Investigating Test Purpose Pluralism and Test Retrofitting in High-Stakes Language Proficiency Testing

Melissa McLeod, Queen's University, Canada

Abstract: There is a lack of clarity in current language testing practices regarding using tests for multiple purposes and test retrofitting. Many existing tests are used in cases that extend far beyond their original intent. This literature review examines how, despite the availability of three similar procedures, there is little publicly available research describing test retrofits. It provides an overview of the three procedures for retrofitting (Fulcher & Davidson, 2009), retrofitting for diagnostic assessment (Lee & Sawaki, 2009), and repurposing (Wendler & Powers, 2009). Then, it gives a brief discussion of the available research, including the use of academic tests for high-stakes immigration decisions, and concludes with suggestions for future research.

Keywords: Test Retrofitting, Test Repurposing, Test Purpose, Test Design, Test Development

Introduction

The number of labor migrants is growing quickly. Approximately 3.5% of the global population is moving, already surpassing the 2.6% projected for the year 2050 (McAuliffe & Khadri, 2019). Many countries, including Canada, now require language proficiency test scores as part of an immigrant's application. However, many of the tests used for this purpose were originally developed for other, unrelated purposes such as entry to post-secondary studies. Test score use in such a high-stakes context has to be defensible, just, and valid because the test scores contribute to the decision of allowing or denying access to a better future. "Sleight of hand" refers to some form of deception, often connected with magic tricks, and yet that is how Fulcher and Davidson (2007) describe using an existing test for a purpose other than that for which it was originally validated. However, creating and validating new tests is a long and expensive process (Lamprianou & Athanasou, 2009). The Educational Testing Service (ETS) in the United States argues customers often do not want to wait a long time for a new test to be developed and their large bank of test items can be repurposed as long as standards are followed (Wendler & Powers, 2009).

Cronbach (1984) put forth that a single test cannot contain all of the desired components; rather, the specific purpose and context drive the choice of test used. Chalhoub-Deville and Fulcher (2003) argued "different test purposes entail different test design considerations and require differentially targeted validation research […] a test that suits all purposes creates validation chaos" (p. 502). However, Newton (2017) argued that tests are already multipurpose and all the different purposes need to play a central role in the test development process. Such divergent views of test purpose purism are evident in how language proficiency tests are being used by testing companies and test score users for immigration. For example, The International English Language Testing System (IELTS) is a large-scale test designed for entry to academic study but by the mid-90s, it was being used beyond its original intent by governments for immigration applications. While large-scale tests such as IELTS have been used for different purposes than their original intent, language assessment scholars and testing companies have only recently defined theories such as test retrofitting (Fulcher & Davidson 2007) and test repurposing (ETS, 2009) to underpin multiple uses and purposes of tests. The purpose of this literature review then is to examine some of the retrofitting and repurposing theories regarding using tests for multiple purposes. Additionally, it will contrast these theories against current practices in the language testing field of using language proficiency tests designed for academic purposes in the high-stakes context of immigration.

To address the objectives of this study, literature searches of the ERIC academic search engine on the EBSCO platform for the terms "*test retrofit", "test retrofitting", "reverse engineering", "test repurposing"*, and "*assessment repurposing"* were conducted. These searches yielded only 11 results, with Fulcher authoring or co-authoring five works and the test development company ETS publishing two research reports. Searches on the IELTS website for the above terms as well as "*immigration"* and "*Canada"* yielded only 2 related research reports[2].

Based on the identified studies and research reports, this literature review begins with Newton's (2017) discussion of designing new tests for multiple purposes. Next, it gives an overview and comparison of three procedures for using existing tests for multiple purposes. Finally, it provides an overview of studies of tests that are used for high-stakes decisions without being retrofitted and then concludes with suggestions for future research.

[2] Articles from this search are marked with an asterisk in the reference list.

Test Purpose Pluralism

Newton (2017) claimed that tests can be multipurpose. Purpose pluralism "idealizes the principle that assessment design should be driven by a multiplicity of assessment purposes simultaneously" (Newton, 2017, p. 5). To support the claim in the title of his article, *There is More to Educational Measurement than Measuring*, Newton gives examples of how teaching professionals experience purpose pluralism. Classroom teachers have already experienced how the grades they give carry more than one meaning for their students and often have more than one impact on their lives. Instead of purpose purism, Newton (2017) argued having only a single perspective to inform test design was not as productive as considering multiple interactive perspectives to inform test design. This multiplicity in perspectives enables test developers to ensure tests represent the wide spectrum of input from administrators, teachers, parents, and other stakeholders and the pros and cons of each purpose are better balanced.

Newton (2017) has claimed that three perspectives are important for informing test design and answering the question of why a test is needed in the first place: the information perspective, the expertise perspective, and the engagement perspective. The information perspective involves how an assessment's results inform decisions. The expertise perspective involves the expertise and social capital that is gained from performing well on an assessment. The engagement perspective involves empowering both teachers and test-takers as they prepare for an upcoming assessment. These three perspectives relate to the personal, institutional, and/or societal values associated with the test and they are not meant to be viewed in isolation. It makes sense to consider the three perspectives simultaneously throughout the design process since they do not occur on their own and their interaction also informs any compromises between them.

Newton's arguments focus more on incorporating multiple purposes into new test design rather than taking an existing, single-purpose test and using it for other purposes as is often the case in high-stakes language proficiency testing. The next section shifts the discussion towards a description of procedures that could be used to expand the uses of existing single-purpose tests.

Procedures for Test Retrofitting & Repurposing

Upgrade & Change Retrofitting

Fulcher and Davidson (2009) defined two types of test retrofitting: *upgrade retrofits* and *change retrofits*. When a test is revised to use new technology or to meet revised standards, it is thus better suited to its intended purpose and has been upgraded. When a test is used for a new purpose or with users that differ from its original purpose, this is a change retrofit (Fulcher & Davidson, 2009). Fulcher and Davidson (2009) identified six components of test architecture that could change and need to be addressed by test retrofitting procedures (see Table 1 below). Of these components, it is the test's interpretive argument that requires retrofitting to maintain its validity argument.

Table 1. Test Architecture Components as Related to Retrofitting

Test Architecture Components	Relationship with Retrofitting
Test items	Commonly changed or revised as the test evolves
How the test is built (the number of items, text length)	Easier changes that do not typically change the level of difficulty of the tasks
Interpretive argument	Majority of test retrofits involve changes made at this level
Test appearance & delivery (ex. computer vs. paper-based)	Only ever changed out of necessity as these are very visible changes that can lead to apprehension among stakeholders
Test construct (what exactly is it that the test aims to measure) as well as things like the test brand, name, & scores	Typically difficult to change
Models of test design (theoretical underpinnings of language learning & use)	Typically little to no change

Note. Adapted from Fulcher and Davidson (2009).

The authors outline an eleven-step procedure to conduct an upgrade or change retrofit (see Table 2 below). The first two steps involve assembling a team of experts and gathering the appropriate documentation to inform their

decisions. In the third step, the expert panel determines if the retrofit is required and which type of retrofit is needed. If a change retrofit is needed, it is at this point that the panel decides if it would be more appropriate to develop a new test. Next, they examine other tests available for the same purpose and consider how a retrofitted test may address the new needs and be accepted by the users. Then, after fully consulting test score users, policymakers, test-takers, and other stakeholders, the decision is made to retrofit or not. Step 7 is where the actual retrofit process begins with a completed and detailed outline of the entire project, including timelines, supporting research requirements, and resources. The plan is checked against current standards documents and/or available guidelines. The experts prioritize the evidence required to make new inferences from the retrofitted test, thus determining what research they need for their validation argument. Finally, the plans for the retrofit are made public and their entire process is documented. Unfortunately, Fulcher and Davidson (2009) were only able to reference a few available examples of documented upgrade retrofits from language test providers. They found no examples of change retrofits.

Test Retrofitting for Diagnostic Assessment

One area that does have a reasonable amount of documented test retrofits is diagnostic assessment. Retrofitting existing language tests with the procedures of cognitive diagnostic approaches could examine "the extent to which useful diagnostic information could be extracted from existing assessments before delving into an expensive, time-consuming process of designing a new diagnostic test" (Lee & Sawaki, 2009, p. 174). There are, however, challenges associated with retrofitting for diagnostic purposes because "this process goes directly counter to the inferences attempted to be made from the original assessment" (Liu et al., 2018, p. 358). Liu et al. (2018) proposed a continuous 4-step procedure for conducting a retrofit specific to cognitive diagnostic assessment (CDA) that is similar to that of Fulcher and Davidson (2009). Jang (2009) defined CDA as having "explicit links between learners' competencies in skills constituting the latent construct of interest and the characteristics of test items used to elicit the skills" (p. 210).

The first step is to gather information about the test-takers, the test itself, the item responses, and the test users. Having items that have already been developed and previous test tasker responses available for analysis is imperative to validate the inferences made after the retrofit. Second, the specific skills, knowledge, and competencies, or attributes, need to be defined for each item and inter-skill relationships and the number of attributes tested are considered at this stage. Then, diagnostic classification models (DCMs) must be retrofitted to the existing items to get diagnostic feedback from test items designed for classical test theory or item response theory. DCMs are multidimensional models used to classify latent variables. Test retrofitters also need to examine DCM fit statistics, attribute correlations, and reliabilities. Finally, the main goal of retrofitting a test for diagnostic purposes is to generate diagnostic feedback for each test-taker. This feedback could also be aggregated to inform planning and decision making, to suggest a sequence in which the attributes are learned, to improve construct knowledge, and to compare test-takers' results on the original test versus their multidimensional diagnoses (Liu et al., 2018).

Test Repurposing

"Failure to provide an explicit validity argument for a retrofitted test, especially when no modifications have taken place, should alert score users to the likelihood of invalidity and test misuse" (Fulcher & Davidson, 2007, p. 375). If those scores are used for high-stakes decisions, it may indeed be unethical to use "instruments that are not demonstrably relevant, useful, and sufficient for the defined purpose" (Fulcher, 2013, p. 5809). Fulcher and Davidson (2007) put forth that score users are likely to put their trust in testing companies to offer tests appropriate to their use, but they dismiss test repurposing, claiming it is only a way for testing companies to increase business. However, ETS researchers Wendler and Powers (2009) have claimed that it is not that simple. In addition to the competition between testing companies, customers typically want a high-quality test with reliable results that can be made available in a short amount of time. "It is wasteful not to take advantage of the good work carried out to support the original development of an assessment" (Wendler & Powers, 2009, p. 3).

ETS, however, does not use the term change retrofit. Instead, they use the term repurposing which they define in a near-identical way to Fulcher & Davidson's 2009 definition of a change retrofit. Test repurposing is defined as "using a test either for test-takers or for purposes that are different from those for which the test was originally developed" (Wendler & Powers, 2009, p. 1). The actual test questions are one of the most valuable resources when repurposing a test because they can be reassembled with a clear purpose, and test specifications are less tangible and concrete. Wendler and Powers (2009) also outline ETS' validity standards for repurposed tests and their overlap with

validity generalizability from the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA] et al., 2014).

Validity generalizability is how well the evidence used for one validation argument can be applied to a new validation argument without any further examination of the new context (AERA et al., 2014). The Standards outline three scenarios in which generalizing a past validation argument may be appropriate, namely "where the meta-analytic database is large, where the meta-analytic data adequately represent the type of situation to which one wishes to generalize, and where correction for statistical artifacts produces a clear and consistent pattern of validity evidence" (AERA et al., 2014, p. 18). A large quantity of past research may be more informative than a single new study with small sample size. So, when considering if a new validation study is required for the new context, it is important to consider the informational value of a new study versus the meta-analytic data available. A meta-analysis of related validation argument findings is an important part of any validation study (AERA et al., 2014). Fulcher (2013), however, argued that, while test scores might be generalizable, validation arguments are not. This is especially true for tests whose scores were used for life-changing decisions such as immigration. Validity through test design is what gives meaning to test scores, not validity through test retrofits (Fulcher & Svalberg, 2013).

Wendler and Powers (2009) outlined only four steps for repurposing a test (see Table 2 below). The reduced number of steps does not, however, suggest the procedure is any less complex. First, they determine the differences between the test's original use and its new, proposed use. In particular, they focus on the gap between the evidence gathered for the first validity argument and the new intended use. Next, they create a credible argument for why the new test-takers will interact with the test in the same way as the original intended audience. To do this, they consider if the test items and the test as a whole work as expected. Third, they make a validity argument by gathering as much evidence as possible in the short term and making a plan to gather evidence in the long term in order to solidly justify the repurposed test's use and score interpretation. Lastly, they promptly address any issues that arise which threaten the validity argument. Although this may seem like a linear process, it is in some ways cyclical given that they expect to examine the long-term use of the test and make changes as needed.

Procedure Comparison

When comparing the 3 discussed procedures for retrofitting and repurposing, it is easy to see many similarities (see Table 2 below). Although Fulcher and Davidson (2009) provided a more detailed procedure, each procedure starts with extensive information gathering and examination of the test items, its current uses and users, and its proposed use.

Table 2. Side-by-side Comparison of Procedures for Test Retrofitting and Repurposing

Fulcher & Davidson (2009)	Wendler & Powers (2009)	Liu et al. (2018)
(1) Set up a team of experts	(1) Determine the differences between the use the test was originally designed for and its new, proposed use	(1) Gather information about the test-takers, the test itself, the item responses, and the test users
(2) Gather relevant documentation	(2) Create a credible argument for why the new test-takers will interact with the test in the same way as the original intended audience	(2) Define the specific skills, knowledge and competencies, or attributes for each item
(3) Identify the purpose & type of retrofit	(3) Make a validity argument	(3) Retrofit diagnostic classification models (DCMs) to the existing items
(4) Research other tests used for the new purpose, decide if the retrofit will meet the new purpose, & consider the test-taker and social consequences	(4) Be prepared to deal with the unexpected	(4) Generate diagnostic feedback for each test- taker
(5) Consultations with stakeholders		
(6) Decide if the retrofit will proceed		
(7) Create a detailed retrofit plan		
(8) Confirm the retrofit plan meets current industry standards		

(9) Set a research agenda for new validation research
(10) Market the retrofit to the public
(11) Keep detailed documentation & records

Note. Adapted from Fulcher & Davidson (2009), Wendler & Powers (2009), and Liu et al. (2018)

Two of the procedures mentioned conducting validation research. Where they differ is in how the procedures are used and who has access to the information. Liu et al. (2018) aimed to provide low-stakes feedback for individual test-takers. Along with Lee and Sawaki (2009), they fully acknowledged the difficulties in retrofitting for diagnostic purposes. Liu et al. (2018) maintain that their procedure is about gaining insight into existing tests, and steps two and three examine how they may fit with diagnoses. The other procedures are used for tests that may have high-stake impacts on test-takers. Fulcher and Davidson's (2009) procedure is the only one that includes a step on whether or not a retrofit or repurposing should proceed. None of the procedures include any consultation with test-takers, either before or after a test is repurposed. Only Fulcher and Davidson (2009) made clear mention of stakeholder consultations, record keeping, and public accountability, and yet no available studies are using their procedures.

Publicly Available Examples

The examples below were all identified within the literature illustrating the three procedures.

Upgrade Retrofitting

The Test of English as a Foreign Language internet-based Test (TOEFL iBT) is "the largest upgrade retrofit to any language test" (Fulcher & Davidson, 2009, p. 135). ETS's TOEFL working paper outlined the rationale and procedures taken to update the exam based on user needs and new technology (Jamieson et al, 2000). The computer-based TOEFL included changes to virtually every component laid out in Table 1. However, ETS does not use the term "retrofit"; instead, ETS refers to the test as a new computer-based TOEFL test (Jamieson et al., 2000; Chapelle et al., 2008) and provides a new validity argument.

Test Retrofitting in Diagnostic Assessment

Jang (2009) stated serious difficulties arose from retrofitting a test for CDAs when its original purpose was not diagnostic. 11 ESL students of various academic and linguistic backgrounds sat reading comprehension sections from an established ETS test while simultaneously following think-aloud protocols in English. Trained raters identified the primary reading skills needed to answer each reading item. Jang analyzed these skills with test data from over 2000 test-takers and identified nine categories of reading skills. She found it is not possible to get a balanced distribution of attributes or to remove inter-skill relationships. So, she advocated for "new test development procedures [and] a principled test design for cognitive diagnostic assessment" (p. 235).

Liu et al. (2018) also found "there is no doubt that developing an assessment under the diagnostic framework is a better way to obtain diagnostic information than retrofitting" (p. 361). However, they argued retrofitting was best used as a way to learn more about the test construct or to make low-stakes decisions. Using their procedure, Liu et al. (2018) had 422 test-takers sit a placement test with 51 mock TOEFL listening items. Even though the results informed test-takers' and instructors' decisions on where to focus in a TOEFL preparatory course, the authors warned "while retrofitting presents a feasible approach to gain more actionable information from existing assessments of other psychometric frameworks under certain circumstances, much caution is needed to use and interpret DCMs [diagnostic classification models] appropriately" (p. 378).

Test Purpose Pluralism and High-Stakes Decisions

Fulcher and Davidson (2009) claimed that the use of an academic English university entrance test for immigration selection was an example of tests that have either not gone through a change retrofit and do not have any publicly available literature of the retrofit. Fulcher (2013) said it was increasingly common for test developers to carry out

studies that map their tests to documents like the Common European Framework of Reference or the Canadian Language Benchmarks and then claim their tests can be used for different purposes at the set cut scores[3]. The International English Language Testing System (IELTS) and the TOEFL iBT are two such tests that are being used for immigration purposes in various countries including Australia, New Zealand, the United Kingdom, and Canada.

The original purpose of IELTS was "to assess whether candidates were ready to study or train in the medium of English" (Merrylees, 2003, p. 2) but the test was soon used for immigration to New Zealand and Australia, so IELTS researchers wanted to consider how these immigrant test-takers contributed to their original test-taker profile. As the only immigration-related research report on the IELTS website, Merrylees' (2003) study arose from the need to gain insight into how IELTS had been repurposed by customers to meet a political agenda. The study examined the results and attitudes of test-takers taking IELTS for access to secondary education and those taking IELTS for immigration. Demographic data and test results of 379 participants who took IELTS for immigration were compared against 4,675 post-secondary test-takers. Then, 188 immigrant participants completed a Likert-scale questionnaire with statements related to ease of understanding the instructions and/or different accents, the difficulty of the tasks, and timings. Merrylees (2003) found the immigration test-takers had higher scores and generally liked IELTS although they wished for more time for the reading and writing sections and felt it should have only Australian accents in the listening section. This IELTS research report does not meet any of the proposed retrofitting or repurposing procedures in that there was no extensive information gathering on the test, its items, or the new context for use before repurposing IELTS. It did, however, allow for a rather extensive consultation with a particular stakeholder group: the test-takers themselves. Consultation with test-takers is not explicitly named in any of the three procedures described in this paper.

In a case study of two IELTS test-takers who had already been living and working in Australia for nearly 10 years, Hoang et al. (2017) reported a heavy toll on their participants' employment, financial, and psychological well-being as they struggled to get the test scores required for their immigration applications. Hoang (2019) surveyed and interviewed both IELTS and TOEFL test-takers who reported feeling a lack of trust towards the tests topics, scoring systems, and cut scores. However, test-takers felt the score-based judgements of their language proficiency were fair. Some participants argued that governments should accept other types of evidence of language proficiency and that the immigration language policies are too rigid. Hamid et al. (2019) surveyed 430 participants who sat IELTS about their perspectives on its fairness and validity, and its use for immigration. Many reported feeling that while it did not measure their language proficiency, it was fair overall. Also, they were overwhelmingly critical of IELTS as both a money-driven business and a barrier to immigration. These findings highlight the need for more research on the appropriateness of language proficiency test scores used as a way of allowing people to move and resettle. This is especially important since it was the test score users who implemented this policy and began using the tests in unintended ways.

Discussion and Implications for Future Research and Practice

Newton's (2017) argument for a multi-perspective approach to test design is to ensure that the multiple purposes inherent in all assessments are taken into consideration from the beginning of the development stage. While persuasive, this argument does not address current practices with high-stakes language tests in the context of immigration. Furthermore, as Fulcher (2013) claimed, there does seem to be a paucity of available literature regarding test retrofitting and repurposing and these procedures seem unlikely to gain wider use given the current practices in the field. What is not clear is exactly why this is the case.

It does seem then that using tests designed for a single purpose for multiple unrelated purposes is here to stay for high-stakes testing contexts such as immigration. IELTS has officially been in use for immigration for over 20 years in New Zealand and in 2010, the Canadian government-mandated tests used for immigration, such as IELTS, be mapped to the Canadian Language Benchmarks (Merrifield, 2008). However, there is still a lot of room for research into the consequences of this reality, particularly for high-stakes testing contexts such as immigration. Shohamy et al. (2009) stated that data for such research can be difficult or impossible to access and such research may need to focus on qualitative methods since it would likely involve an analysis of the policies and culture surrounding immigration policies (McNamara et al., 2011). As the examples provided in this paper show, examining the test-taker perspective is one way to shed light on current practices and contribute to a broader discussion of the validity of such language

[3] Cut scores are points on the score scale of a test that are used to classify a test-taker's performance into different levels of achievement.

testing practices. The test-takers and their families are the ones most impacted by this potential test misuse. However, it is immigration policymakers who start this practice and seek out testing companies to implement it. Studies presenting their perspective would also be critical in this discussion.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. AERA.

*Chalhoub-Deville, M. & Fulcher, G. (2003). The oral proficiency interview: A research agenda. *Foreign Language Annals, 36*(4), p. 498-506. doi:10.1111/j.1944-9720.2003.tb02139.x

*Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a Validity Argument for the Test of English as a Foreign Language*. Routledge Taylor Francis Group.

Cronbach, L. J. (1984). *Essentials of Psychological Testing Fourth Edition*. Harper & Row.

*Fulcher, G. (2013). Test design and retrofit. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 5809 - 5817). Wiley Blackwell.

*Fulcher, G. & Davidson, F. (2007). *Language Testing and Assessment: An advanced resource book*. Routledge.

*Fulcher, G. & Davidson, F. (2009). Test architecture, test retrofit. *Language Testing, 26*(1), p.123-144. doi:10.1177/0265532208097339

*Fulcher, G. & Svalberg, A. (2013). Limited aspects of reality: Frames of reference in language assessment. *International Journal of English Studies, 13*(2), p. 1-19.

Hamid, M. O., Hardy, I., & Reyes, V. (2019). Test-takers' perspectives on a global test of English: Questions of fairness, justice, and validity. *Language Testing in Asia, 9*(16), p. 1-20.

Hoang, N. T. H. & Obaidul Hamid, M. O. (2017). 'A fair go for all?' Australia's language-in-migration policy. *Discourse: Studies in the Cultural Politics of Education, 38*(6), p. 836-850. doi:10.1080/01596306.2016.1199527

Hoang, N. T. H. (2019). Building a validity argument for the use of academic language tests for immigration purposes: Evidence from immigration-seeking test takers. *Language Education & Assessment, 2*(3), p. 135-154. doi:10.29140/lea.v2n3.148

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 Framework: A working paper.* Educational Testing Service.

*Jang, E. E. (2009). Demystifying a Q-matrix for making diagnostic inferences about L2 reading skills. *Language Assessment Quarterly, 6*(3), p. 210-238. doi:10.1080/15434300903071817

Lamprianou, I. & Athanasou, J. A. (2009). *A Teacher's Guide to Educational Assessment*. Sense Publishers.

*Lee, Y. W & Sawaki, Y. (2009). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), p. 172-189. doi:10.1080/15434300902985108

*Liu, R., Huggins-Manley, A. C., & Bulut, O. (2018). Retrofitting diagnostic classification models to responses from IRT-based assessment forms. *Educational and Psychological Measurement, 78*(3), p. 357-383. doi:10.1177/0013164416685599

McAuliffe, M & Khadri, B.(Eds.).(2019). *World Migration Report 2020.* International Organization for Migration. https://publications.iom.int/system/files/pdf/wmr_2020.pdf

McNamara, T. & Ryan, K. (2011). Fairness versus justice in language testing: The place of English literacy in the Australian citizenship test. *Language Assessment Quarterly, 8*, p. 161-178. doi:10.1080/15434303.2011.565438

*Merrifield, G. (2008). *The use of IELTS for assessing immigration eligibility in Australia, New Zealand, Canada, and the United Kingdom.* IDP: IELTS Australia and British Council. https://www.ielts.org/-/media/research-reports/ielts_rr_volume13_report1.ashx

*Merrylees, B. (2003). An impact study of two IELTS user groups: Candidates who sit the test for immigration purposes and candidates who sit the test for second education purposes. In R. Tulloh(Ed.), *IELTS Research Reports 2003 Volume 4*(pp.1-58). IELTS Australia Pty Limited. https://www.ielts.org/-/media/research-reports/ielts_rr_volume04_report1.ashx

*Newton, P. E. (2017). There is more to educational measurement than measuring: The importance of embracing purpose pluralism. *Educational Measurement: Issues and Practice, 36*(2), p. 5-15. doi:10.1111/emip.12146

Shohamy, E. & McNamara, T. (2009). Language tests for immigration, citizenship, and asylum. *Language Assessment Quarterly, 6*(1), p. 1–5. https://doi-org.proxy.queensu.ca/10.1080/15434300802606440

*Wendler, C. & Powers, D. (2009). What does it mean to repurpose a test? *R & D Connections, 9*, p. 1-8. https://www.ets.org/Media/Research/pdf/RD_Connections9.pdf

ABOUT THE AUTHOR

Melissa McLeod is a second-year PhD student in the Faculty of Education at Queen's University. She taught ESL/EFL for 17 years in Canada and abroad. Her primary research interests are refugee education and refugee integration.