RECOMMENDING LANGUAGE TEST SCORES FOR WORKPLACE READINESS

Who recommends higher language test score for workplace readiness — Language specialists or employers?

By Andrea Strachan, Touchstone Institute: Competency Assessment Experts

Abstract

Regulatory bodies are mandated through the Ontario Regulated Health Professions Act to certify only practitioners who demonstrate the knowledge and skills to practice in a safe and effective manner. These standards protect the public by ensuring that only those applicants who meet the prescribed minimum standard are able to gain employment and serve the public. Internationally educated health professionals, although previously certified in other jurisdictions, must complete professional practice exams to demonstrate their professional competence and often they must also provide scores on an acceptable language proficiency exam to demonstrate their communicative competence. Scores on international standardized language proficiency tests are used in this context. This investigation sought to determine whether cut score recommendations differed between health professionals and language specialists acting as expert panellists in a standard-setting study. Cut score recommendations by each of these expert groups were collected using a standard-setting instrument and the cross-group discussion periods between judgment rounds were recorded and analysed. Results showed that language specialists recommended higher cut scores than health professionals. The transcript analysis indicated that language specialists contributed information about language testing that helped the panel understand the language testing process, and health professionals contributed workplace examples that helped the panel understand the language demands of the workplace.

Keywords: standardized language tests, cut-scores, standard-setting

Go Back to Menu

Successful immigrants contribute substantially towards strong labour-force growth, which in turn enhances Ontario's potential economic growth rate (Drummond Report, 2012). However, there is serious concern about the underutilization of immigrant skills (Alboim, Finnie, & Meng, 2005). Many policy recommendations have been made for improved credential-recognition programs (Alboim & Cohl, 2007) but challenges faced by immigrants along the professional registration and licensure pathway continue to be documented (Baumann & Blythe, 2009; Johnson & Bauman, 2011; Cheng, Spaling, & Song, 2013). To address potential unfair practices in this area the Fair Access to Regulated Professions Act of 2006 (FARPA) promotes transparent, objective, impartial, and fair registration practices with the intent to positively impact the experience of international applicants for professional registration in Ontario. Scores on international standardized language proficiency tests are accepted as evidence of language ability, and cut scores (the passing threshold) must be defensible and in keeping with Ontario's fair-access law.

Setting Standards for Language Proficiency

Regulators have had language proficiency policies since the Ontario Regulated Health Professions Act (RHPA) was enacted in 1991. The RHPA is the legislation that governs Ontario's health professional regulators (i.e., regulatory colleges). These colleges determine each profession's scope of practice, define what is allowed and disallowed (controlled acts), and establish entry-to-practice licensure, certification of registration requirements. This includes setting exams and required scores.

Language Proficiency Standards in the Legislation and Regulatory Policy

Internationally educated professionals who apply for professional registration or licensure in Canada often must meet a language-proficiency standard. These standards may be defined in the regulation, in regulatory policies, or through by-laws, with wording and interpretation that differ across the professions. In medicine, the English-language requirements are defined in the Medicine Act itself: "an applicant is reasonably fluent in English or French if the applicant, obtains a score of 220 on the Test of Spoken English and a score of 580 on the Test of English as a Foreign Language of the Educational Testing Service" (Medicine Act, 1991 O. Reg. 93/12 ONTARIO REGULATION 865/93 REGISTRATION). In contrast, the legislated requirement for Medical Laboratory Technologists is much more general: "The applicant must have reasonable fluency in either English or French" (Medical Laboratory Technology Act, 1991, S.O. 1991, c. 28). Similarly, the more recent Traditional Chinese Medicine Act, 2006, requires that "the applicant must be able to speak, read and write either English or French with reasonable fluency" (Traditional Chinese Medicine Act, 2006 O. Reg. 27/13 ONTARIO REGULATION 27/13 REGISTRATION).

The requirements described above demonstrate some of the challenges regulators face with language proficiency standards. One issue is the need for language proficiency standards to be dynamic rather than static given the changing nature of testing. For example, the Test of English as Foreign Language (TOEFL) has changed significantly since 1991. Its

Go Back to Menu

speaking component, the Test of Spoken English (TSE), has been discontinued and the "580" is a score that refers to the paper-based TOEFL, which was replaced in 2006 by the Internet-Based TOEFL. On the other hand, the Traditional Chinese Medicine and Medical Laboratory Technology acts are very general and do not describe the requirements in a way that defines how language proficiency is to be measured.

The Question of Defensible Language Proficiency Standards

The 2006 Fair Access to Regulated Professions Act (FARPA) imposed oversight of regulatory bodies to ensure that registration practices were transparent, objective, impartial and fair. The question of whether existing language standards were appropriate was raised around this time by regulators and employers who suspected that the language difficulties experienced by applicants on qualifying examination and in the workplace, even after language proficiency standards had been met, were based on the weakness of the standard itself. Regulators, being familiar with standard setting as a reliable method for setting cut scores on competency exams, began to consider this approach for language assessments.

Standard Setting as a Defensible Method to Set Standards

Cut scores on professional competency examinations are commonly determined through standard setting. Standard setting is a psychometric process used as a policy-making tool that is legally defensible and can be associated with the requirements of the target situation (Cizek, 1996, 2001; Johnson, Squires, & Whitney, 2002; Kane, 1994). Standard-setting studies for educational standards or professional competency exams are commonplace and represent the acceptable practice in educational and competency measurement and evaluation (Popham, 1992; Plake, Impara, & Potenza, 1994; Smee & Blackmore, 2001; Skorupsky & Hambleton, 2005). The most widely used standard-setting process relies on the judgment of subject matter experts (expert panellists or judges) who analyse test items and recommend a point on the score scale that, in their view, represents the threshold that distinguishes between examinees who are competent and those who are not (Cizek, 1996; Zeiky, 2001).

Norcini and Shea (1997) describe two types of validity evidence required to support credible and valid standards resulting from a standard-setting process. One is that the composition of the panel must demonstrate the appropriate qualifications, numbers, and variety, and the methods used must be supported by research. The credibility of a standard is dependent on the quality and performance of panellists who represent the communities of interest (Norcini & Shea, 1997, 2002; Raymond & Reid, 2001). The criteria established for the panel include: subject matter expertise, an understanding of the examinee population, the ability to estimate item difficulty, knowledge of the instructional environment, appreciation of the consequences, and representation of the communities of interest. It could be argued, therefore, that language specialists should be included in the panel in order to support greater validity of the cut scores on international standardized language proficiency tests are accepted evidence of language ability. Standard-setting studies for language proficiency tests have appeared in the academic literature (O'Neill, Tannenbaum, & Tiffen, 2005; O'Neill, Buckendahl, Plake, & Taylor, 2007), and although these studies demonstrate a reasonable variety of panellists, including locally and internationally trained nurses representing 16 different languages, they exclude language specialists.

Methods

This study investigated cut score recommendations made by health professionals and language specialists who participated as expert panellists in a standard-setting study. The study was designed to evaluate the impact of the participations of these diverse panellists. An existing methodology developed by Educational Testing Services was utilised as a framework for the standard-setting activity.

Standard-Setting Procedures

The TOEFL®iBT Standard Setting Manual was developed by Educational Testing Service in 2006 to coincide with the launch of the Internet-based TOEFL (TOEFL®IBT). The Manual provides a validated and adapted set of tools to run a modified Angoff standard-setting study and adheres to the prescribed methodology such as the summary of the accepted practice and the general steps provided by Hambleton (2001):

1. Choose a large and representative panel;
2. Choose a standard-setting method;
3. Train panellists to use the method;
4. Compile and analyse item ratings;
5. Conduct a panel discussion on the proposed ratings;
6. Compile item ratings a second time;
7. Compile final ratings;
8. Present consequences to the panel;
9. Revise ratings if necessary; and
10. Compile validity evidence.

The TOEFL®iBT Standard Setting Manual is a structured package that includes materials and instructions for execution of all these steps (Educational Testing Services, 2006).

Research Design

A mixed methods design was applied (Onwuegbuzie & Teddlie, 2003) in order to facilitate analysis of both the qualitative and quantitative data collected in the standard-setting session. The quantitative analysis focused on differences in voting patterns of occupational experts and language specialists while the qualitative data was found in the transcription of the discussion, thus helping with the identification of factors that might have influenced the decision making process of the two groups. A concurrent triangulation approach was applied to the data analysis (Creswell & Plano Clark, 2007). This allowed for a separate collection of quantitative and qualitative data during the same timeframe throughout the standard-setting session. Afterwards, the data was analysed and merged in an interpretive stage.

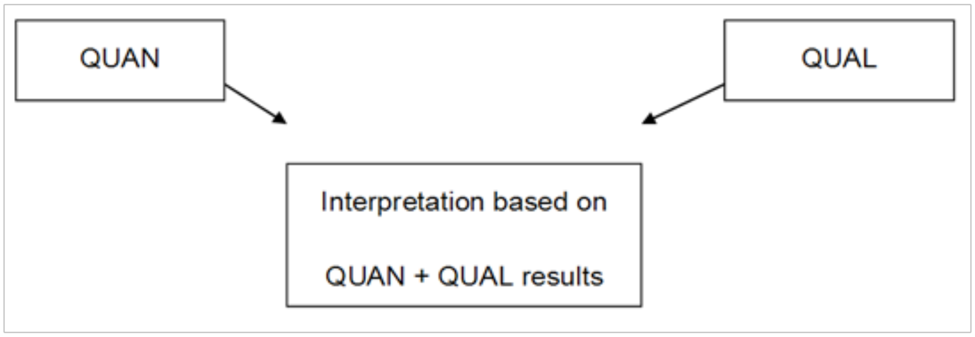

Figure 1. A visual model of a triangulation mixed-method design.

Study Participants (Panellists)

Fifteen panellists were selected through a purposeful sampling procedure. They represented the communities of stakeholders that met the requirements recommended by Raymond and Reid (2001) and included: two employers, two ESL specialists, two patient advocates, one trainer, three internationally educated medical technologists, three Canadian trained technologists, one other health care professional, and one representative of the regulatory board. All panellists met the qualifying criteria established Raymond & Reid (2001) and demonstrated subject matter expertise, an understanding of the examinee population, the ability to estimate item difficulty, knowledge of the instructional environment, appreciation of the consequences, and representation of the communities of interest.

Data Collection

The standard-setting instrument included in the TOEFL®iBT Standard Setting Manual (Educational Testing Service, 2006) was used as the data collection instrument. Quantitative data was collected and recorded in pre-programmed Microsoft Excel tally sheets provided within the manual. The qualitative data comprised transcriptions of the recordings of the four between-round discussion periods, each for one section of the exams (speaking, writing, reading, and listening).

The quantitative data contained the first and final cut score recommendations by each panellist for each of the skill areas tested (speaking, listening, reading, writing). Cut score recommendations from each voting session were recorded in the tally sheets. Pre-programmed embedded formulas then generated descriptive statistics (minimum, maximum. mean, median, and standard deviation) and also converted raw score recommendations into a scaled score which resulted in a final overall score.

The recordings of four between-round discussion periods were recorded and transcribed, each representing the discussion related to the cut score of a specific component of the test (speaking, writing, reading, and listening). Data preparation and analysis were modelled on a grounded-theory analysis process documented by Harry, Sturges, and Klingner (2005) which include the following phases:

1. assign open coding to key points, recurring ideas and concerns that emerged more than once throughout the discussion
2. assign conceptual categories
3. develop themes.

Study Results

Data Analysis

The quantitative results will be presented first, with specific results by test section, followed by a section outlining the qualitative data.

Quantitative data

The table below illustrates the data collected during the rounds of voting. Round 1 represents the aggregate results from the individuals in the two groups: the English as a Subsequent Language (ESL) specialists and the medical technology group (Techs). The final round represents the aggregate votes after the group discussion which emerged from a review of the first recommended score. Differences were observed between the groups for each test section (speaking, writing, listening, and reading). In most test sections, the ESL group often recommended higher cut scores than the Techs group. The notable exception is the speaking test section where recommended cut-scores were highest for both groups, and indeed closer between the groups. Despite these differences, both groups' cut scores represented the intermediate levels as described by TOEFL®iBT test score descriptors.

Table 1

Scaled Score Recommendations of Language Specialists and Content Experts

TEST SECTION		G1 (ESL) SCALED SCORE OUT OF 30	G2 (TECHS) SCALED SCORE OUT OF 30	DIFFERENCE (G1 - G2)
Speaking				
	Round 1	23	24	1
	Final round	24	24	0
	Change	+1	0	
Writing				
	Round 1	22	21	1
	Final Round	21	18	3
	Change	-1	-3	
Listening				
	Round 1	17	15	2
	Final round	17	15	2
	Change	0	0	
Reading				
	Round 1	19	17	2
	Final round	19	18	1
	Change	0	+1	

Notes. Round results are reported as scaled scores out of 30.

The overall results represent the final overall recommended cut scores. The overall final cut score recommendation made by the ESL group was 81 out of a total of 120 whereas the recommendation from medical technologists was 76 out of a total of 120. The ESL group's overall recommendation is noticeably higher than the Techs.

Go Back to Menu

Table 2

Overall Scaled Score Recommendations of Language Specialists and Content Experts

ROUNDS	G1 (ESL) SCALED SCORE OUT OF 120	G2 (TECHS) SCALED SCORE OUT OF 120	DIFFERENCE (G1 - G2)
Round 1	81	76	5
Final Round	81	75	6
Change	0	-1	
Change	0	-1	

Notes. Round results are reported as a scaled score out of 120.

Qualitative data

The qualitative data analysis included the assignment of open coding to key points, recurring ideas and concerns that emerged more than once throughout the discussion. These categories were then developed into themes. Seven main themes were identified in the discussion. Of these, two were areas of expertise for the language specialists (Language & Language test), two were areas of expertise for the regulators (Credentialing & Candidate experience), two related to the workplace (Patient care & Professionalism) and one was related to the standard-setting process. These themes of discussion represented the specialty areas of participating panellists. The continuing discussion on the standard-setting method itself indicates that panellists and the facilitator were continually ensuring that the established protocol was being followed appropriately.

Table 3

Codes and Categories Observed in the Discussion Transcripts

THEMES	CATEGORY DESCRIPTION	OPEN CODES
1. Language	Synonyms, near lexical equivalents, phrases, or paraphrases that were inferences, interpretations or summarized meanings related to language use and language proficiency	Vocabulary, comprehensibility, tone, accuracy, grammar, overall meaning, confidence, inference, coherence, factual details, fluency, pronunciation, punctuation, spelling
2. Credentialing	Synonyms, near lexical equivalents, phrases, or paraphrases that were inferences, interpretations or summarized meanings related to the credentialing process	Fairness (requirements for internationally trained versus Canadian trained technologists), cross-cultural awareness of the need for the process (different from other countries), and the credentialing process itself (tests, time, cost), purpose of language testing

Go Back to Menu

THEMES	CATEGORY DESCRIPTION	OPEN CODES
3. Standard setting	Synonyms, near lexical equivalents, phrases, or paraphrases that were inferences, interpretations or summarized meanings related to the standard-setting process	Standard-setting process (clarification or reminders of the protocols), definition of the minimally competent practitioner
4. Patient care	Synonyms, near lexical equivalents, phrases, or paraphrases that were inferences, interpretations or summarized meanings related to patient safety and needs, as well as the level of confidence of the patient about the treatment being received	Patient safety, patient needs, patient confidence
5. Professionalism	Synonyms, near lexical equivalents, phrases, or paraphrases that were inferences, interpretations or summarized meanings related to professional standards, workplace tasks and performance, concerns about human resource shortages, legal repercussions, clinical errors, interpersonal skills appropriate to the workplace, technical skills required at work.	Professional standards, workplace tasks, human resource shortages, legal repercussions of workplace performance, clinical errors, interpersonal skills, technical skills
6. Language test	Synonyms, near lexical equivalents, phrases, or paraphrases that were inferences, interpretations or summarized meanings related to the language test (TOEFL®iBT)	Test relevance/irrelevance, test purpose, critic of test design, description of test design, passages in the, test taking strategies, testing condition, test administration protocols, test location, and test in a computer lab
7. Candidate experience	Synonyms, near lexical equivalents, phrases, or paraphrases that were inferences, interpretations or summarized meanings related to the candidate experience or perspective	Personal experience, advocacy, sympathy, impact, and consequences of the cut score

Conclusions

Through their language standards, regulators make decisions about the level of language skills required for the workplace. Their duty is to protect the safety of the public by ensuring that certified and registered professionals are capable and effective. Language skills play a critical role in the provision of safe and effective health services as they support the specialized communication skills which facilitate quality care. Canada's client-centred approach requires health practitioners to understand patients' needs and desires for their health care and wellbeing, as well as to communicate with patients and their families about shared health care goals and priorities. Good communication also underlies effective teamwork that enables efficient delivery of inter-professional health care services. Furthermore, sound communication skills help health practitioners meet legal and ethical requirements related to documenting patient interviews, assessments, care plans, and treatment outcomes. In the case of IEHPs, there is an assumption that a minimum English language proficiency standard will enable these professional communication competencies.

This study showed that expertise of each group was a contributing factor to the discussion in this standard-setting study. Language specialists supplied information about language testing that helped the panel understand the language testing process while medical technologists added workplace examples that helped the panel understand the language demands of the workplace. Indeed, a language standard that requires cut scores for speaking, listening, reading, and writing skills can be defensibly linked to safe professional communication only if the perspectives of both language specialists and medical technologists considered.

While some similarities were observed in the cut score recommendations made by these two different groups, there were also some similarities. Both groups recommended the highest cut-scores for speaking. This was indeed the area where recommendations were the closest between the groups, demonstrating a shared understanding of the critical importance of spoken communication in the health care setting. The overall cut score recommendation did, however, differ. The ESL group recommended a total score of 81 whereas the Techs recommended a total score of 76 out of a possible total of 120. Unlike other tests, the TOEFL does not provide bands of proficiency, but analysis of the results in comparison to the scores on the International English Language Testing System (IELTS) shows that 81 represents IELTS 6.5, whereas 76 represents IELTS 6.0. Although the TOEFL and the IELTS language proficiency scales are different scales, they are often compared because regulators commonly accept both tests as evidence of language ability. This difference is noteworthy because many health care regulators require 6.5 as a minimum standard. This could indeed make the difference between success and failure in gaining a license of becoming certified. What is interesting about this is that the ESL professionals are the group recommending the higher cut score, yet the professionals are arguably better at defining the needs of the workplace.

Go Back to Menu

References

Alboim, N. & Cohl, K. (2007). *Shaping the future: Canada's rapidly changing immigration policies*. Toronto, ON: Maytree Foundation.

Alboim, N., Finnie, R., & Meng, R. (2005). The discounting of immigrants' skills in Canada: Evidence and policy recommendations. *IRPP Choices, 11*(2).

Baumann, A. & Blythe, J. (2009). *Integrating internationally educated health professionals into the Ontario workforce (2009).* Toronto, ON: Nursing Health Services Research Unit, McMaster University for the Ontario Hospital Association.

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*(1), 733–750. doi: 10.1007/s12134-012-0263-3.

Cizek, G. J. (1996). Setting passing scores. *Educational Measurement: Issues and Practice, 15*(1), 12–21.

Cizek, G. J. & Bunch M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. London: Sage Publications.

Commission on the Reform of Ontario's Public Service (2012). *Public Services for Ontarians: A Path to Sustainability and Excellence*. Toronto, ON: Queen's Printer for Ontario, 2012.

Creswell, J. W. & Plano Clark, V. L. (2007). *Designing and conducting mixed methods research*. London. Sage Publications.

Educational Testing Service (2006). *TOEFL® iBT standard setting manual*. Princeton, NJ: Educational Testing Service.

Fair Access to Regulated Professions and Compulsory Trades Act, 2006, S.O. 2006, c. 31. Retrieved from: https://www.ontario.ca/laws/statute/06f31?search=Fair+Access+to+Professions+Act

Ferdous, A. & Plake, B. (2005). Understanding the factors that influence the decisions of panelists in a standard-setting study. *Applied Measurement in Education, 18*(3), 257–267.

Harry, B., Sturges, K. M., & Klingner, J. K. (2005). Mapping the process: An exemplar of process and challenge in grounded theory analysis. *Educational Researcher, 34*(2), 3–13.

Johnson, K. & Baumal, B (2011). *Assessing the workforce integration of internationally educated health professionals*. Hamilton, ON: Canadian Society for Medical Laboratory Sciences

Johnson, R., Squires, J., & Whitney, D. (2002). Setting the standard for passing professional certification examinations. *Financial Management Online,* April 22, 1–12.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Medical Laboratory Technology Act, 1991, S.O. 1991, c. 28. Retrieved from: https://www.ontario.ca/laws/regulation/940207

Norcini, J. & Shea, J. (1997). The credibility and comparability of standards. *Applied Measurement in Education, 10*(1), 39–59.

Norcini, J. & Shea, J. (2002). The reproducibility of standards over groups and occasions. *Applied Measurement in Education, 5*(1), 63–71.

O'Neill, T. R., Buckendahl, C. W., Plake, B. S., & Taylor, L. (2007). Recommending a nursing-specific passing standard for the IELTS examination. *Language Assessment Quarterly, 4*(4), 295–317.

O'Neill, T. R., Tannenbaum, R. J., & Tiffen, J. (2005). Recommending a minimum English language proficiency standard for entry-level nursing. *Journal of Nursing Measurement, 13*(2), 129–145.

Ontario Fairness Commission. (2006). Fair Access to Regulated Professions and Compulsory Trades Act, 2006.

Plake, B. S., Impara, J. C., & Potenza, M. T. (1994). Content specificity of expert judgments in a standard setting study. *Journal of Educational Measurement, 31*(4), 339–347.

Popham, W. J. (1992). Appropriate expectations for content judgments regarding teacher licensure tests. *Applied Measurement in Education, 5*(4), 285–301.

Raymond, M. R. & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In Cizek, G. (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 119–157). Mahwah, NJ: Lawrence Erlbaum Associates.

Skorupsky, W. & Hambleton, R. (2005). What are panelists thinking when they participate in standard-setting studies? *Applied Measurement in Education, 18*(3), 233–256.

Zeiky, M. J. (2001). So much has changed: how the setting of cut scores has evolved since the 1980s. In Cizek, G. (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 19–51). Mahwah, NJ: Lawrence Erlbaum.

Go Back to Menu