

Setting Language Proficiency Standards for Accreditation in Health Care Professions

*A Guide on How to Plan, Deliver and Evaluate Standard Setting Procedures to
Recommend Cut Scores on Tests of English Language Proficiency for
Professional Accreditation*

Based on the Project

'Language Proficiency Testing for Internationally Educated Medical Laboratory
Technologists: Validating Cut Scores and a New Testing Tool'

Prepared by



March 2009

© 2009 LCRT Consulting and the Canadian Society of Medical Laboratory Sciences (CSMLS)

The information in this How to Guide reflects the experience of the standard setting team for the project *'Language Proficiency Testing for Internationally Educated Medical Laboratory Technologists: Validating Cut Scores and a New Testing Tool'*. Recommendations and resources provided in this guide are offered as a reference only.

Contents

Introduction.....	2
Step 1 ▶ Select a standard setting method	3
What is the recommended practice?.....	3
What tests will be reviewed?.....	3
What method is appropriate for the tests and the standard setting purpose?	4
What references and resources were helpful?	5
Step 2 ▶ Plan the standard setting format and delivery	6
What delivery model will be most effective?	6
Who will plan and administer the sessions?	6
What costs will be involved?	6
How much time will be necessary?	6
What are some specific recommendations for planning and delivery of the sessions?	8
What references and resources were helpful?	9
Step 3 ▶ Select and recruit the panel of experts	10
What will be the composition of the expert panel?.....	10
How will the expert panelists be selected?	11
What references and resources were helpful?	11
Step 4 ▶ Collect and prepare the standard setting materials	12
What materials are needed?.....	12
What additional Information must be researched?	13
Where can the materials be procured?	14
How should they be reproduced?	14
What references and resources were helpful?	14
Step 5 ▶ Evaluate the standard setting process and results.....	15
What is the recommended practice?.....	15
How will the quality of the standard setting process be known?.....	15
How will the reliability and validity of the standard setting results be known?	15
How will the standard setting results be recorded and analyzed?	16
What references and resources were helpful?	17
Step 6 ▶ Set the final cut scores and communicate the results.....	18
What is the recommended practice?.....	18
How are the panel's recommendations communicated to the decision makers?.....	18
What follow up activities are needed?.....	19
What references and resources were helpful?	19
Resource List.....	20

Introduction

The goal of a standard setting procedure is to determine a point on a score scale that distinguishes between candidates who meet a minimum requirement and those who do not. This point is called a cut score. Cizek (2001) defines standard setting as the “task of deriving levels of performance on educational or professional assessments, by which decisions or classifications of persons will be made” (Cizek, p.3).

This How to Guide describes the steps that were taken in preparing and delivering three standard setting sessions through the Canadian Society of Medical Laboratory Sciences (CSMLS). The purpose of these standard setting sessions was to establish English language proficiency cut scores for professional accreditation in medical laboratory sciences. The process involved identifying cut scores on different language proficiency tests that would distinguish between applicants with levels of English language proficiency that are minimally acceptable for safe and effective practice from those who fall below that level.

Three tests were reviewed during this process: the Internet Based TOEFL (TOEFL iBT), the International English Language Testing System – General Training (IELTS GT), and the Michener English Language Assessment (MELA). The same standard setting methodology was applied to all sessions.

In this manual we describe six steps that summarize how the sessions were planned, delivered and evaluated, as follows:

- Step 1 – Select a standard setting method
- Step 2 – Plan the standard setting format and delivery
- Step 3 – Recruit the panel of experts
- Step 4 – Collect and prepare the standard setting materials
- Step 5 – Evaluate the standard setting process and results
- Step 6 – Set the final cut scores and communicate the results

For each of these steps we describe the recommended practice and the approach and procedures that we followed. We also discuss the challenges and lessons learned from the project and offer specific recommendations and references related to each step.

Step 1 ► Select a standard setting method

What is the recommended practice?

Standard setting is a widely used process for establishing cut scores on tests, either for educational or for professional purposes. Standard setting guidelines are outlined in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) as well as in academic literature (Cizek, 2001; Cizek & Bunch 2007). Our project began with a detailed literature review. We noted that numerous methods have been developed and validated since the early days of standard setting (the 1950s), with the most widely used technique being the Angoff method. It is clear that the method must be selected based on the design of the test and the purpose for which the test is being used.

The objective of the CSMLS project was to recommend cut scores on language proficiency tests required as proof of English language proficiency for medical laboratory technologists (MLTs). Specifically, we sought recommendations regarding a performance standard that internationally educated medical laboratory technologists (IEMLTs) must demonstrate in order to become eligible for professional registration in Canada. Within the context of the project, standard setting involved identifying the point on score scales of three different English language proficiency exams that distinguishes between examinees whose language skills fall below a minimally acceptable standard for practice as an MLT and those who meet or surpass that standard.

Keeping in mind that a language proficiency test “aims to establish a candidate’s readiness for a particular communicative role, e.g. in a work or educational setting” (McNamara, 2000, p. 135) we sought a standard setting method that would provide a way to associate the communication demands of the MLT workplace to performance levels on each of the language proficiency tests under consideration.

What tests will be reviewed?

Once we had clearly articulated the purpose for which the standard setting session was being planned, we identified the tests that would be reviewed. This decision was guided by the current CSMLS language standards policy. Several English language proficiency tests had already been adopted as proof of language proficiency for the CSMLS prior learning assessment (PLA) process; two of these were selected for this process. Additionally, CSMLS was interested in exploring a made-in-Ontario, occupation-specific test of language proficiency for medical professionals, so this test was also included. In all, three test were selected:

- The **Internet Based Test of English as a Foreign Language (TOEFL iBT)** is an academic-purpose English language test that measures the test taker’s ability to communicate in English in a college or university context. It is an American test most often used to measure the English language proficiency of foreign students who wish to study in the USA, but regulatory bodies here also use it as a test to provide proof of language proficiency for credentialing purposes. TOEFL iBT is one of the tests that the CSMLS currently accepts as proof of language proficiency for IEMLTs seeking registration.
- The **International English Language Testing System (IELTS)** is a test that measures the test taker’s ability to communicate in English for the purposes of study (IELTS Academic) or work (IELTS General Training) where English is the language of communication. IELTS is a British-Australian test designed primarily for those who wish to work or study in the UK or Australia, but it too is used for a variety of purposes including professional credentialing, employment and immigration in many English speaking countries. It is one of the tests that the CSMLS had adopted prior to this standard setting project as proof of language proficiency for IEMLTs seeking registration. Because Citizenship and Immigration Canada (CIC) accepts the

IELTS General Training version as part of the immigration system, that a particular version of the test was chosen for the standard setting with a view to reduce duplication in testing for the candidates.

- The **Michener English Language Assessment (MELA)** is a test of English language proficiency for allied health professionals seeking admission to upgrading, training or bridging programs and professional certification. This is a small-scale, recently developed test based on the Canadian Language Benchmarks. At the time that this project was undertaken, the MELA was being used by the Michener Institute of Applied Health Sciences in Toronto. It had not yet been approved by the CSMLS as proof of language proficiency for IEMLTs seeking registration. This project represents one step in the validation process of that tool's use for professional registration purposes.

What method is appropriate for the tests and the standard setting purpose?

In reviewing available methodologies, the project team determined that a modified Angoff approach would be most appropriate for the three target assessments. This is one of the most familiar, flexible and adaptable methodologies, and it has been used successfully for standard setting on a range of language testing tools. In our search for an appropriate method we found that the test developers for both TOEFL iBT and IELTS recommend that cut scores be set through standard setting methods, and both organizations provide standard setting guides for potential users.

The *TOEFL iBT Standard Setting Manual* is available free of charge from Educational Testing Services (ETS), the organization that developed the test. The manual includes a sampling of test materials along with detailed instructions on how to prepare and deliver a standard setting session. IELTS offers a short guide entitled *Setting Standard for IELTS Scores* and a questionnaire for test users in their *IELTS Scores Explained DVD*. The DVD contains test samples as well. MELA test developers have a standard setting procedure that was used to set cut scores for their various programs at the Michener Institute of Applied Health Sciences. The MELA standard setting materials are not available to the public but were made available for this research study.

We employed the procedures and materials recommended by the *TOEFL® iBT Standard Setting Manual* (ETS, 2006). In this method, a panel of experts first defines the minimally acceptable performance for a hypothetical candidate in all skill areas (speaking, writing, listening, and reading). Once this has been completed, the panel receives an orientation about the test and training on the judgement process. The panel then reviews all sections of each language test (speaking, writing, listening, and reading) and votes on recommended cut scores in two rounds. In round 1, the panelists independently evaluate each test sample or item and determine the first recommended cut score. These initial cut scores are recorded and summarized (mean, standard deviation, minimum, and maximum) and reported to the panelists. A discussion of the results of round 1 then follows. The panelists are given time to reconsider their first recommendations independently. The second, and final, cut scores are recorded, summarized, and analyzed after the end of the standard setting meeting. This process was applied to all three assessments.

Standard 4.21 of the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) states that when cut scores are based on direct judgements about the adequacy of test items or test performance levels (as was the case in this project) the judgemental process should be designed so that judges can bring their knowledge and experience to bear in a significant way. The first consideration then in selecting a method and preparing the materials should be that they are targeted and accessible to the expert panel.

What references and resources were helpful?

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (1999) *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Cizek, G.J. (2001). *Setting Performance Standards: Concepts, Methods and Perspectives*. New Jersey: Lawrence Erlbaum Publishers.

Cizek, G.J., & Bunch, M.B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. London: Sage Publications, Inc.

Educational Testing Services, ETS (2006). *TOEFL® iBT Standard Setting Manual*. Princeton, N.J.: Educational Testing Services.

International English Language Testing System, IELTS (n.d.). *IELTS Scores Explained DVD*. Cambridge: UCLES.

McNamara, Tim (2000) *Language Testing*. Oxford University Press.

Step 2 ▶ Plan the standard setting format and delivery

What delivery model will be most effective?

The format and delivery of the standard setting sessions was guided by the method selected and the tests being reviewed. There were two levels considered: the planning of the standard session and the operational aspects related to it. Our selected delivery model required two full days for each language test reviewed. In planning, we had to consider the availability of the panel members and the schedule of the facilitator and CSMLS representatives to meet for two consecutive days on three separate occasions. We also had to consider how to distribute the training, group work, test review and judgements so that panelists did not feel a cognitive overload. As expected, this was a complex administrative process.

Who will plan and administer the sessions?

The project was administered by the regulatory body (CSMLS). This proved to be efficient for many reasons. The regulator has the contacts required to recruit panelists, can provide access to in-house administrative resources, and, in our case, also provided a comfortable professional setting for the panel session. In many cases, regulators may also have in-house expertise in testing and perhaps even standard setting, as is the case with CSMLS (which conducts Angoff standard setting sessions to set cut scores on the professional practice exams). However, making use of external standard setting facilitators specialized in English language testing contributes to the credibility of the process. CSMLS sub-contracted the standard setting design, implementation, facilitation and analysis to a group of language testing experts, and managed the operational aspects of the sessions.

What costs will be involved?

There are considerable costs involved in planning and delivering standard setting sessions. There is the cost of administration, the preparation of the session materials, the venue, food and accommodation, consulting fees, printing, postage, and other expenses. It is important to consider how the sessions will be funded before engaging in the process. Panelists will be more likely to attend if they can recover lost wages and expenses incurred for the session. These costs must be budgeted.

How much time will be necessary?

When considering the time factor, it is helpful to think of the standard setting process as having three phases: planning, delivery, and reporting. The planning component of the project included the operational aspects (conducted by the regulator) and the research component (conducted by the subcontractor) and required six months to complete. The delivery of the session was planned over a three-month period (a two-day standard setting session planned for each month), but because it was difficult to find two consecutive days where a group of professionals were available, the sessions were distributed over five months. We discovered that the recruitment and scheduling component of the project must begin as early as possible so that invited panelists can set aside time in their work schedules. It is critical to recruit and get commitments from panelists as early as possible. Finally, an additional six months should be planned for the analysis and reporting phase of the project. This was not the case in this project; we allowed for three months but, because the sessions could not be scheduled within that period, we did not have adequate time to complete the reports. Allowing enough time for reporting is particularly important when the regulatory chooses to partner with a team of experts because time is needed for consultations

and review of the final report.

Based on the *TOEFL® iBT Standard Setting Manual* (ETS, 2006) each standard setting session required two full days to examine each of the four sections of each test (speaking, writing, reading, and listening). During Day 1, the panelists described a minimally acceptable performance standard for health care, received training on the standard setting process, and recommended passing scores for the speaking and writing sections of the test. During Day 2 the panelists recommended cut scores for the listening and reading parts of the test, and evaluated the standard setting process. The standard setting training and the definition of the minimally acceptable performance were completed in the first two-day session (the TOEFL iBT session) so the subsequent meetings were shortened since only a refresher was required.

The following were the agendas for Day 1 and Day 2:

Day 1: Standard Setting Agenda	
Standard setting begins at 8:30 sharp	
8:00 – 8:30	Refreshments
8:30 – 8:45	Confidentiality agreement and consent forms
	Welcome and introductions/ Panel introductions
	Overview by facilitator
8:45 – 9:30	Definition of minimally acceptable speaking ability
9:30 – 10:30	Introduction to speaking prompts and samples
	Review of speaking scoring procedures
	Training on making judgements about speaking
10:30 – 10:45	Refreshments
10:45 – 11:45	Round 1 speaking judgements
11:45 – 12:30	Group discussion
	Round 2 speaking judgements
12:30 – 1:15	Lunch
1:15 – 2:00	Definition of minimally acceptable writing ability
2:00 – 2:45	Introduction to writing prompts and samples
	Review of writing scoring procedures
	Training on making judgements about writing
2:45 – 3:00	Refreshments
3:00 – 4:00	Round 1 writing judgements
4:00 – 4:45	Group discussion
	Round 2 writing judgements
4:45 – 5:00	Introduction to Day 2 agenda and activities
	Explanation of productive and receptive skills
	Introduction to listening and reading test content
	Homework assignment
	Wrap up and adjourn

Day 2: Standard Setting Agenda	
Standard setting begins at 8:15 sharp	
8:00 – 8:15	Refreshments
8:15 – 9:15	Definition of minimally acceptable listening ability
9:15 – 10:15	Review of listening tasks and items
	Review of listening judgement procedure
	Practice judgements on listening
	Debriefing, discussion, and clarification
	Quick washroom break if needed
10:15 – 11:00	First round listening judgements
11:00 – 11:30	Refreshments
	Facilitators tally results for discussion
11:30 – 12:15	Listening discussion and final judgements
12:15 – 1:00	Definition of minimally acceptable reading ability
1:00 – 1:45	Lunch
1:45 – 2:30	Review of reading tasks and items
	Review of reading judgement procedures
	Practice judgements on reading
	Debriefing, discussion, and clarification
	Refreshments and first round judgements
2:30 – 3:45	As panelists finish, facilitators enter results
	Group discussion and final reading judgements
3:45 – 4:45	Standard setting evaluations
4:45 – 5:00	Wrap up and adjourn

What are some specific recommendations for planning and delivery of the sessions?

The *TOEFL® iBT Standard Setting Manual* (ETS, 2006) outlines a detailed timeline for the preparation of the TOEFL iBT session, and their material has provided the basis for what follows. Below is our adapted version of their timeline, which reflects our experience.

Six months prior to the standard setting session:

- Identify and invite panelists.
- Select dates and confirm panelist attendance.
- If there are out-of-town panelists, make arrangements for travel and accommodations.
- Select an adequate location (comfortable and professional – remember that panelists will be working in the same room for 2 days, at three different instances), make room arrangements and plan for catering.

One month prior to the standard setting session:

- Confirm panelist attendance, and send them an agenda and pre-session assignment.
- Prepare session materials (print panelist booklets, review power point presentation, name tags for participants, flip charts and markers).

- Reserve equipment (computer, LCD projectors, good quality speakers).

One week prior to the standard setting session:

- Facilitators meet to review the procedures and review the panelist booklet and the power point presentation, and to check the sound files.
- Decide on the seating of the panelists. To facilitate small group discussions, divide the panel into three groups. Be sure that each group represents a mix of stakeholders. Prepare name tags.
- Check that all room arrangements are in order and that all equipment is in place and has been tested (or that a technician is available to ensure this will be the case).
- Confirm catering and refreshments ordered.

On the day of the standard setting session:

- Distribute panelist confidentiality forms (for test materials and panelist discussion and findings).
- Distribute expense forms (as required).
- Conduct evaluation form at the end of each day.

What references and resources were helpful?

Cizek, G.J., & Bunch, M.B. (2007). Chapter 3: Common Elements in Setting Performance Standards in *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. London: Sage Publications, Inc. p 59-63.

Educational Testing Services, ETS (2006). Standard Setting Steps in *TOEFL® iBT Standard Setting Manual*. Princeton, N.J.: Educational Testing Services.

International English Language Testing System, IELTS (n.d.). Setting Standards for IELTS Scores in *IELTS Scores Explained DVD*. Cambridge: UCLES.

Step 3 ► Select and recruit the panel of experts

What is the recommended practice?

According to the format and method we adopted for this project, the *TOEFL® iBT Standard Setting Manual* (ETS, 2006) the composition of the panel is described as “a panel consisting of no fewer than 12 and no more than 20 members who are familiar with language and other issues confronted by first-year students whose native language is not English.” (*TOEFL® iBT Standard Setting Manual - Standard Setting Steps*, 2006, p.1).

Clearly this definition, based on an expectation of academic usage, was inadequate for the purpose of professional accreditation in health care, so we sought a definition of the ideal composition of a panel in the standard setting literature. In their standard setting guide, Cizek and Bunch (2007) list as a first step the selection of a panel that is large and representative of the stakeholders. They also suggest that the definition of a “qualified panelist” is necessarily related to the purpose of the standard setting task and emphasize that the quality of the group of persons selected to participate in the standard setting procedure can affect the eventual standard recommended even more than the standard setting method used (Cizek & Bunch, p. 49). Raymond and Reid (2001) describe the specific qualities of panelists and argue that participants for standard setting panels should be: a) subject matter experts; b) have knowledge of the range of proficiency and individual differences in the examinee population; c) appreciate the consequences of the standards; and d) collectively represent all relevant stakeholders.

What will be the composition of the expert panel?

We considered the criteria described in the literature in relation to the purposes of the CSMLS standard setting sessions to recommend language proficiency standards for medical laboratory technology and developed a description of an ideal panel as follows:

Position	Number	Qualifications
Employers	2	Individuals who hire and manage IEHPs (managers, supervisors, human resource management/consultants) from a range of workplaces who have had exposure to IEHPs for 3 years or more.
ESL specialists	2-3	ESL professionals who have worked in a related occupation-specific ESL initiative for a period of at least 7 years.
Public representatives	2-3	Patients/clients/consumers or their representatives who have dealt with IEHP practitioners in this profession.
Faculty/ Trainers	2	Individuals with content expertise who have participated in the development and implementation of bridging programs for IEMLTs.
IEHPs	2-3	Individuals with at least three years of work experience in this profession, who are both content experts and peers (not managers).
HPs trained in Canada	2-3	Individuals with at least three years of work experience in this profession, who have worked alongside IEHPs, and who are both content experts and peers (not managers).
Other healthcare professionals	2	Doctors, nurses, and other health care professionals who communicate with people in this profession on a regular basis and who have at least three years of experience in health care.
Regulator/ Association	2	Representatives of regulatory bodies or professional associations whose decisions may be affected by the outcomes of the standard setting session. Consider provincial and national organizations.
Total (ideal)	19	(Acceptable range is 12 – 20).

NOTE: IEHPs – Internationally Educated Health Professional

How will the expert panelists be selected?

In this project, thirteen panelists, representing a wide range of stakeholders, were convened to participate as experts for the standard setting sessions. In addition to the qualifications mentioned by Raymond and Reid (2001) above, panelists for this standard setting project also had to be: willing to commit to 2-3 hours of reading and reflection on materials to be sent at least ten days before the session and be available for six days (three sets of two consecutive days) between September and December 2008, possibly January. Only panelists who met these criteria were confirmed. Both medical laboratory technology and English language specialists were included in the panel. Ten were native speakers of English. Two of the participants had taken an English language proficiency assessment. Four had participated in previous standard setting studies.

Expert panelists were recruited by CSMLS through a targeted letter-writing campaign. The target group included qualified CSMLS members, MLT employers, IEMLTs, and ESL specialists. The final group was selected based on their qualifications and availability for three two-day sessions. Consulting fees were paid to individuals whose attendance was not supported by their employers.

We considered it critical to engage an appropriate mix of stakeholder expertise to bring varied perspectives to the definition of minimally acceptable ability. Additionally, having panelists who were interested in the issue guaranteed their full participation in the completion of pre-session tasks and in the in-session discussions.

What references and resources were helpful?

Raymond & Reid (2001) Who Made Thee a Judge? Selecting and Training Participants in Standard Setting in Cizek, G. *Setting Performance Standards*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers pp. 119-157.

Cizek, G & Bunch, M. (2007) *Standard Setting: A Guide to Establishing and Evaluation Performance Standards on Tests* Thousand Oaks: Sage Publications.

Step 4 ► Collect and prepare the standard setting materials

What materials are needed?

The materials required for a standard setting procedure will be determined by the method selected and the tests being reviewed. What the literature recommends is that the process be documented clearly (AERA/APA/NCME, 1999; Hambleton, 2001; Cizek & Bunch, 2007) so it is helpful to have a systematic approach to the collection, organization and storage of materials and data related to the standard setting sessions planned.

The following materials were prepared for each standard setting session:

- **Agenda:** A sample of the standard setting agenda appears in *Step 2 - Plan the standard setting format and delivery*. This agenda was adapted from the one available in the *TOEFL® iBT Standard Setting Manual* (ETS, 2006). We carefully considered the panel's overall unfamiliarity with ESL and language testing theory and practice when setting out the agenda so that the participants would not be overwhelmed by the new information they were receiving. We were also aware that they were being asked to take the test, which is, in itself, a cognitively demanding task, and we planned around that.
- **Power Point Presentation:** Although the *TOEFL® iBT Standard Setting Manual* (ETS, 2006) provides a very thorough presentation in the manual, the intended audience was so different from ours that we decided to create a custom power point presentation for each session to reflect both our audience and each test's unique design. Reminders were embedded throughout the presentation to ensure that all of the information the panel needed to make their judgements was included, that in-process evaluations were conducted periodically (to verify that the panel understood the instructions) and that the definitions of minimally acceptable language competency were referenced throughout. We found the power point presentation model effective because it assisted the facilitator in keeping the group on task while maintaining the necessary pace to cover all the needed steps and information. In addition, this approach allowed us to embed sound and video in the slides. One down side to the power point presentation was that the noise generated by the computer and the LCD projector sometimes interfered with the playing of the audio files.
- **Test materials (speaking, listening, reading, writing):** Booklets for each section of each test were prepared including the test description, tasks and answer keys. In the case of TOEFL, all materials were supplied with the manual, IELTS materials were available in the IELTS standard setting package, but we opted to use the *IELTS Handbook* and *Official IELTS Practice Materials* because they were more thorough than the materials in the *IELTS Scores Explained* DVD. MELA standard setting materials were also adapted to the methods used with the CSMLS.
- **Sound files:** These may be necessary for listening to test input and interview/speaking test samples where appropriate. It is critical to have good quality sound files and speakers so that panelists can do their jobs properly. We tested the sound in the room where the standard setting session would take place to ensure clarity and proper volume.
- **Judgement forms:** These forms should be prepared specifically for the panel and standard setting task. The judgement form in the *TOEFL® iBT Standard Setting Manual* (ETS, 2006) asks panelists to answer the question, "has this student demonstrated the minimally acceptable speaking skills adequate for first-year undergraduate/graduate school studies at your university?" The forms that we created asked panelists to answer the question: "Does this person demonstrate the minimum level of speaking ability that an MLT needs for safe and effective practice?"

- **Session Evaluation:** The validity of a judgement-based standard setting procedure is vested on the participant's confidence in the process. Participant feedback is commonly gathered through an evaluation instrument. We developed an evaluation form for this purpose (See Step 5 – Conduct an evaluation).

The greatest challenge with the *TOEFL® iBT Standard Setting Manual* (ETS, 2006) was the fact that the materials assumed the standard setting would be implemented within a college or university context to set standards for admissions into an academic program. Substantial revisions and adaptations were necessary to contextualize the materials to suit a professional certification purpose and to best utilize the background and expertise of the invited panel. The methodology, however, was fully adequate for the purpose.

What additional information must be researched?

In addition to the materials needed to conduct the standard setting sessions, the literature recommends that information provided to the panel and decision makers about the final cut score recommendations needs to be collected (Cizek & Bunch, p. 54-56). This includes:

- **Normative information:** data that permits each panelists to perceive how his or her ratings compare to other panelists' judgements. We did this by providing the mean, standard deviation, minimum, and maximum for the first round of discussion. This data was calculated automatically on a data collection Excel spreadsheet (see below).
- **Reality information:** this is data to help panelists perceive the accuracy of their judgements and includes information such as statistics on item difficulty (item p values, mean task performance). The *TOEFL® iBT Standard Setting Manual* (ETS, 2006) provides sufficient data for this purpose, but the same data was not available for the other two tests, so we used this data sparingly.
- **Impact information:** is data to help panelists understand the consequences of their decisions. In the context of these standard setting sessions this included the existing cut scores, anecdotal information about candidate success, and cut scores set by other organizations.

The *TOEFL® iBT Standard Setting Manual* (ETS, 2006) provided detailed statistics on the TOEFL test in general as well as for the particular version enclosed as a test sample. Reality data, or statistics such as p values and the standard Error of Measurement (SEM) are important to the final decision-making (this is discussed in more detail in Step 5 – Conduct an evaluation). These statistics were not readily available for IELTS or MELA. Finding the appropriate statistics for IELTS required additional research, and SEM for MELA were unavailable. The availability of data had an impact on our decisions about feedback given to panelists during the session and to the final decision-making panel. Although the TOEFL iBT package provided reality data, we decided to use the normative data as the main feedback to panelists. This included mean, standard deviation, minimum, and maximum in relation to panelists' judgements. Impact data was also discussed.

There were two reasons for this decision. The first was that not all tests we reviewed had the same level of test taker data available. Second, the testing audience for the general statistics is a younger university applicant group whereas the group that we were conceptualizing in this standard setting process is a more mature, adult group with many years of experience in the health care sector. We felt that the data would not be helpful in providing a picture of test taker performance that reflected our target audience (IELTMs). Third, the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) standard 14.7 states that "the level of performance required for passing a credentialing test should be dependent on the knowledge and skills necessary for acceptable performance in the occupation or profession and should not be adjusted to regulate the number or proportion of persons passing the test".

Where can the materials be procured?

The *TOEFL iBT Standard Setting Manual* is available free of charge through Educational Testing Services (ETS). The manual includes a presenters manual, a participant booklet, a ready-made power point presentation, audio files and supporting documentation. We ordered the DVD from ETS online (See resources.) The *IELTS Scores Explained DVD* includes sample test materials and a short guide for *Setting Standard for IELTS Scores* as well as a questionnaire for users. The DVD can be ordered at www.ielts.org. MELA test developers have a standard setting procedure that they had used at the Michener Institute of Applied Health Sciences to set cut scores for their various programs. The MELA standard setting materials are not available to the public but were made available for this research study.

How should they be reproduced?

The main consideration in reproducing the test materials is copyright infringement. We respected the copyright of the *TOEFL® iBT Standard Setting Manual* (ETS, 2006) and used the materials solely for the purposes of the standard setting sessions. The panelist manuals were shredded after this use. Similarly, materials for IELTS and MELA were used for the sole purpose of standard setting then shredded. Proper acknowledgment of the source materials was made throughout the project and in the reporting process.

What references and resources were helpful?

Cizek, G.J., & Bunch, M.B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. London: Sage Publications, Inc.

Educational Testing Services (ETS) (2006). *TOEFL® iBT Standard Setting Manual*. Princeton, N.J.: Educational Testing Services.

International English Testing System (IELTS) (2007). *Handbook*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).

International English Testing System (IELTS) (N/A). *IELTS Scores Explained DVD*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).

International English Testing System, IELTS (2003). *Official IELTS Practice Materials – Updated January 2005*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).

Step 5 ► Evaluate the standard setting process and results

What is the recommended practice?

Standard setting outcomes are validated by gathering data about the design and implementation of the session and the level of confidence of the panelists about the training they received, their understanding of the role and the process, their ability to follow the standard setting procedures and in the performance standards that they recommend (Cizek & Bunch, p. 59). The standard setting procedures as well as the standard setting results must be evaluated. Standard setting evaluations are recommended while the sessions are in progress to verify whether panelists have understood their task, right after the session to verify whether panelists felt confident about their work, and at the end of the process, to evaluate the overall effectiveness.

How will the quality of the standard setting process be known?

The principal evaluation instrument for this standard setting study was a questionnaire administered to the panelists at the end of each standard setting session. This is a standard and recommended procedure (Cizek & Bunch, 2007; AERA/APA/NCME, 1999). The questionnaire asked the panelists to evaluate the standard setting process (training, time, etc.) and to indicate how comfortable they were with the final cut score. In this way, the standard setting process and organization, as well as the comfort and confidence of panelists with the process and its results, could be assessed. We also conducted informal in-process evaluations throughout the sessions by checking verbally with panelists about their comprehension of the process.

How will the reliability and validity of the standard setting results be known?

The various elements of standard setting evaluation practice can be categorized into three areas: procedural, internal and external (Cizek & Bunch, p. 60). The reliability and validity of standard setting sessions is measured using these three levels of data. The following questions related to each of these elements should be considered in planning the sessions and in conducting the evaluation:

Procedural

- Explicitness: Were the standard setting purposes and procedures clearly articulated at the onset of the project?
- Practicability: Was the implementation of the procedures and data analysis easy to follow? Were the procedures credible and interpretable to relevant audiences?
- Implementation: Was the selection of participants, definition of performance and data collection reasonably and systematically conducted?
- Feedback: Did the participants have confidence in the process and the resulting cut scores?
- Documentation: Were the features of the study reviewed and documented for evaluation and communication purposes?

Internal

- Consistency within method: Were the estimates of the cut scores precise?
- Intraparticipant consistency: Were participants able to provide ratings that were consistent with empirical item difficulties? How much did the ratings change across rounds?
- Interparticipant consistency: Were item ratings and cut scores consistent across participants?
- Decision consistency: How will the extent to which repeated application of the identified cut scores yield consistent classification of candidates measured?
- Other measures: Was there consistency of cut scores across item types, content areas, and cognitive processes?

External

- Comparisons to other standard setting methods: What would the agreement of cut scores be across replications using other standard setting methods?
- Comparison to other sources of information: What is the relationship between decisions made using the test to other relevant criteria?
- Reasonableness of cut scores: Are the cut score recommendations feasible or realistic?

The procedural validity can be evaluated as described above, through a questionnaire administered to participants. A sample evaluation form is provided here. An adaptable version of an evaluation form can be found at www.sagepub.com/cizek/evaluationform (Cizek & Bunch, p. 61).

Please indicate the level of agreement with each of the following statements and add any additional comments you have on the process at the bottom of the page.

Statement		Strongly agree	Disagree	Agree	Strongly Disagree
1	The orientation provided me with a clear understanding of the purpose of the meeting.				
2	The workshop leader clearly explained the task.				
3	The training and practice exercises helped me to understand how to perform the tasks.				
4	Taking the test helped me to understand language proficiency testing.				
5	The performance level descriptors were clear and helpful.				
6	The large and small group discussions aided my understanding of the process.				
7	There was adequate time provided for discussion.				
8	There was an equal opportunity for everyone in my group to contribute his/her ideas and opinions.				
9	I was able to follow the instructions and complete the rating tasks successfully.				
10	The discussions after the first round were helpful to me.				
11	The discussion after the second round of ratings was helpful to me.				
12	The information and feedback between rounds was helpful to me.				
13	I am confident about the defensibility and appropriateness of the final recommended cut scores.				
14	The facilities and food service helped create a productive and effective work environment.				
15	Comments:				

Source: Cizek, G & Bunch, M. (2007) *Standard setting: A Guide to Establishing and Evaluation Performance Standards on Tests* Thousand Oaks: Sage Publications p. 62

How will the standard setting results be recorded and analyzed?

Internal evaluation deals with the quality of the data collected, and can be measured by looking at standard deviation and standard error or judgement. We found it essential to have language testing and measurement experts on the standard setting team that could conduct this type of analysis for us. We found that having a standard setting manual that contained ready-made data

collection tools very helpful. The *TOEFL® iBT Standard Setting Manual* (ETS, 2006) includes a pre-programmed Microsoft Excel Spreadsheet for the collection and calculation of data. This spreadsheet automatically calculated the mean, median, minimum, and maximum for the panel's judgements. It was important to have these automatically generated, as we needed this information as feedback to participants. In the future we might consider a more interactive approach such as the use of audience response technology (e.g., Data On The Spot, DOTS, <http://www.dataonthespot.ca/>) as a way to collect panelist judgements and receive immediate results for feedback during the sessions.

The test measurement expert on the team was able to use for the TOEFL spreadsheets as a model for both the IELTS and MELA sessions. However, the score conversion charts available for IELTS were less detailed than needed for the ready-made spreadsheets, so we had to seek a set of test materials and a conversion chart that would be appropriate. We also relied on the test measurement expert to analyse the data collected and calculate the standard deviation and standard error of judgement (SEJ) statistics that would allow us to evaluate the reliability of the standard setting results.

External validity is the most difficult to gather as it requires longitudinal or comparative research that would require additional human and financial resources.

What references and resources were helpful?

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999) *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Cizek, G & Bunch, M. (2007) *Standard Setting: A Guide to Establishing and Evaluation Performance Standards on Tests* Thousand Oaks: Sage Publications.

Hambleton, R. K. (2001) *Setting Performance Standards on Educational Assessments and Criteria for Evaluating the process* in Cizek, G.J. *Setting Performance Standards*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

www.dataonthespot.ca

www.sagepub.com/cizek/evaluationform

Step 6 ► Set the final cut scores and communicate the results

What is the recommended practice?

According to standard setting scholars and practitioners, the last step in standard setting should be to compile validity evidence and technical documentation (Hambleton, p. 94). In their standard setting manual, Cizek and Bunch (2007) describe the last step as the assembly of “documentation of the standard setting process and other evidence, as appropriate, bearing on the validity of the resulting performance standards (Cizek & Bunch, p. 36). The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) state that the recruitment, training and evaluation of panelists must be clearly documented (Standard 1.7); that where cut scores are specific for the selection of candidates, the standard error of measurement should also be reported (Standard 2.14); and that interpretations involving one or more cut scores, the rationale and procedures used for establishing those should be clearly documented (Standard 4.19).

For each standard setting session conducted we prepared a report that followed these recommendations and contained the following:

- A description of the standard setting process;
- A description of the panel, including the number of participants and their affiliation;
- Tables reporting the panelist votes, by round, including mean, minimum, maximum, median, standard error of measurement for the test (as reported by the test developer) and standard error of judgement of the standard setting process.

How are the panel's recommendations communicated to the decision makers?

The *TOEFL® iBT Standard Setting Manual* (ETS, 2006) includes a Summary Memo Template to be used in communicating the standard setting results to stakeholders and decision makers. The memo outlines the purpose of the study, gives a description of the test, the standard setting process and the participants, and lists the recommended cut scores. The *TOEFL® iBT Standard Setting process* recommends that this memo be distributed to the standard setting panel and to the decision-makers in the institution so that the final cut scores may be established.

In keeping with this recommendation and, in addition to the reports specific to each test (described above), the standard setting team prepared a thorough report that included the categories recommended by the *TOEFL® iBT Standard Setting Manual* (ETS, 2006):

- Purpose of the study;
- Description of the tests reviewed;
- Documentation of the standard setting process;
- Documentation describing the selection and recruitment of panelists;
- A statement of the recommended cut scores; and
- Information for policy makers including normative, reality and impact data.

As recommended by Perie (2006) this information would ideally be discussed at an articulation meeting with key stakeholders and decision-makers, the purpose of which is to review the results of the standard setting workshops in the context of, in the case of this project, the credentialing process. In our project, the expert panel brought their knowledge and experience to bear on the judgement making process. In the context of this project, however, the final cut scores are a policy decision to be made by the regulator. The standard setting team prepared detailed technical documentation on the process, the cut score recommendations and additional relevant data and information for the regulatory body to consider in making the final cut score decision.

Once a final decision on cut scores is made, a policy statement briefly describing the due diligence process undertaken would also be prepared and released publicly. A good example of this is offered by the National Council of State Boards of Nursing in the USA (See resource list).

What follow up activities are needed?

It is advisable to track the standard by means of longitudinal data collection and studies that establish the predictive validity of the recommended cut scores over time. Standards should be reviewed periodically, and this data would not only provide external validity evidence to support the standards but also provide data to support any necessary changes.

What references and resources were helpful?

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999) *Standards for Educational and Psychological Testing*. Washington, DC: AERA.

Cizek, G & Bunch, M. (2007) *Standard setting: A Guide to Establishing and Evaluation Performance Standards on Tests*. Thousand Oaks: Sage Publications.

Hambleton, R. K. (2001) Setting Performance Standards on Educational Assessments and Criteria for Evaluating the process in Cizek, G.J. *Setting Performance Standards*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

National Council of State Boards of Nursing (December 2005). *Position Statement: NCSBN Recommends English Language Proficiency Standards for Internationally Educated Nurses*. Available at: www.ncsbn.org.

Perie, M. (2006). *Convening an articulation panel after a standard setting meeting: A how-to guide*. Center for Assessment, The National Center for the Improvement of Educational Assessment. Available at: http://www.nciea.org/publications/RecommendforArticulation_MAP06.pdf.

Resource List

- American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999) Standards for Educational and Psychological Testing. Washington, DC: AERA. Cizek, G.J. (2001). *Setting Performance Standards: Concepts, Methods and Perspectives*. New Jersey: Lawrence Erlbaum Publishers.
- Cizek, G.J., & Bunch, M.B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. London: Sage Publications, Inc.
- Educational Testing Services (ETS) (2006). *TOEFL® iBT standard setting manual*. Princeton, N.J.: Educational Testing Services.
- Educational Testing Services (ETS) (2007). TOEFL® iBT Score Reliability and Generalizability. Available at: http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Score_Reliability_Generalizability.pdf.
- Educational Testing Services (ETS) (2007). *Understand your TOEFL® iBT scores*. Available at: <http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509?vgnextoid=0f8c6e873ee77110VgnVCM10000022f95190RCRD&vgnnextchannel=5d76a1e13bf36110VgnVCM10000022f95190RCRD>.
- Hambleton, R. K. (2001) Setting Performance Standards on Educational Assessments and Criteria for Evaluating the process in Cizek, G.J. *Setting Performance Standards*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.
- International English Language Testing System, IELTS (2003). *Official IELTS Practice Materials – Updated January 2005*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).
- International English Language Testing System (IELTS) (2007). *Handbook*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).
- International English Language Testing System (IELTS) (2007). *Information for Candidates*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).
- International English Language Testing System (IELTS) (2003). *Official IELTS Practice Materials – Updated January 2005*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).
- International English Language Testing System (IELTS) (N/A). *IELTS Scores Explained DVD*. Cambridge: University of Cambridge Local Examinations Syndicate (UCLES).
- International English Language Testing System (IELTS) (n.d.). *Test performance 2007*. Available at: http://www.ielts.org/teachers_and_researchers/analysis_of_test_data/test_performance_2007.aspx.
- National Council of State Boards of Nursing (December 2005). Position Statement: NCSBN Recommends English Language Proficiency Standards for Internationally Educated Nurses. Available at: www.ncsbn.org.
- Perie, M. (2006). Convening an articulation panel after a standard setting meeting: A how-to guide. Center for Assessment, The National Center for the Improvement of Educational Assessment. Available at: http://www.nciea.org/publications/RecommendforArticulation_MAP06.pdf.