Institution: Lancaster University		
Unit of Assessment: 26, Linguistics		
Title of case study: Transforming the testing of listening and reading ability in a second or foreign language		
Period when the underpinning research was undertaken: 2009 to 2020		
Details of staff conducting the underpinning research from the submitting unit:		
Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Tineke Brunfaut	Professor	January 2009 to present
Luke Harding	Professor	December 2010 to present
Period when the claimed impact occurred: August 2013 to December 2020		
Is this case study continued from a case study submitted in 2014? N		

1. Summary of the impact

Poorly designed listening and reading tests can lead to inaccurate and biased measurement of ability, threatening the fairness of educational and professional opportunities for learners of second or foreign languages. Research led by Brunfaut and Harding into the nature, methods and equity of testing listening and reading has positively transformed educational testing systems around the world and ensured fairer assessments of these skills. Specifically, the research resulted in i) the creation, revision, and much improved evaluation of listening and reading tests of major international and government test providers, constituting fairer testing of over 7.5million language learners worldwide, and ii) enhanced knowledge and capability in approximately 1,215 language educators and testing professionals across 10 countries through evidence-based training to create fairer tests of listening and reading ability.

2. Underpinning research

It is common to test listening and reading in a second or foreign language for various reasons, for example, to assess achievement and progress in language education; to determine readiness to undertake academic study in a second or foreign language; and, for certification of language proficiency for professional purposes. A major challenge in test design has been limited understanding of what is entailed in listening and reading in a second or foreign language, and thus what should be tested. In addition, some tests have been constructed based on outdated views of the nature of listening and reading comprehension or have been designed without sufficient insight into the factors which determine effective and fair measurement of reading and listening skills. The use of unfair test scores in educational or professional decision-making can affect the life chances of second or foreign language learners. The research programme conducted by Brunfaut and Harding, comprising various projects and collaborations with Alderson, Kormos, Kremmel, McCray, Michel and Révész (see Section 3), addressed the above issues in three main strands:

A. *What is listening and reading ability?* Conducted from 2009 onwards, this research into the nature of listening and reading in a second or foreign language identified linguistic variables (e.g., multiword expressions, phonological variation, contractions, overlap between utterances) and learner variables (e.g., first-language background, skill-related anxiety, working memory) that impact on listening and reading performance in complex ways [R2, R3, R5, G1, G2, G5]. The findings also revealed specific cognitive processes which drive listening and reading comprehension in a second or foreign language (e.g., the differential role of lower-and higher-level processes depending on performance level) [R6, G4].

B. *How to measure these skills effectively and fairly?* This body of research, undertaken since 2009, determined the impact of different task characteristics (e.g., text genre, task-related motivation) and modes of delivery (e.g. paper v. computer) on the measurement of listening and reading proficiency [R2, R3, G1, G2, G5, G6]. Further, Brunfaut and Harding, together with Alderson, derived a set of 5 diagnostic principles from interviews conducted in 2012 with diagnosticians across a range of professions. For example, they found a common diagnostic procedure involving 4 stages: listening/observation, initial assessment, hypothesis checking, and decision-making. From these findings, in 2015 they published a ground-breaking theoretical framework with concrete implications and procedures for diagnostic assessment of second/foreign language listening and reading, including guidance for building diagnostic tools based on a processing view of comprehension [R1, R4].

C. *How to evaluate the quality of listening and reading tests?* These studies pioneered the development of cutting-edge test validation methodologies, including establishing key metrics to investigate linguistic variables and cognitive and non-cognitive learner characteristics in the context of language testing [R2, G1, G2, G5, G6] (since 2009), the novel development and use of eye-tracking metrics specifically adapted for language testing [R6, G4] (2011-2018), and a new approach to standard setting which involved a twin panel design, and a modified approach to the “basket method” [G3] (2012-2014).

The research programme was funded through the government, non-profit and commercial bodies listed in Section 3. In addition, Brunfaut and Harding oversaw a research group of ten people (Language Testing Research Group) who aided in piloting new research methodologies and materials between 2011-2018, e.g., methods to identify information necessary to comprehend texts, to measure interactive listening skills, and to set standards for listening exams. Furthermore, the Linguistics Eye-Tracking Lab, coordinated by Brunfaut, assisted in developing innovative approaches to researching listening and reading, their measurement and validation, and providing high-spec hardware and software.

3. References to the research

[R1-R6] all peer-reviewed

[R1] Alderson, J.C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics*, 36(2), 236-260. https://doi.org/10.1093/applin/amt046.

[R2] Brunfaut, T., & Révész, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, 49(1), 141-168. https://doi.org/10.1002/tesq.168

[R3] Harding, L. (2012). Accent, listening assessment and the potential for a shared-L1 advantage: a DIF perspective. *Language Testing*, 29(2), 163-180. https://doi.org/10.1177/0265532211421161

[R4] Harding, L., Alderson, J.C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing*, 32(3), 317-336. https://doi.org/10.1177/0265532214564505

[R5] Kremmel, B., Brunfaut, T., & Alderson, J. C. (2017). Exploring the role of phraseological knowledge in foreign language reading. *Applied Linguistics*, (38)6, 848-870. https://doi.org/10.1093/applin/amv070

[R6] McCray, G., & Brunfaut, T. (2018). Investigating the construct measured by banked gap-fill items: evidence from eye-tracking. *Language Testing*, 35(1), 51-73. https://doi.org/10.1177/0265532216677105

International prizes awarded for the quality and pioneering nature of the research:

- [R1] won Best Article Award 2015, by the International Language Testing Association
- [R5] won Caroline Clapham IELTS Masters Awards, by IDP:Australia, the British Council & Cambridge English, and was top 5-shortlisted for the ILTA Best Article Award 2017
- TOEFL Outstanding Young Scholar Award 2015, by Educational Testing Service (Brunfaut, complete body of work)
- Best Research Award 2018, by the e-Assessment Association (Brunfaut & Harding)

Grants in support of the research (*peer-reviewed, open calls; **commissioned studies):

[G1] **Brunfaut, T. & Révész, A.: *‘Tasks and assessing second language listening comprehension’,* Trinity College London (2009-2010) GBP15,000.

[G2] *Brunfaut, T. & Révész, A.: *‘Tasks, proficiency and assessing L2 listening comprehension’,* Pearson (2010-2011) GBP10,000.

[G3] *Brunfaut, T. & Harding, L.: *‘Linking the GEPT listening test to the Common European Framework of Reference’,* Language Training and Testing Centre (Taiwan) (2012-2014) GBP19,914.

[G4] *Brunfaut, T. & McCray, G.: *‘Looking into Reading I’* and ** Brunfaut, T.: *‘Looking into Reading II’,* British Council (2013-2016) GBP19,807.

[G5] *Kormos, J., Brunfaut, T. & Michel, M.: *‘The effect of working memory & task motivation on performance in the TOEFL Junior Comprehensive test’,* Educational Testing Service (ETS; USA) (2017-2018) USD44,802.

[G6] **Brunfaut, T. & Harding, L.: '*An investigation of the impact of delivery mode on the Integrated Skills in English exam*', Trinity College London (2017-2018) GBP60,818

4. Details of the impact

Brunfaut and Harding's research-based expertise has prompted Ministries of Education, commercial and non-profit language learning and testing organisations, and professional bodies to adopt the research findings in order to improve existing testing practices. The listening and reading tests concerned are used around the globe and have been taken by over 7.5million language learners.

Creating new, fairer tests for secondary schools in Austria and Luxembourg

The research programme under strand A and B led to invitations from the Austrian and Luxembourg Ministries of Education for Brunfaut and Harding to consult on the creation of new tests designed to be fairer and more reflective of best-practice approaches. Working with teams of language teachers in Austria and Luxembourg, 3 new tests were designed and implemented.

- *E8 Standards Test* (80,000 learners, Austria; impact since 2019)

Based on the research on diagnostic listening and reading assessment [R1, R4], the Austrian Ministry of Education commissioned Brunfaut and Harding to design a new national English test for Year 8 with diagnostic reporting. As attested by the IQS Division Head at the Ministry [S1], the new test was a marked improvement on previous assessments, providing better information for teachers to use in assessing learning and for policy makers to use in educational system monitoring. The work on this test led to ongoing collaboration, including current work on a new, online diagnostic reading and listening test for English language learners in Years 7 and 8.

- *Épreuve Commune test – Anglais* (14,221 learners, Luxembourg; impact between 2013 and 2018)

The research findings on the linguistic, learner, and task variables that affect listening and reading test performance [R2, R3, R5, G1, G2] were adopted in the development of a new Year 8 listening and reading test for English, supported by the Luxembourg Ministry of Education. The relevant Director at the Ministry explains [S2] that "*the test results provided standardized information against an international benchmark on English listening and reading competence development ... [enabling the Ministry] to make evidence-based policy decisions regarding English education.*" The test also impacted on classroom teachers, giving them "*concrete and dependable information on learner progress and on achievement*."

- *Examen de fin d'études* - *Anglais* (768 learners, Luxembourg; impact since 2020)

Brunfaut and Harding's work with the Ministry of Education in Luxembourg led to the development of a new, high-stakes end-of-secondary school leaving exam, the oral interaction component of which was administered for the first time in June 2020 (with the reading test planned for May 2021). According to the Ministerial Director [S2]: "F*eedback from students and teachers on the first live edition of the oral interaction part of the exam last school year (768 school leavers), confirmed the effectiveness and suitability of the new exam in evaluating learners' English proficiency and in certifying outcomes achievement in a comprehensive and fair manner for individual learners*." The test also provided better academic and professional opportunities for school-leavers based on the certification.

Improving the efficacy of English language tests around the globe

The uptake of Brunfaut and Harding's research (strand A and B) by international English language test providers has resulted in revisions to the design of these established international English language tests:

- *Aptis listening test & Workplace Literacy and Numeracy test* (test-taker numbers confidential; impact since 2016)

Years of collaboration with the British Council, and uptake of the findings on the nature of listening comprehension and its fair assessment [R3, R4] have culminated in the British Council changing their Aptis listening test in May 2020. This included modifying test design specifications to provide more appropriate listening input to better assess learners at high proficiency levels. The specific impact of Harding's research on the effect of speaker accent on listening assessment was also extended to local test development projects conducted by the British Council's Assessment Research Group, including the recruitment of speakers with local, Singaporean accents in the Workplace Literacy and Numeracy (WPLN) test, developed in collaboration with SkillsFuture Singapore. These innovations increased test quality and authenticity, leading to enhanced positive impact for test-takers as confirmed by the Manager of the Assessment Research Group (ARG) at the British Council [S3].

- *ISE – Integrated Skills of English exam* (test-taker numbers confidential; impact since 2017)

The user-interface of the ISE has been redesigned to accommodate Brunfaut and Harding's findings on the effect of task characteristics on the measurement of listening and reading proficiency [G6]. The ISE is an international English test used for high-stakes academic and migration purposes, developed by Trinity College London - an international exam board and education charity. The research was described by Trinity's Director of Language [S4] as "*a seminal piece of research for Trinity, and for the ISE test suite more specifically*" which provided "*empirical foundation from which to develop [Trinity's] response to the crisis of education caused by the COVID-19 pandemic*". Brunfaut and Harding continue to provide strategic and evidence-based advice on Trinity's assessment materials and procedures through the established Trinity-Lancaster Partnership.

- *TOEFL - Test of English as a Foreign Language* (approximately 4.4million test-takers based on in excess of 35million test-takers since 1964; impact since 2014)

The research on the nature of listening comprehension [R3, R4] has informed revisions to the listening section of the TOEFL test, one of the two most widely used English proficiency tests in the world. Harding's research-based expertise also informed the development of a roadmap envisaging how TOEFL listening assessments will change in the future, resulting in a position paper co-authored between Harding and R&D staff at the Educational Testing Service (the US-based non-profit organisation which develops and administers the TOEFL) [S5]. There has been sustained engagement with the TOEFL programme through Harding's appointment to the TOEFL Committee of Examiners (between 2015 and 2019), with an ETS Senior Research Scientist stating: "*Professor Harding's research has had an important impact on the TOEFL program, contributing to changes in listening assessment policy and practice and leading to fairer measurement*" [S5].

Improving the evaluation of test quality with new methods

The research under strand C has provided innovative insights that have informed and improved quality control practices across large-scale language testing provision:

- *Aptis reading test* (approximately 700,000 test-takers; impact since 2015)

Brunfaut's innovations in eye-tracking metrics, developed in collaboration with McCray [R6, G4], led to one of the first pieces of validation work on the Aptis reading test, introduced by the British Council in 2012. This work confirmed that the test elicited reading-relevant cognitive processes, supporting claims made about test-takers' reading scores, and providing a sounder basis for decision-making in educational and professional contexts. It also revealed problems with one particular Aptis reading task, leading to the development of a novel task type, the quality of which, in turn, was confirmed using the innovative eye-tracking techniques. This new task type has subsequently been introduced, improving live testing. The British Council's ARG Manager [S3] affirmed that they "*can now be confident that the Aptis reading test scores provide a valid basis for educational and professional decisions about test-takers' reading proficiency in English*", with Lancaster University's research also leading to "*important improvements to the test development cycle adopted in British Council English language assessment work*."

- *GEPT – General English Proficiency Test* (approximately 2.4million test-takers; impact since 2014)

Brunfaut and Harding's newly developed approach to standard setting [G3] was adopted to set standards for the GEPT listening test – an international English language test developed by the Taiwanese Language Training and Testing Center (LTTC) and used for high-stakes university admission, placement, and graduation purposes. The method led to new, fairer performance standards, an outcome that affects the opportunities afforded to millions of test-takers for further study and employment around the globe. Specifically, linking the GEPT to the Common European Framework of Reference [CEFR] has meant that tertiary-level students now have access to a local English proficiency test, so that "*GEPT scores are used by prestigious universities in Hong Kong, Japan, France, Germany, the UK, and the USA*" (R&D Program Director, LTTC) [S6].

- *GCSE French, German, Spanish* (268,723 test-takers; impact since 2019)

In 2019, Ofqual applied Brunfaut and Harding's standard setting approach [G3] to investigate GCSE Modern Foreign Language [MFL] qualifications to "*inform its policy decision of whether to intervene and adjust grading standards in MFL GCSE qualifications*" (Ofqual/19/6559/1) [S7].

- *TOEFL Junior Comprehensive test* (test-taker numbers confidential; impact since 2018)

The research on suitable and effective metrics of cognitive and non-cognitive learner characteristics in the context of language testing [R2, G2] and modes of test delivery [G6] was used in an evaluation of the TOEFL Junior Comprehensive test. As asserted by an ETS Research Scientist [S8]: "*This key research allowed us to shift the standard version of TOEFL Junior® reading and listening tests to a new online mode in Spring 2020, while still guaranteeing the fairness of our assessment procedures*." Furthermore, "*It has informed the accountability of our work at ETS, which is crucial to securing the organization's future position and operations in educational testing, and its work in advancing fair educational assessment for learners across the world*."

Enhancing knowledge and skills amongst language testing professionals

Language testing literacy has improved vastly amongst more than 1,215 professionals (language teachers, teacher trainers, policy makers, test item writers, test developers, administrators and government employees) due to extensive training carried out by Brunfaut and Harding. Between 2013 and 2020, they provided 157 research-informed training workshops to professionals across Austria, China, Indonesia, Luxembourg, Qatar, Slovenia, Sri Lanka, Switzerland, Taiwan, and the UK. The training resulted in increased language assessment literacy amongst these key stakeholders by improving their own testing practices, leading to better-quality listening and reading tests (e.g. Austrian and Luxembourg Y8 exams; Qatar university-level tests), improved diagnostic test reporting for the Government, schools, teachers, parents and learners in respect of Austrian Y8 exams, and benchmarked score interpretations (e.g. Austrian and Luxembourg Y8 exams) [S1, S2, S9]. A teacher in Slovenia attests to making the following changes to their professional practice: "*When I'm teaching a class, I am generally more attentive to what I want my pupils to learn, which I later include in the exam. My exams have changed thoroughly… If I used to test 70 different examples of grammar, I now include all four skills in each exam*" [S9]. In another example, a workshop participant responsible for the development of aviation English tests at DWR Aviation Engineering GmbH wrote: "*The course had a huge impact on our methods to assess listening … we practically revised all our test prompts. To conclude: yes, there was a huge impact*" [S9].

The extensive nature of the impact described here derives from numerous research-informed interactions with partners which have affected millions of people around the globe, and which are both replicable and sustainable.

5. Sources to corroborate the impact

[S1] Testimonial from Division Head of the Austrian Ministry of Education, IQS corroborating impact on E8 testing (2020).
[S2] Testimonial from the Director of Luxembourg Ministry of Education, SCRIPT corroborating impact on educational reform of English as a foreign language testing (2020).
[S3] Testimonial from the Manager of the Assessment Research Group, British Council corroborating impact on Aptis tests (2020).
[S4] Testimonial from Director of Language, Trinity College London corroborating impact on the Integrated Skills in English exam and operational benefits (2020).
[S5] Testimonial from Managing Senior Research Scientist, ETS (USA) corroborating impact on the TOEFL program (2020).
[S6] Testimonial from Program Director R&D, Language Training & Testing Centre (Taiwan) on impact relating to the General English Proficiency Test (2020).
[S7] Ofqual/19/6559/1 report 'Investigating standards in GCSE French, German and Spanish through the lens of the CEFR' by Curcin M. and Black B (2019).
[S8] Testimonial from Research Scientist, ETS (USA) corroborating impact on TOEFL Junior tests (2020).
[S9] Post-training feedback and appraisal messages received via personal e-mail (2017 to 2018).