff. In 2010 Fulcher was also invited to provide training to test writers in the assessment division of [text removed for publication].

Our research is also important because of its social awareness and justice components. The technical research into improved scoring and decision making is only effective if it can be used by teachers and examination boards to improve their practices. Such impact leads to fairer decision making when using test scores.

Fulcher's Leverhulme-funded research into assessment teaching and training has led to a range of web-based tools (http://languagetesting.info) that has impacted upon improved teaching of language testing around the world. For example, in 2011 the [text removed for publication] approached us to use these resources to train teachers in rural China (**E**). A programme of subtitling was undertaken (http://languagetesting.info/videos/subs.html) and embedded within their training scheme (**F**). He was distinguished visiting lecture at Temple University in Japan, delivering public lectures on testing in society, and conducted a successful webinar for teachers across the Americas through San Diego State University in April 2012 (**G**). Fulcher also edits the prestigious journal *Language Testing* (Sage) and produces a quarterly podcast to make content accessible to the general public, as well as disseminating research in the popular press. We believe that impact should extend beyond academia to engage a wider audience in important social policy issues that involve assessment practices.

5. Sources to corroborate the impact

A. Award citation from the International Language Testing Association for Fulcher, G. and Davidson, F. (2009). Test Architecture. Test Retrofit. *Language Testing* 26(1), 123 - 144.

B. Information from [text removed for publication], dated 2012, indicating article on Performance Decision Trees was one of the top downloads of the year from their website.

C. Web site statistics for languagetesting.info show increased use of assessment literacy resources on a year-by-year basis. Figures for 2012: 84,348 unique visitors, up from 65,027 in 2011.

D. [text removed for publication] to corroborate the claim made with regard to Fulcher's role in the development of new speaking tests.

E. Email from [text removed for publication], indicating the role and use of Leicester video training materials for teachers in China.

F. Subtitled Language Testing Video downloads via [text removed for publication]: 172 unique users during 2012, and 1,228 video viewings in total.

G. Feedback from participants in the Americas Webinar, 2012. Podcast available for use at: http://larc.sdsu.edu/podcasts/category/testing-and-assessment-webinar/. On likert items, this dissemination scored consistently over 4.3 average on all quality measures (spreadsheet available for inspection).

H. Read, J. (2011). Book Review: G. Fulcher (2010). Practical language testing. London: Hodder Education. *Language Testing* 28(2), 302 – 304. Available for free download at http://ltj.sagepub.com/content/28/2/302.full.pdf+html.