Internationally Educated Nurses and the Canadian English Language Benchmark Assessment for Nurses: A Qualitative Test Validation Study of Test-Taker Accounts[1]

Stefanie Baldwin
D2L

Liying Cheng
Queen's University

Abstract

This qualitative validation study examines sixteen Internationally Educated Nurses' (IENs') accounts of the Canadian English Language Benchmark Assessment for Nurses (CELBAN) at two testing centres (Toronto and Hamilton). This study adopts both focus groups and one-on-one interviews to investigate the inferences drawn from the test, and its consequences. Focus groups and interviews were conducted using an adapted interview guide utilized in the TOEFL iBT investigation of test-taker accounts of construct representation and construct irrelevant variance (DeLuca et al., 2013). While construct representation describes the degree of authenticity in the presentation of Canadian English language nursing tasks, construct irrelevant variance refers to potential factors impacting the test-taking experience which might contribute to a score variance that was not reflective of test-taker knowledge of the testing constructs (Messick, 1989, 1991, 1996). In this study, test-taker accounts of construct representation and construct irrelevant variance constituted the data which were coded and analyzed abductively via the sensitizing concepts derived from DeLuca et al., and Cheng and DeLuca (2011) on examining test-takers' experience and their contribution to validity. Seven themes emerged, answering four research questions: How do IENs characterize their test experience? How do IENs describe the assessment constructs? What, if any, sources of Construct Irrelevant Variance (CIV) do IENs describe? Do IENs feel the language tasks are authentic? Overall, participants reported positive experiences with the CELBAN, while identifying some possible sources of CIV. Given the CELBAN's widespread use for high-stakes decisions (a component of nursing certification and licensure), further research of IEN-test-taker responses to construct representation and construct irrelevant variance will remain critical to our understanding of the role of language competency testing for IENs.

Résumé

Cette étude qualitative de validation examine les témoignages de seize infirmières et infirmiers formés à l'étranger (IFE), à propos du test *Canadian English Language Benchmark Assessment for Nurses* (CELBAN), dans deux centres d'examen (Toronto et Hamilton). Cette étude adopte à la fois des groupes de discussion et des entretiens en tête à tête afin d'enquêter sur les inférences tirées du test et ses conséquences. Les groupes de discussion et les entretiens ont été menés avec un guide d'entretien adapté, celui-ci utilisé dans l'enquête TOEFL iBT consistant de témoignages des candidats au CELBAN sur la

représentation des construis et la variance non pertinente des construis (DeLuca et coll., 2013). Tandis que la représentation des construis décrit le degré d'authenticité qui se trouve dans la présentation des tâches des infirmières et infirmiers en anglais canadien, la variance non pertinente des construis fait référence aux facteurs potentiels qui pourraient influencer l'expérience de passation du test, et qui pourraient contribuer à une variance de résultats ne reflétant pas les connaissances des construis d'évaluation que possède le candidat (Messick, 1989, 1991, 1996). Dans cette étude, les témoignages des candidats sur la représentation des construis et la variance non pertinente des construis constituent les données codées et analysées de manière abrégée à travers les concepts de sensibilisé dérivés de Cheng et coll. et de Cheng et DeLuca (2011) sur l'expérience des candidats et leur contribution à la validité. Sept thèmes ont émergé, qui répondaient à quatre questions de recherche : comment les IFE représentent-ils l'expérience de passation du test ? Comment les IFE décrivent-ils les construis d'évaluation ? Quelles, s'il y en a, sources de la variance non pertinente des construis (VNC) les candidats décrivent-ils ? Les IFE pensent-ils que les tâches de langue sont authentiques ? Dans l'ensemble, les participants ont rapporté des expériences positives avec le CELBAN, en identifiant aussi quelques sources possibles de VNC. Étant donné l'usage répandu du CELBAN pour des décisions à enjeux élevés (une composante de la certification et du permis d'exercer des infirmières et infirmiers), d'autres recherches sur les réponses des candidats IFE à la représentation des construis et de la variance non pertinente des construis resteront cruciales à notre compréhension du rôle de l'évaluation des niveaux de compétence linguistique chez les IFE.

Internationally Educated Nurses and the Canadian English Language Benchmark Assessment for Nurses: A Qualitative Test Validation Study of Test-Taker Accounts

The Canadian Nurses Association has indicated concern that as large numbers of Canadian nurses reach retirement age, a "critical shortage of nursing professionals" could result (Epp & Lewis, 2009, p. 286). Further, overreliance on immigration alone may prove costly, as challenges in meeting certification and licensure testing requirements, combined with non-recognition of international credentials, continue to remain "the largest barriers to successful integration into the workforce, which costs the Canadian economy as much as $5 billion a year" (Cheng et al., 2013 p. 734). The Canadian English Language Benchmark Assessment for Nurses (CELBAN) was designed to address the "target language use" (Epp & Lewis, 2009, p. 285) of the nursing profession, nation-wide. Either the CELBAN and International English Language Testing System (IELTS) may be used to demonstrate English language competency to the College of Nurses of Ontario (CNO, 2013). This qualitative study examines the experiences of International Educated Nurses (IENs) who completed the CELBAN in pursuit of their licensure and certification to work in the nursing profession, in Canada, including their accounts of: the assessment as a whole, its testing constructs, possible sources of irrelevant test score variance, and the authenticity of the construct representation of nursing communicative tasks. The findings of the study add to the critical understanding of what is tested by the CELBAN and the role of large-scale language testing in the licensure and certification experience of IENs in Canada.

Literature Review

The theoretical and empirical evidence that inform this study include the discussions of high-stakes, large-scale assessment (DeLuca et al., 2013; Fox & Cheng, 2007, Nagy, 2000), and stakeholder accounts of language testing (Cheng & DeLuca, 2011; DeLuca et al., 2013; Malone & Montee, 2014; Messick, 1996). Test validation in relation to construct representation and construct irrelevant variance have further informed and guided the study (Cheng & DeLuca, 2011; DeLuca et al., 2013; Messick, 1991, 1996). Emerging research in validity has argued for the need to include values-based and consequential validity evidence within validity arguments (Cheng, 2014; Koch & DeLuca, 2011). There is a growing emphasis on collecting and analyzing a variety of validity evidence through multiple methods and multiple stakeholders including test-takers. Haertel (2013) suggests including focus groups, surveys, and interviews, as a viable approach for validation activities that aim to account for value-based and consequential validity evidence. Moss et al. (2006) further describe how qualitative methods could contribute meaningful information to validation studies. The analysis of test-taker as stakeholder accounts formed the data for this study, in a manner consistent with the qualitative data collection process described by Polkinghorne (2005) who suggests "the purpose of data gathering in qualitative research is to provide evidence for the experience it is investigating. The evidence is in the form of accounts people have given of the experience" (p138). Therefore, prior research of high-stakes, large scale assessment as explored through test-taker accounts has informed the research design of this study on the CELBAN.

High-Stakes Large Scale Assessment and Stakeholder Accounts

DeLuca et al. (2013) define high-stakes testing with respect to the decisions made, as a result of test measurement, which has a significant impact on test-takers, as in the case of the following types of examples: "use of [a] test for high-stakes decisions (e.g., university admissions, scholarship, promotion)" (p. 664). Fox and Cheng (2007) define large-scale testing as "used to measure and ensure student competency or provide system accountability" (p. 11). Nagy (2000) argues assessment, inclusive of large-scale assessment, serves three functions: gatekeeping, accountability, and instructional diagnosis. Collectively, these terms are used to describe language assessments like the CELBAN, which are used to make a decision about a test-taker that will have a large impact on them personally (high-stakes, gatekeeping for licensure and certification), and ensure system accountability (large-scale assessment). Nagy's final assessment function, "instructional diagnosis tool" (p. 262) is also fitting in terms of the CELBAN where, test-takers are afforded the opportunity to use the results of their first attempt at the CELBAN to improve upon areas of professional language weakness, and re-take the assessment (CELBAN, n.d.b; The CELBAN Centre, 2018).

The Test of English as a Foreign Language internet-based test (TOEFL iBT) is a similar type of assessment to the CELBAN, in its categorization as a high-stakes, large-scale assessment. A 2014 study (Malone & Montee, 2014), contributing to the validity argument for the TOEFL iBT, utilized focus groups, survey data, and stimulated recall to sample stakeholder accounts of task authenticity in regard to construct representation. Malone and Montee explain that "stakeholder beliefs about the correspondence between

test tasks and the skills they purport to test provide important support for test validity" (p. 2). This is consistent with the views of Messick (1996) in the suggestion that "test preparation practices emphasizing test familiarization and anxiety reduction may actually improve validity: scores that formerly were invalidly low because of anxiety might now become validly higher" (p. 6). The interaction, or experience, of the test, is especially important in the case of the CELBAN, where assessment tasks are intended to model target language use of the profession. Where it is not clear to test-takers which target language use they are meant to demonstrate when completing an assessment item, the faith we may have in the score reported by that assessment question – as it relates to their skill level may be undermined.

Although research in assessment has shown that "all tests vary in their ability to represent the testing construct to some extent, validity is also affected by the way in which the test-takers approach common test formats" (Baldwin-Bojarski, 2016). p. 28). For example, Rupp et al. (2006) have offered evidence to support the theory that test-takers approach multiple-choice questions as problem-solving tasks rather than as comprehension tasks. When test-takers respond to comprehension tasks (the testing construct) in a manner consistent with process-of-elimination problem-solving strategies, the validity of that testing format may be called into question.

Validity, Consequences and Construct Irrelevant Variance

This paper relies heavily on the works of Messick (1989, 1991, 1996) to define test validity, and threats to test validity because they are seminal works as cited by (Cheng & DeLuca, 2011; DeLuca et al., 2013; Fox et al., 2013; Haladyna & Dowling, 2004). Validity may be understood as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores of other modes of assessment" (Messick, 1991, p. 1). While tests may be used for a variety of evaluative purposes in the case of the CELBAN, test results guide the decisions to progress or hold back an application for licensure and certification.

Messick (1991) identifies two major threats to test validity: (1) construct underrepresentation and (2) construct irrelevant variance. Construct underrepresentation is defined as an occurrence whereby a test "…is too narrow and fails to include important dimensions or facets of the construct [the knowledge or skill being assessed]" (p.14). Construct irrelevant variance is defined as an occurrence whereby a test …"is too broad and contains excess reliable variance associated with other distinct constructs as well as method variance making items of tasks easier or harder for some respondents in a manner irrelevant to the interpreted construct" (Messick, 1991, p.14). This concept is expanded upon in a discussion of "performance assessment" (Messick, 1996, p.1) as it pertains to language teaching. Messick (1996) suggests, "in the case of language testing, the assessment should include authentic and direct samples of the communicative behaviours of listening, speaking, reading, and writing of the language being learned" (p. 1). In accordance with this view of validity, this study examines test-taker accounts of the skills tested by the CELBAN, and their experience completing the assessment. Therefore, this study contributes information to our understanding of the evaluative judgements made from CELBAN scores.

Test validation, not only in the case of the CELBAN but in all assessments, is an ongoing process of gathering and evaluating data. As such, validity evidence may be considered to be always incomplete. As Messick (1991) suggests, "validation is essentially a matter of making the most reasonable case, on the basis of the balance of evidence available, both to justify current use of the test and to guide current research needed to advance understanding of what the test scores mean and how they function in the applied context" (p. 2).

Cheng and DeLuca's (2011) study of the relationship between aspects of the testing experience and test-takers' perspectives of test validity and use reveal a number of construct-irrelevant factors that can negatively affect the perception of validity within a high-stakes, large-scale testing environment. Eight themes emerged from their study including: test administration and testing conditions; timing, test structure and content; scoring effects; preparation and test-taking strategies; test purpose; psychological factors; external factors; and consequences (Cheng & DeLuca, 2011). This study revealed that environmental factors largely impacted test-takers' accounts of test validity (Cheng & DeLuca, 2011).

A 2013 study on the factors affecting student performance on the TOEFL iBT also shed light on construct-irrelevant variance and its potential impact on test-taker scores (DeLuca et al., 2013). DeLuca et al. identified three themes, which emerged from the focus groups conducted with key informants following the completion of the TOEFL iBT under the pattern of Construct Irrelevant Variance: testing environment, test design, and score reports. Of particular significance for the CELBAN study, were the findings on testing environment factors. The testing environment factors found to have contributed to construct irrelevant variance in the TOEFL iBT included: the amount of travel time required to access a testing location, the time of day at which the test began and its duration, ambient noise in the test-taking facility, administration procedures that interrupted test-takers, and security protocols that led to personal discomfort e.g., being required to remove shoes in a chilly test-taking room (DeLuca et al. 2013).

The findings of Cheng & DeLuca (2011) and DeLuca et al. (2013) have emphasized the importance of: the testing environment, test design, and score reporting on the effects of construct irrelevant variance. These, in combination with the eight themes that emerged from the Cheng & DeLuca study, constitute sensitizing concepts applied in this study, within a wider abductive approach to the coding of focus group and interview data following the completion of the CELBAN.

Although the CELBAN is quickly approaching its 20th anniversary, relatively few publications have addressed its significant role in widespread use for high-stakes decisions (a component of licensure and certification) (Baldwin-Bojarski, 2016; Epp & Lewis, 2009; Lewis & Kingdon, 2016), and fewer still on test-taker accounts of the testing constructs (construct representation) and their experience with the assessment (possible sources of construct irrelevant variance) which are critical to our understanding of the inferences made from CELBAN test scores. This study seeks to further our understanding of the role of professional language competency testing in the licensure and certification of IENs, specifically with respect to the CELBAN test-taking experience. Therefore, this study addresses the following four questions in relation to IENs taking the CELBAN test:

1. How do IENs characterize their assessment experience?
2. How do IENs describe the constructs (English reading, writing, listening, and speaking proficiency) measured through the CELBAN?
3. What, if any, potential sources of construct irrelevant variance do IENs describe based on their assessment experience?
4. Do IENs feel that the CELBAN tasks provide a good reflection of the types of communicative tasks required of a nurse?

Method

This study employed a qualitative approach using focus group discussions and one-to-one interviews. Three, one-and-a-half-hour focus groups were conducted with sixteen IEN test-takers who were required to complete the CELBAN as a component of the licensure process in Canada. Additionally, nine half-hour, one-on-one interviews were conducted following the completion of test-taker speaking assessments. Focus groups and interviews used the interview guide adapted from a pre-existing tool utilized in the TOEFL iBT investigation of construct representation and construct irrelevant variance (DeLuca et al., 2013). Table 1 provides an overview of the focus group and interview guides and how they align with the research questions of this study.

Table 1
Research Questions and Instrument Alignment

Research Questions	Instrument Alignment
1) How do IENs characterize their CELBAN experience?	Section 1 of both Focus Group and Interview Guides ex. "Tell me in general about your experience with the CELBAN" and "Describe your experience with the oral interview"
2) How do IENs describe the constructs (reading, writing, listening, and speaking proficiency) measured by the CELBAN?	"How do you feel the assessment tested your English language skills?" and "How do you feel the oral assessment tested your English language skills?"
3) What, if any, potential sources of construct irrelevant variance do IENs describe based on their CELBAN experience?	"Were there any technical difficulties during the test?" and "Were you aware of any observers in the room while you completed the assessment?"
4) Do IENs feel that the assessment tasks provide a good reflection of the types of communicative tasks required of a nurse in Canada?	"Do you feel the exam tasks were similar to what you would have to do as a nurse while working in Canada?" and "Do you feel these types of tasks are realistic in terms of the interactions you have in a Canadian hospital?"

Participants

A purposive sampling technique was used to contact potential participants who met the criteria of having registered to write the CELBAN at an assessment centre in Ontario in the winter of 2015. Touchstone Institute Competency Evaluation Experts, formerly CEHPEA, distributed a copy of an e-mail inviting participation in the study, as well as a Letter of Information and Consent Form to IENs, registered to take part in the CELBAN at either the Toronto or Hamilton test centres.

Sixteen IENs participated in the study, comprised of eleven female, and five male participants ranging in age from twenty to forty years old. The majority of participants (11) were originally from India. The unit of analysis in this study was the individual IEN. A pseudonym was given to each participant for anonymity. Information was gathered from individual responses to the research questions included in the semi-structured focus group and interview guides (see Table 1).

Instruments

The CELBAN seeks to recognize the IEN's ability to use English to accomplish communicative tasks associated with nursing; an acknowledgement that IENs are already trained in the practice of nursing in another country (Epp & Lewis, 2009). As a national testing instrument of professional language competency, the CELBAN is used to evaluate language proficiency in four skill areas: reading, writing, speaking and listening (Lewis & Kingdon, 2016).

In the year 2000, a survey of fifty nursing profession stakeholders was conducted by the Centre for Canadian Language Benchmarks (CELBAN, n.d.a). This feasibility study indicated a "strong need for a specialized English language assessment tool to evaluate the English language communication skills of internationally educated nurses seeking registration in Canada" (CELBAN, n.d.a). Introduced in 2004, the CELBAN is intended to measure the professional language competencies of IENs. In 2014, as this study was being conducted, the CELBAN underwent a new phase of growth and development as it transitioned from the Canadian English Language Assessment Service at Red River College in Winnipeg, Manitoba, to the CELBAN Centre at Touchstone Institute in Toronto, Ontario (Lewis & Kingdon, 2016). Additionally, this time also marked the beginning of the CELBAN test renewal project which sought to, "develop additional forms of the test that would retain the salient features of the original model and introduce some new task types and fresh content" (Touchstone Institute, 2019b, p. 2). This study examines CELBAN test-taker experience and feedback on test iterations in circulation in the winter of 2015.

Currently, in the year 2020, and at the time of this study, CELBAN requires approximately three and one-half hours to complete (CELBAN, n.d.a; Touchstone Institute, 2019a). It is a task-based, or genre-specific evaluation, which is scored according to the Canadian Language Benchmarks (English). Test-takers can attempt the CELBAN a maximum of three times, but they must wait a minimum of three months between test attempts (CELBAN, n.d.a; The CELBAN Centre, 2018). To successfully complete the CELBAN, test-takers must achieve the following Canadian Language Benchmark scores: Listening 10, Speaking 8, Reading 8, Writing 7, (CNO, 2013; The CELBAN Centre, 2018).

If an IEN is unable to meet the minimum level for a pass on the listening section, or on any of the other sections of the test taken in a group setting (reading and writing), then the test-taker must re-take this entire section of the exam (CELBAN, n.d.b; The CELBAN Centre, 2018). The content and format of the CELEAN is illustrated in Table 2.

Table 2
Content and format of the CELBAN in 2015

Skill Area	Skills Assessed	Format	Length[a]
Speaking Assessment	Candidates are expected to demonstrate their ability to -narrate -describe -summarize -synthesize -state and support an opinion -advise	1 oral interview 2 role-plays 2 assessors	30
Listening Assessment	Candidates are expected to demonstrate their understanding of conversations in the following settings: -hospital -home -clinic -medical office Conversations may be between nurses and: -patients -family members -other professionals	5 video scenarios 4 audio scenarios multiple choice	45
Reading Assessment	Candidates are expected to demonstrate comprehension of these text formats: -charts -patient notes -manuals -information texts related to health issues	1 skimming and scanning section (short answer) 1 reading comprehension section (multiple choice and cloze)	50
Writing Assessment	Candidates are expected to demonstrate their knowledge of the following areas: -conventions of form filling including the use of correct spelling, legibility of writing, and the use of point form -inclusion of necessary information including main ideas and supporting details -conventions of a narrative report including use of vocabulary, effectiveness of the report, and correct grammar	1 form filling section 1 report writing section	30

[a]The length is given in minutes.

Focus groups and interviews used an adapted interview guide utilized in the TOEFL iBT investigation of test-taker accounts of construct representation and construct irrelevant variance (DeLuca et al., 2013) (see Table 1 for the interview guide in relation to the research questions). In addition, focus groups and interviews began with general questions on the assessment experience, for example: "How would you describe your test-taking experience today?"

The combination of both focus groups and individual interviews benefited the data collection process in a variety of ways. Focus groups are fundamentally a way of listening to participants and learning from them as a group (Krueger & Casey, 2000) as they share a common experience, the CELBAN test-taking experience. The focus group method supported participant sharing, in the form of thoughts, feelings, attitudes, and ideas about their experiences with the CELBAN, within a small group of IEN peers. This was especially beneficial for this study focus which asked participants to consider aspects of the assessment that were not their primary focus during the assessment. The individual interviews which took place following the speaking assessment provided participants with an additional opportunity to contribute thoughts or feelings they may not have shared in a group setting, for those whose schedules permitted one-on-one interviews.

Data Collection Procedures

To prepare for the data collection, a semi-structured interview guide was piloted with an Advanced English Language Learner. This individual is also an Internationally Educated Professional with previous direct experience of large-scale, high-stakes English language testing for the purposes of admission into a post-graduate educational program in Canada. McMillan & Schumacher (2010) have advised that "a pilot test is necessary as a check for bias in the procedure, the interviewer, and the questions" (p. 206). The pilot encouraged revisions to the order, structure, and clarity of language in the interview guide, particularly for lengthy questions which the piloted participant noted as being challenging to respond to.

Data collection was comprised of three, one-and-a-half-hour-long focus groups as well as nine half-hour-long individual interviews. Data collection took place at two testing centres (Toronto and Hamilton) on the day of test-takers' CELBAN assessment. Due to the format of the assessment's two-part structure, focus groups were intended to follow the group section of the assessment with individual interviews following the speaking section. In Toronto, three participants discussed their entire CELBAN test-taking experience in the format of a Focus Group to accommodate their schedules and availability In Hamilton, one participant shared their views on the entire CELBAN experience in an interview setting, due to a scheduling conflict that prevented their participation in a later focus group.. Focus groups and interviews took place in unoccupied classrooms or assessment rooms.

Each session began with an explanation of the purpose of the study, information about the recording of the focus groups and interviews, and a request for the participants to complete a brief demographic survey, except in cases where this information was reviewed and collected with the participant in a prior focus-group setting. Focus groups and interviews were recorded on two recording devices, to mitigate technical recording issues, i.e., indistinct responses, or background noise interference. The focus groups and interviews were audio-recorded and transcribed verbatim to facilitate analysis.

Data Analysis

Focus group and interview transcripts were coded abductively via the sensitizing concepts derived from Cheng & DeLuca (2011) and DeLuca et al. (2013) on examining test-takers' experiences and their contribution to validity. The abductive coding method "combines the deductive and inductive modes of proposition development and theory construction" (Patton, 2002, p. 470). This method is especially useful in allowing researchers to be guided by the sensitizing concepts which further "sensitizes the analyst[s] to the possibility of a category or behaviour that either has been overlooked in the data or is logically a possibility in the setting but has not been manifested" (Patton, 2002, p. 470). The inductive component of the overall abductive method was utilized in identifying emergent analytic codes not addressed by the sensitizing concepts in a manner consistent with the description offered by Patton.

In total, eight sensitizing concepts derived from Cheng & DeLuca (2011) and DeLuca et al. (2013) were used to guide the inductive coding of focus group and interview transcript data: testing environment, test design, score reports, test structure and content, preparation and test-taking strategies, test purpose, psychological factors, external factors and consequences. Data analysis was first conducted independently. The transcripts were reviewed, and the codes applied, guided by the sensitizing framework. Where a concept or description central to a research question was not addressed by a code offered by the sensitizing framework, one was created to address this concept. Then, previously coded transcripts were reviewed to confirm if this concept was described there, as well. This inductive component of the overall abductive method was utilized in identifying emergent analytic codes not addressed by the sensitizing concepts in a manner consistent with the description offered by Patton (2002).

In total 21 codes emerged to describe the data. These codes were reviewed and grouped under broader themes which more broadly described them. Codes and themes were then reviewed by both researchers, with themes aligned to the guiding research questions of this study. In total 7 themes describe the data. The coding and thematic structure was further reviewed by another researcher to explore the opportunity to more specifically define themes and the appropriate grouping of codes. Following this review, the structure outlined in Table 3 was determined to best represent the data according to the themes outlined. This approach to collaboration on the review and analysis of codes allowed for additional refinement of theme definition and was intended to offer evidence of efforts toward establishing inter-rater reliability. NVivo 10 was used to facilitate the recursive process of applying research-driven themes to fit deductive and inductive codes.

Results

Seven themes along with twenty-one codes emerged from the data analysis. Themes and codes are illustrated in Table 3 according to their correspondence with the four research questions of this study. Frequency of the occurrence of a code is also included in Table 3 to denote the commonality of these shared experiences, however, frequency was not used as a means by which to arrive at the themes. In some instances, less commonly shared views may be represented by a low frequency of coding in the data. For example, the comments made on cultural differences (see code 21 in Table 3) occurred less

frequently, but the significance of these comments merit their prioritization as a significant contributor to a theme because it sheds light on a research question. The results are sequentially reported according to the four research questions posed for this study.

Table 3

Research Questions, Themes, Codes & Frequencies

Research Questions	Themes and Codes	Frequency
1. How do IENS characterize their assessment experience?		
	• Code 1 CELBAN assessment as preference	23
	• Code 2 Ease	52
	• Code 3 Challenges	111
	• Code 4 Discomfort	54
	• Code 5 Assessment aspects described positively	96
	• Code 6 Next steps for self	12
2. How do IENS describe the constructs measured by the CELBAN?		
	• Code 7 Language skills assessed	123
	• Code 8 Nursing skills assessed	61
3. What, if any, potential sources of Construct Irrelevant Variance do IENS describe based on their experience		
	• Code 9 Immigration and CELBAN	42
	• Code 10 Psychological factors	108
	• Code 11 Licensure process	55
	• Code 12 Family pressure	8
	• Code 13 Financial stressors	14
	• Code 14 Knowledge of CELBAN	86
	• Code 15 CELBAN resources	36
	• Code 16 Timing	51
	• Code 17 Fairness protocols	10
	• Code 18 Assessment structure and design	155
	• Code 19 Administrative challenges to completion	72
	• Code 20 Exceptional administrative organization	31
4. Do IENS feel the assessment tasks provide a good reflection of the types of communicative tasks required of a nurse?		
	• Code 21 Cultural Differences	5

Research Question 1: How Do IENs Characterize Their Assessment Experience?

Test-takers were asked how they would describe the CELBAN and their test-taking experience. Responses like those following describe a positive test-taking experience and a common preference for the CELBAN over other language tests. Mary explained that, in her opinion, the CELBAN:

> … asked us to write something we are used to writing. Because we write nurse's report[s]. Always we document things – so that's something nice – because they're not asking you to explain a graph … which is what exactly happens in IELTS where they give you some population graph or things which – it's ok it's good – but it's not as related to nursing whereas here everything [in CELBAN] was nursing.

This response describes the CELBAN in positive terms based on the approximation of the tasks to the intended purpose, to effectively communicate medical information. As noted in Table 3, participants occasionally contrasted their positive CELBAN experience with less positive experiences with other language proficiency tests that were not focused on nursing. Sandy also described the CELBAN positively in contrast to her experience with IELTS saying: "The scenarios are much … easier, and from your life experience, patient care area. IELTS is different." In Toronto, Emily's comments agreed with those of Mary and Sandy in Hamilton saying: "In general I think it was very good because it's all from medical, from nursing content. It's everyday hospital scenario. So, I think it's a good way to assess nursing language, English language for nurses." This type of feedback from test-takers suggests that a professional language assessment offers a closer approximation of the Target Language Use (TLU) tasks, and thus greater authenticity of the task (Grabowski & Dakin, 2014), consistent with the goals of developing the CELBAN, specifically with regard to "face validity" (Epp & Lewis, 2009 p. 296). These findings concur with those of Cheng & DeLuca (2011) that construct underrepresentation may occur when a test designed for English academic purposes or university entrance is used for professional certification purposes (p. 106).

Some test-takers had less positive experiences. For example, Penelope questioned the validity of her CELBAN experience based on the sequence of certification assessments, as she explained:

> I think in Ontario they d[o]n't require [you] to pass the English exam first but – in other province[s] before you start studying your nursing you have to clear your English exam…if you clear English that will also help with your nursing exam as well… I already clear[ed] all [my] exam[s] for nursing – I didn't clear this exam. If I didn't clear this exam there is [no] meaning of [the] nursing exam and that is all a waste of time and money. So, I think CNO should [make it a requirement to] take this exam first and then the nursing exam.

Penelope's comments highlight how the wider certification process can affirm or undercut the "face validity" (Epp & Lewis, 2009, p. 296) of the individual assessments. In Penelope's case, a "pass" on her professional nursing assessment, administered and written in English, was not enough to signify her eligibility to work in her field, because she had

not yet passed her English language exam. Her comments suggest that, in her view, the certification experience could be improved by establishing the English language assessment as a pre-requisite to the professional competency exam to maintain the face-validity of the overall certification experience of IENs. All applicants must apply first to the National Nursing Assessment Service to be eligible to engage with the certification process to work as a nurse in Ontario (CNO, 2015). This means, theoretically, that an IEN may complete and pass the certification exam that assesses professional competency to practice but fail the English language assessment. As in the case of Penelope, this certification experience inclusive of the CELBAN may lead to a decrease in the perception of face validity (Epp & Lewis, 2009) of the IEN certification and licensure process as a whole.

Research Question 2: How Do IENs Describe the Constructs Measured by the CELBAN?

Participants were asked about the skills assessed by each section of the CELBAN. Their responses highlight a sense of confidence in the fairness of CELBAN, but also identify areas where test-takers were uncertain of the assessment's objectives. When asked about which skills he was tested on in the reading section, Jake explained:

> … there was a part of the skimming and scanning where … I was able to look for specific information from the information given and was also … able to extract information from the passage, which is … good because it will test my knowledge in skimming and scanning and comprehension also.

Participants identified the language skills tested by the assessment and associated these skills with particular questions, question types, or sections of the assessment. That test-takers saw a relationship between assessment sections and language skills suggests congruence between the types of tasks assessed by the CELBAN and what Messick (1996) deems "the content aspect of construct validity" (p. 248). This information is critical to our understanding of authenticity and fairness in the assessment of language skills because the intended purpose of the CELBAN is to address the language demands or "target language use" of the nursing profession, nation-wide (Epp & Lewis, 2009, p. 285).

Although the CELBAN is intended to assess language skills, the nursing-specific context of the test at times offered moments of confusion to some participants. Some test-takers described sections of the CELBAN where they prioritized their nursing knowledge over their language knowledge as a test-taking strategy. Sandy explained:

> …for [a patient whom that has] rheumatoid arthritis … there was a blank, what should we put under the knee? Like we have to select either knee, wound, or what. So if I was clear about rheumatoid arthritis, [it] is [a] joint problem … that's why I selected knee. … Like because it's a joint disease so I selected the ah one pillow should – shouldn't be under the knee because it's a joint problem – knee's a joint – so because of my nursing knowledge I feel I selected the correct thing, the correct choice.

While it is unclear which answer Sandy selected, what is clear, is that her answer selection was influenced by her professional knowledge rather than focusing specifically on her language skills. This type of test-taking strategy is similar to that described by Rob, who agreed with Sandy's approach in the focus group, stating:

> So what she said is entirely correct because if you don't have proper knowledge about that topic, you will not be able to fill the form. Because like what she said in the question is like, '… pillow should be placed under dash' and the options are head, joints and knee. So we know that rheumatoid arthritis is a joint problem so we'll of course select the answer "knee." But if you don't have that knowledge about rheumatoid arthritis – because we [are] usually placing pillows under the head – you might choose the option like head.

The type of question Sandy and Rob described in these quotations is a multiple-choice cloze exercise, which appeared in the reading section of the CELBAN. The Centre for Canadian Language Benchmarks' (2014) *Test Taking Strategies* document outlines the following instructions to complete cloze exercises: "Read the entire passage once through quickly, ignoring the blanks to get the gist (the general content of the text). Then read again for meaning, line by line to select the best option from the multiple choices provided to fill in the blanks" (p. 6). This instruction indicates that the best test-taking strategy is to focus primarily on the context of the reading passage. This suggests Sandy and Rob might have benefitted from adjusting their test-taking strategy from one that prioritized their sense of patient need, to one that identified the context clues of the paragraph.

Research Question 3: What, if any, Sources of Construct Irrelevant Variance (CIV) do IENS Describe?

Participants described challenges in completing the CELBAN due to the availability of test locations and dates, access to test preparation materials, and administration of the assessment. During data collection, the process for writing the CELBAN was undergoing significant changes affecting when, and where the assessment might be written. Test-takers like Jake and Janet flew from Manitoba to Ontario to complete their CELBAN assessments. While others may not have had to fly in, some like Emily were required to drive for most of the day to get to a test-taking center and home:

> Interviewer (I)- "4:30 in the morning? So, you left at 4:30 in the morning you were here by 8:15 and your test started here at 9 and then when is your speaking exam?"
>
> Emily- "At 3"
> I- "At 3 O'clock so by 3:30"
> Emily- "3:30 or 4 probably if I go"
> I- "You're gonna get back in the car?"
> Emily- "Think so, I guess so"

From April to November 2016 there were no test dates listed for any of the test centres in Ontario, outside of Toronto. The limited number of testing locations available then for writing the CELBAN is a factor outside of the testing construct itself that could affect test-takers travelling long distances to complete their assessment. This commitment to travel extensively to complete the CELBAN is evidence of the deep appreciation many IENs have for this assessment. Although limited test-taking locations had a high impact on a few test-takers, several test-takers indicated frustration due to the challenge associated with locating appropriate assessment preparation materials.

The most frequently cited concern in regard to CELBAN preparation was the challenge in locating CELBAN-specific study materials:

> I went into the internet and the libraries to look for resources and there were none-so that really scared me. But then, … my friend told me that because we don't have any idea, it's better that we take a prep course. Which I did take, paying a lot of money, but I think it was no – It was no help. Because I think no one really knows what CELBAN is. Even the people claiming to be teaching CELBAN have no idea what the CELBAN is at all. Because what I studied in the class and what I did today was totally, entirely different (Mary).

At the time of data collection, there were no CELBAN study guidebooks akin to the types of preparation books available for the IELTS or TOEFL test. While IELTS and TOEFL cater to a global population with a wide variety of professional and educational goals, the CELBAN's audience is far more limited: IENs in Canada. Mary indicated that due to the challenges she experienced locating CELBAN resources, she chose to pay for a preparation course that was not endorsed by The CELBAN Centre. Mary explained that the course she took was not reflective of the exam structure or questions.

Comments from test-takers at both the Hamilton and Toronto assessment centres highlighted the challenges of administering long, complex, paper and pencil assessments, especially for provisioning test-taking supplies. For example, as a paper and pencil test, the CELBAN featured a question booklet and an answer booklet. Some test-takers described frustration at not being allowed to take notes on their question booklet to guide their responses in their answer booklets:

> We had … answer booklet and question booklet … the thing that I didn't like about this listening exam was they … didn't allow us to write anything on the question paper. Because there is a video going on, two people are having a conversation – and you are doing a multitasking skill in this listening … we're looking at the screen … you're listening what they're talking, and you're also reading the questions. And the question had all the options and the answer booklet had where we need to mark the answers right. (Mary)

Mary highlighted an important point in regard to the structure of the exam. Several tasks, such as the listening task she described are integrated tasks – requiring the test-takers to listen, observe, read questions and indicate a response. These tasks require the test-takers to perform several tasks at once or multitask.

As with many other types of high-stakes assessment, test-takers are given very strict instructions for what they can or cannot bring into the exam room with them. Some test-takers spoke about the challenges they experienced with writing down their responses, not only due to the limited space for note-taking but also due to the restriction on the number of pencils permitted to each test-taker. Penelope explained "But they didn't give the pencils – they only give one pencil, … and whenever we need[ed a] pencil, … we have to raise the hand so that takes time and that distract[s] other people as well right. So at least they [should] have to give two pencils and one sharpener" (Penelope). In Penelope's view, the provision of one more pre-sharpened pencil would help to reduce the downtime experienced by a test-taker whose lead has broken in their pencil, as well as the possible disruption to other test-takers as they await the receipt of another pencil with which to complete the test.

Research Question 4: Do IENs feel the language tasks are authentic?

Test-takers were asked whether they believed the assessment tasks and questions reflected the types of communicative tasks required of a nurse. These types of questions were used to learn more about test-taker accounts on the authenticity of the language used, and assessed, by the CELBAN. Some participants observed that listening activities included recordings, which also captured background noises commonly heard in a nursing setting. This was described as similar to what often occurs in a clinical setting when conversing with patients. For example, Mary explained, "… you need to be good with your communication, [with] your listening because this is exactly what happens in a nursing setup. You're listening, there's a lot of noise going on … you cannot assume things." The inclusion of background sounds in the listening assessment was noted by some participants as a point of contrast to other types of English language tests they had completed in the past, for example, the IELTS. As Joe explained, "… the clarity of the recordings [of the CELBAN] compared to IELTS. … For IELTS … maybe the questions are difficult but the clarity of the questions or the speaking qualities – come more than compared to CELBAN." On this point, I felt it was important to probe for additional information on the type of clarity that was noted by Joe. At this point in the focus group I ("I-" the interviewer), asked the following:

> I- Ummhmm so can you tell more about that, 'the clarity' – do you mean the kinds of voices you heard?
> Joe- No it's like there are lots of other noises are there in the listening
> I- In the recording
> Joe- Ya, in the recordings. I think they're recording from the life situations.
> I-Ya
> Joe- Not like for the other exams
> I- Why do you think they did that?
> Joe- We can hear … some background noises that happen around that situation, in the general atmosphere. Or same like if you are recording something out from here, there are some noises of [a] teacher, or the movements of the agendas, same [as] we can hear in the CELBAN exam too.

The observation made by Joe that background noises seem more likely to be included in the CELBAN listening section as opposed to other general assessments of English language skill, in conjunction with Mary's remarks that in a nursing setting "there's a lot of noise going on" suggests that the listening section of the CELBAN exam more authentically captures the demands placed on the listening skills of a nurse while working on the job. Although background noise may be considered a source of score variance in a general test of English listening skills, it is interesting to note that within a nursing-specific context, it is described by some participants as an assumed reality of the challenges of job performance.

When asked about how realistic the test was compared to working as a nurse, many test-takers responded with comments on the oral component of language assessment. This part of the assessment asks test-takers to participate in a role-play exercise in which they gather patient history and offer discharge instructions to a patient. Although many test-takers recognized this component's attempt to incorporate on-the-job types of tasks into the exam, the setting also introduced some limitations. For example, Jake appreciated the acted responses of the assessors, but also indicated the tension that was created between the test environment and the approximation of the work environment:

> … it was a little bit hard to end it. You know like in a real situation there's a procedure on how you can end a conversation of the nurse-patient interaction-but in that form [the role-play] I think it was nothing like that. So, I had to … figure out, what, how can I conclude here? … I was not sure how to- where do I send the patient? (Jake).

In this case, Jake recognizes the value of structured role-play and is willing to act along with it, but the value of the exercise is undermined when the structure of the role-play misses a procedural step that would follow from the conversation; there was no location to send the patient to. It was Jake, in fact, who would shortly be required to leave the room, but in the typical work scenario, Jake would be tasked with sending a patient either to another area of the hospital, a pharmacy, or an alternate location.

Discussions

Participants in this study often indicated their preference for the CELBAN as an assessment which they viewed as a more appropriate assessment of their professional language skills. This preference was consistent with the original vision of the CELBAN test developers to establish an assessment that "stakeholders believe … measures what it is supposed to measure" (Epp & Lewis, 2009, p.296). Although, some IENs like Penelope, indicated how the wider licensure process might undercut the face validity of the CELBAN. Where the sequence of assessments may suggest sufficient professional nursing knowledge (assessed in English) but insufficient English language knowledge (assessed through the approximation of professional nursing tasks), IENs may be left questioning the validity of the construct representation of the assessments that comprise their wider licensure evaluation.

As of November 2019, there are seven CELBAN test locations across Canada, with future dates posted as far into the future as November 2020 for some locations (The

CELBAN Centre, 2019a). This is an impressive expansion from the two CELBAN test locations in the province of Ontario (Toronto and Hamilton) that served the needs of IENs in the winter of 2015. While the need to travel to limited and for some, distant, locations has been mitigated, as a result of these positive changes, the distraction, and mental exhaustion associated with that type of travel prior to the start of a high-stakes large-scale assessment remains a potential source of construct-irrelevant variance for some. Messick (1991) suggests that this type of variance occurs when the assessment is "too broad containing excess reliable variance that is irrelevant to the interpreted construct" (p. 14). In the case of the CELBAN test-taking experience for those who live in remote areas, their assessment experience risks the inclusion of irrelevant factors, such as staying awake and alert through the assessment despite travelling hundreds of kilometres prior to the beginning their test at the nearest assessment centre.

Since 2015, the CELBAN Centre has published four, new Practice Handbooks (The CELBAN Centre, 2019b). However, in the absence of official CELBAN preparation courses, enterprising individuals will continue to advertise themselves as experts in CELBAN preparation for a significant fee, despite the poorly understood, and inaccurate information presented in these courses, as described by Mary. This risk represents a negative washback effect that Messick (1996) describes by saying: "… if the test employs unfamiliar item formats … to the detriment of communicative competence, teachers might pay undue attention to overcoming the irrelevant difficulty as opposed to fostering communicative proficiency" (p. 14). The official CELBAN handbooks offer independent study aids that help to clarify the test item format for each skill area, an option consistent with Messick's suggestion that a defence against this negative washback effect is to "provide test familiarization and preparation materials to reduce the effects of construct irrelevant difficulty and attendant test anxiety, but the best defence is to minimize such irrelevant difficulty in the first place … (p.14)." Since the time of data collection, there are now more resources to help test-takers become familiar with the structure and skills that must be demonstrated in each section of the CELBAN assessment. However, in the winter of 2015, test-takers like Rob and Sandy described confusion regarding how best to respond to some test item formats specifically with regard to the decision to prioritize general communicative competence over, specific nursing knowledge. A point acknowledged by Rupp et al. (2006) who argue that where test-takers respond to a reading comprehension question in a manner inconsistent with their thoughts about the passage at the time of their reading, their response is dictated not necessarily by their understanding, but by their understanding of the purpose of the question itself. In particular, what this piece of information may call into question is the potential for test-takers to attempt to problem solve beyond the construct of reading comprehension in order to select answers that they feel are more representative of their professional knowledge This test-taker feedback emphasizes the important role test-taker preparation materials play in preparing IENs to demonstrate the best of their abilities for a given task.

DeLuca et al. (2013) have noted, "the difference between the student's true score and his/her observed score on the test as attributed to a variety of factors" (DeLuca et al., 2013, p. 663). Factors affecting score variance impact our understanding of validity, where it is defined as, "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (Haladyna & Dowling, 2004, p. 18). In other words, our choice of what is measured as a factor of score variance, and what is not, can

affect the claims made about the degree to which theory and evidence concur, or validate the testing instrument. High-stakes testing situations demand increased evidence of validity as the outcomes which follow the interpretation of a given test score can be life-altering. Reducing or eliminating threats to validity establishes greater confidence in the interpretation of test scores (Haldyna & Dowling, 2004). While the CELBAN continues to evolve (the separate question and answer booklet described by Mary, has now been combined into one workbook), consideration might be given to future online versions of the CELBAN that might eliminate some of the administrative challenges noted by some test-takers regarding access to physical testing centers, pencils, and scrap paper.

Conclusion

This study represents the qualitative findings of focus group discussions and individual interviews with sixteen participants at two CELBAN testing locations in Ontario in the winter of 2015. Future research and administration of the CELBAN might consider whether the accounts of test constructs, conditions, and administration described in this study are shared by test-takers in other test locations and provinces since the CELBAN Test Renewal Project was launched in 2014. A survey of a large sample of CELBAN test-takers may also help to determine if the results of this study are generalizable to a wider population, particularly in regard to the reported CELBAN administration and preparation experience.

The data for this study were coded primarily by one researcher in consultation with the guiding framework. The deductive codes that were applied emerged from the literature on large-scale, high-stakes English language assessment and inductive codes were created as the data was reviewed. A subset of the data were reviewed by another researcher, and their coding structure was compared with that of the primary researcher. This process helped to establish inter-rater reliability, and encouraged the reduction of emergent codes to key differences in test-taker experiences and descriptions. Future research might benefit from collaboration with nursing professionals to review the accounts of test constructs, as well as the role that professional knowledge plays in demonstrating language competency in an assessment presented in the format of profession-specific scenarios and contexts.

The CNO has estimated an attrition rate of 40% among Registered Nurses educated abroad who apply for membership (Blythe et al., 2009). This estimated loss of IENs to the profession suggests a critical need for further research on how to best support federally recruited IENs in their pursuit of re-certification and re-licensure with the CNO, to facilitate economic integration in Canada. The CELBAN is intended to facilitate integration by recognizing the professional knowledge of IENs obtained in their home countries, and thus focuses language testing specifically on the language of the profession (Epp & Lewis, 2009). As a national testing instrument of professional language competency, the CELBAN judges what constitutes language proficiency based on the receptive and productive language skills demonstrated by the test-takers on their given test date. This type of approach to language skill assessment concurs with Kern's (2004) critique on instructional approaches that organize language curriculum by having students work through a sequence of phrases, sentences, paragraphs, and finally extended discourse. Kern acknowledges that these linguistic elements may be "eminently logical, but do not mesh well with the psychological needs of language learners who strive to communicate in

meaningful, whole acts" (p. 10). Recognition of the motivational and psychological factors which may impact the economic integration of Internationally Educated Professionals (IEPs) is particularly important for Canada, where research has reported only 19-40% of IEPs obtain suitable employment in their professional field (Ngo & Este, 2006).

While this study is exploratory in nature, it offers a closer look at IEN test-taker accounts of the CELBAN, which contributes to our understanding of the meaning and inferences we make from its assessment score (DeLuca et al. 2013; Messick, 1996). Additional CELBAN research is needed to continue to support IENs in their attainment of licensure to practice, and most importantly, to support the national vision for economic and social integration of Internationally Educated Professionals into our Canadian communities of practice.

Correspondence should be addressed to Stefanie Baldwin.
Email: stefanie.bojarski@d2l.com

Notes

[1] This paper is based on Baldwin-Bojarski, S. (2016). *Proving my competency one test at a time: Internationally educated nurses and the Canadian English language benchmark assessment for nurses.* [Unpublished master's thesis, Queen's University]. https://qspace.library.queensu.ca/handle/1974/14926?show=full

References

Baldwin-Bojarski, S. (2016). *Proving my competency one test at a time: Internationally educated nurses and the Canadian English language benchmark assessment for nurses.* [Unpublished master's thesis, Queen's University]. https://qspace.library.queensu.ca/handle/1974/14926?show=full

Blythe, J., Baumann, A., Rheaume, A., & McIntosh, K. (2009). Nurse migration to Canada: Pathways and pitfalls of workforce integration. *Journal of Transcultural Nursing. 20*(2), 202-210.

CELBAN. (n.d.a). *CELBAN Overview.* http://www.celban.org/celban/display_page.asp?page_id=3

CELBAN. (n.d.b) *What if I only failed one component of the CELBAN?* http://www.celban.org/celban/display_page.asp?page_id=87

Centre for Canadian Language Benchmarks. (2014). *Test Taking Strategies.* http://celbancentre.ca/celban/media/Celban/CELBAN_TestStrat-Dec-18,-2015.pdf

Cheng, L. (2014). Consequences, impact, and washback. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1130–1146). John Wiley & Sons.

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation use. *Educational assessment, 16,* 104-122.

Cheng, L., Spaling, M., & Song, X. (2013). Barriers and facilitators to professional licensure and certification testing in Canada: Perspectives of Internationally Educated Professionals. *Journal of international migration and integration, 14*, 733-750.

College of Nurses of Ontario (CNO). (2013). *Accepted language proficiency tests*. http://www.cno.org/en/become-a-nurse/new-applicants1/accepted-language-proficiency-tests/

College of Nurses of Ontario (CNO). (2015) *In depth: Language proficiency*. www.cno.org/en/become-a-nurse/registration-requirements/language-proficiency/in-depth-language-proficiency/

DeLuca, C., Cheng, L., Fox, J., Doe, C., & Li, M. (2013). Putting testing researchers to the test: An exploratory study of the TOEFL iBT. *System, 41*, 663-676.

Epp, L. & Lewis, C. (2009). Innovation in language proficiency assessment: The Canadian English Language Benchmark for Nurses (CELBAN) In Susan Danridge Bosher & Margaret Dexheimer Pharris (Eds.), *Transforming Nursing Education: The Culturally Inclusive Environment* (pp. 285-306). Springer

Fox, J. & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education: Principles, Policy & Practice, 14*(1), 9-26.

Fox, J., Cheng, L., & Zumbo, B. D. (2013). Do they make a difference? The impact of English language programs on second language students in Canadian universities. *TESOL Quarterly, 48*(1), 1-29.

Grabowski, K. C., & Dakin, J. W. (2014). Test development literacy. In A. J. Kunnan (Ed.), Companion to language assessment (pp. 869-889). Wiley-Blackwell.

Haertel, E. H. (2013). Getting the help we need. *Journal of Educational Measurement, 50*(1), 84–90.

Haladyna, T. M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues & Practice, 23* (1), 17-27.

Kern, R.K. (2004). Literacy and advanced foreign language learning: Rethinking the curriculum. In Byrnes, H., & Maxim, H.H. (eds) Advanced foreign language learning: A challenge to college programs. (pp. 2-18). Thomson Heinle.

Koch, M., & DeLuca, C. (2011). Narrative case description: An approach to validation in the context of high-stakes assessments. *Assessment in Education: Principles, Policy & Practice, Special Issue: High-stakes Assessments, 19*(1), 99–116.

Krueger, R.A., & Casey, M.A. (2000). Focus groups (3rd ed.). Thousand Oaks, CA: Sage.

Lewis, C., & Kingdon, B. (2016). CELBAN: a 10-year retrospective. *TESL Canada Journal, 33(2),* 69-82.

Malone, M. E., & Montee, M. (2014). Stakeholders' beliefs about the TOEFL iBT test as a measure of academic language ability. *ETS Research Report Series, 2012*(2), 1-51.

McMillan, J., & Schumacher, S. (2010) Research in education: Evidence-based inquiry (7th ed.). Pearson.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher. 31(7)* 28-38.

Messick, S. (1991). Validity of test interpretation and use. In M.C. Alkin (Ed), *Encyclopedia of Educational Research* (6th ed.) (pp.1-29). Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(3), 241-256.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162.

Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. Canadian Journal of Education, 25, 262-279.

Ngo, H.V., & Este, D. (2006). Professional re-entry for foreign trained immigrants. *Journal of international migration and integration.*7, 27-50.

Patton, M.Q. (2002). Qualitative evaluation and research methods (3rd ed.). Sage.

Polkinghorne, D.E. (2005). Language and meaning: Data collection in qualitative research. *Journal of Counseling Psychology, 52(2),* 137-145.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23(4)*, 441-474.

The CELBAN Centre. (2018). *FAQ.* http://www.celbancentre.ca/FAQ1.aspx

The CELBAN Centre. (2019a). *Test locations and dates.* http://www.celbancentre.ca/register/test-locations-and-dates.aspx

The CELBAN Centre (2019b). *CELBAN practice handbooks.* http://www.celbancentre.ca/prepare/Products.aspx

Touchstone Institute. (2019a). *Test information manual.* http://www.celbancentre.ca/celban/media/Celban/CELBAN-Test-Information-Manual-Final.pdf

Touchstone Institute. (2019b). Writing test renewal. *Facts and Figures, 6*, 1-6.