

Adaptation within a language: Considerations for standard setting

Chad W. Buckendahl
Buros Center for Testing
University of Nebraska, Lincoln

Andrew Blackhurst
University of Cambridge ESOL

Elaine Rodeck
Buros Center for Testing
University of Nebraska, Lincoln

Paper presented at the International Test Commission conference, Brussels, Belgium,
July 6-8, 2006

Abstract

Standard setting methods applied to constructed response items often utilize examinee anchor performances as part of the procedures. When implementing these methods, researchers have recommended the importance of representing the range of score points for panelists. However, if there are a variety of ways that a score point might be attained, multiple representations of the score point are likely necessary to illustrate this variation to panelists. This paper describes factors that were considered for a standard setting study that included selecting anchor performances for an English literacy test that is being used as a component of immigration eligibility. The sampling of the performances and other considerations for standard setting are discussed in the context of the *Guidelines for Test Adaptation and Translation* (International Test Commission, 2000).

Keywords: test adaptation, performance assessment, standard setting

Adaptation within a language: Considerations for standard setting

Introduction

Many studies have applied language translation and adaptation methodologies across languages. For example, to expand the population of potential examinees, a testing agency may want to translate a test that was originally developed in English to another language (e.g., Dutch, French, German, Spanish). However, adaptation of a test for a broader population may not be limited to uses across languages. Different linguistic characteristics, conventions or dialects within a given language may contain significant variation that could interfere with interpreting the intended construct. Thus, there may be instances where adaptation within a language is also necessary to represent the intended or anticipated population of test takers.

The purpose of this paper is to describe considerations for test adaptation within a language using the *Test Adaptation Guidelines* (ITC, 2000) as a framework for the discussion. We will then illustrate how these considerations were applied in a standard setting study that recommended cut scores on the International English Language Testing System (IELTS) for use in an immigration eligibility setting.

Test adaptation

The International Test Commission's (2000) *Test Adaptation Guidelines* serve as a framework for test developers and users as they consider uses of tests for broad populations. The *Guidelines* have four distinct strands for developers and users to consider. First, there are contextual factors that should be eliminated to ensure that variability is not related to cultural characteristics or the construct. A related characteristic is that the construct should be substantively similar across intended or anticipated populations. Second, the test development and adaptation strand includes

recommendations to ensure that linguistic and cultural differences are fully considered when constructing or adapting a test. Third, the administration strand provides recommendations for ensuring that the administration strategies or conditions are similar across populations and that characteristics of the administration are not dependent on a particular socio-cultural context. The fourth strand focuses on documentation and score interpretations. This strand recommends documenting any changes that are made for a respective population and asks developers to indicate when there are factors that may change the interpretability of scores across these populations. Guidelines from each of these strands were considered in designing and implementing the standard setting study described below.

Illustrative Study

In this section of the paper we describe how the ITC (2000) Guidelines were considered as part of an operational standard setting study. The University of Cambridge ESOL Examinations organization commissioned a standard setting study in 2005 to gather information that would produce a range of recommended cut scores on the *International English Language Testing System* (IELTS). The cut scores were developed for use by Citizenship and Immigration Canada as part of immigration eligibility for those candidates who took this test to demonstrate English literacy skills. Because the IELTS modules contain both dichotomously and polytomously scored test items, multiple standard setting methods that were unique to the item or task format were used to elicit the panelists' judgments in the study. Some of these standard setting methods required anchor performances to represent the range of score points that candidates may operationally demonstrate. The next section of the paper describes characteristics of the test and the standard setting methodologies and the adaptation needs that were considered at different stages in the project.

Information about IELTS

The *International English Language Testing System* (IELTS) was developed to measure language abilities of candidates ages 16 years and older who need to study or work where English is the primary language for communication (IELTS, 2003). Because of the different uses for the scores, there are two formats that are available to potential candidates depending on their needs. First, there is an Academic format that is appropriate for candidates who want to demonstrate their readiness to study in English at the undergraduate or graduate level. Second, a General Training format is available to candidates who may wish to use their scores for vocational training, work or immigration eligibility requirements. The General Training format emphasizes “basic survival skills in a broad social and educational context” (IELTS, p. 2).

The content specified in the IELTS is represented by four modules: listening, reading, writing, and speaking. Because the intended population of interest for the standard setting study illustrated here focused on potential immigrants to Canada and not for college or University admissions purposes, the General Training version of the reading and writing modules were used. Note that the listening and speaking modules are the same for both Academic and General Training formats.

The listening module contains forty multiple-choice, short answer, and completion items. There are four sections of the module, two of which represent language usage in social need situations; and two of which represent situations that candidates might encounter in typical educational or training contexts. Each item is scored as correct or incorrect and then transformed to an interpretative scale score (i.e., IELTS 9-band scale).

Similar to the listening module, the reading module of the exam is comprised of forty multiple-choice, short answer, and completion items. Within the General Training

format of the reading module there are three sections. The first section includes reading text about “social survival” that would be related to “basic linguistic survival in English” (IELTS, p. 7). In the second section, the content targets “training survival” and involves reading text with more complex language (IELTS, p. 7). The third section is characterized as “general reading” and focuses on “reading more extended prose with a more complex structure. . .” (IELTS, p. 7). Items in this module are also scored as correct or incorrect with the scores then scaled to an interpretative scale score.

The General Training format of the writing module consists of two tasks. Task 1 requires the preparation of a response of at least 150 words. For this task, test takers are asked to write a letter requesting information or explaining a situation. Task 2 requires a minimum of 250 words and is the more heavily weighted of the two writing tasks. For Task 2, test takers are asked to respond to an opinion or problem. The writing tasks are scored using a scoring rubric that corresponds to the IELTS 9-band scale. Examiners award marks (points) for Task 1 scripts on the following criteria:

Task Achievement (that is how clearly the purpose of the letter is presented, and how fully the bullet points are addressed); Coherence and Cohesion (that is how well the information and ideas are organized, and how well the information is linked) Lexical Resource (that is the range of vocabulary used, how accurately and how appropriate it is for the task); and Grammatical Range and Accuracy (that is the range of structures used, how accurately they are used and how appropriate they are for the task).

The first criterion used to assess Task 2 scripts differs. Candidates are assessed according to Task Response (that is how fully and appropriately the candidate has answered all parts of the task; the extent to which the candidate's ideas are relevant, developed and supported; the extent to which the candidate's position is clear and

effective). The other three criteria used are the same as for Task 1: Coherence and Cohesion; Lexical Resource; Grammatical Range and Accuracy.

The speaking module is the same for both Academic and General Training formats and consists of a one-on-one interview with three parts: introduction and interview, individual long-turn speaking, and a two-way discussion. Candidates' speaking module performance is scored on four criteria using a scoring rubric that corresponds to the IELTS 9-band scale. These scoring criteria are fluency and cohesion, lexical resource, grammatical range and accuracy, and pronunciation (IELTS, 2003).

Each candidate receives an overall band score that is calculated from the equally weighted four individual band scores obtained within each module. The overall band score represents a broad profile of the candidate's English literacy abilities.

Information about the panelists

Eighteen individuals served as panelists for the standard setting study. They were selected for their content expertise in English literacy skills instruction and their familiarity with the target population (i.e. those learning English as a second language). Of the panel, five members were IELTS examiners and five were ESL instructors. Panelists represented a range of professional expertise in addition to their content expertise. For example, panelists' experience included positions as a distance education instructor, Canadian Language Benchmarks (CLB) task writer, Director of CanTEST, Director of Second Language Certification Test, Test Development Coordinator (Centre for Canadian Language Benchmarks), IELTS examiner trainer, consultant in adult second language curriculum and testing, CLB assessor, Teachers of English as a Second Language trainer, and several were current or former college professors in related fields. Specialties of the panelists included pronunciation, student acquisition and teacher training, second language writing assessment, English for academic purposes, test

development, ESL writing instruction, and English/Romanian language and literature, and job search skills for foreign-trained professionals, among others. Participants also brought a considerable level of experience in their respective fields to the panel. The mean years of related professional experience for this panel was 17.9 years (SD=8.2) with a range of 3 to 34 years.

The panelists also had a wide variety of linguistic experience. This was judged as important for representing the range of linguistic abilities that potential candidates might possess. As a group they were familiar with over 18 languages including Arabic, French, German, Indonesian, Italian, Japanese, Korean, Latin, Norwegian, Russian, Tagalog, and Ukrainian. Additionally, panelists had worked with students that collectively spoke more than 40 languages. Some of these include Portuguese, Polish, Czech, Persian, Thai, Bengali, Mandarin, Cantonese, Turkish, Vietnamese, Serbo-Croatian, Hindi, Urdu, Cambodian, Farsi, Khmer, Dari, Bulgarian, Ethiopian, Punjabi, and Tamil.

Eight of the panelists had received special honors or awards, some of these included the Excellence in Teaching Award (School of Continuing Studies, University of Toronto, 1996), Instructional Skills Leader (Dubai, United Arab Emirates, 2001), the Wakefield Scholarship (British Columbia Teachers of English as an Additional Language, 1999), Best Teacher of the Year (Romania, 1984), 3M Award for Teaching Excellence & Educational Leadership (2002), Immigrant of Distinction (Professional category, Calgary, 2004), and one has been a recognized Canadian Language Benchmarks Expert since 2003.

Overview of Standard Setting Methods

This standard setting study applied two methods for estimating the panel's recommended range of cut scores for each module. Each method relies on different assumptions and is unique to the item types contained in the assessment (i.e. selected response, constructed response, performance tasks) and the scoring rules for the items (i.e. objectively scored versus subjectively scored). These methods were: a) a modified Analytical Judgment method (Plake & Hambleton, 2000), and b) a modified Angoff (1971) method. Each of these methods is described briefly below.

Analytical Judgment Method

A modification of the Analytical Judgment method (Plake & Hambleton, 2000) was used for the writing and speaking modules in the study. This method entails asking panelists to classify candidates' performance into defined categories. Classification is first at a broad level (e.g., Basic, Intermediate, and Advanced) and then narrowed down to identify the performance that would likely be produced by a target candidate. Selecting these illustrations of candidate performance represent another consideration for adaptation. Although the methodology requires performances to be selected to represent scores in both depth (i.e. multiple examples at a given score point) and breadth (i.e., examples across the full range of score points), with a language literacy test, there are likely additional considerations about how the performances are represented.

When selecting performances to represent score points, particularly in the speaking module, we strived to illustrate different representations of the construct from different regions of the world that might use English speech as one mode of communication. The performances that were selected for this study included representations of English from candidates with the following countries of origin: India, Indonesia, Nepal, Slovakia, Morocco, Poland, Italy, South Korea, Argentina, Hong Kong,

Kuwait, China, Japan, Thailand, and South Korea. This selection provided the panelists with a number of performances from potential candidates that demonstrated a variation in both score and representation of the speaking construct.

Candidate performances were also selected for the writing module to represent the range of score points in the scale. Multiple examples were provided for most score points to illustrate that different performance may achieve the same score. Adaptation considerations for the writing performances included how the papers were scored and the tolerance for different grammar or conventions that may be acceptable within the construct. Catering fully for each of the varied cultural backgrounds of a multi-ethnic candidature is not possible in large scale examinations. Rather, steps must be taken at the test development stage to ensure that no candidate is disadvantaged in relation to other candidates by the socio-cultural content of the test. Tasks and topics should be neutral with regard to candidates and materials are pilot tested in advance to confirm this and to filter out any material which may be culturally inaccessible/ inappropriate. Feedback from administrators and candidates on the accessibility of topics and tasks is elicited during pilot testing and used in modifying materials where necessary. Examiner training is also important in this regard.

Yes/No Variation of the Angoff Method

The Yes/No Variation of the Angoff (1971) method (Impara & Plake, 1997) entailed asking panelists to examine each item on the test and estimate how a typical borderline target candidate would perform on that item. For the IELTS, panelists were asked (after a training activity) to conceptualize specific target candidates (i.e. Initial Intermediate, Initial Advanced) with whom they had worked or supervised. Keeping these target candidates in mind, they were directed to indicate, for each item, whether the target candidates they had in mind would answer the item correctly or not (Right or

Wrong). This was done for the multiple-choice, short answer, and completion items the panelists rated. After an initial rating, actual performance data (item level proportion correct) from a representative sample of IELTS test takers¹ was provided to the panelists. Panelists also received information on their individual recommendations and the group's average recommendations.

After seeing these data, panelists made a second estimate of whether the Target Candidates would answer correctly or not. The second estimates could be either the same or different from their first estimates. These data provide a reality check to ensure that expected performance is not set either unrealistically high or low because the panelist has misjudged how hard or easy the item actually is. The cut scores are based on the second estimates and are calculated by summing, for each panelist, the number of "Right" items and then averaging those values across the panelists. This value typically represents the lower boundary of the recommended cut score range. This method was used for the listening and reading (General Training version) modules of the IELTS for the 2005 study.

For the listening module, the recording that is played for the candidates also represents test adaptation considerations. Specifically, the Handbook (IELTS, 2003) states, "A range of English accents and dialects are used in the recordings which reflects the international usage of IELTS" (p. 6). Inclusion of this range of accents and dialects is a consequence of the historical evolution of IELTS. Its precursor, the English language Testing system featured British accents and reading texts from Britain only. In 1989 the test was extensively revised as a test for broader use among candidates seeking vocational training, further or higher education in the UK, and Australia. There are now item writer teams in the UK, Australia and New Zealand, and they are encouraged to

¹ Test takers in this sample reflect candidates who designated Canada as their intended destination and that

source on listening material and reading texts exhibiting features of “Englishes” used these regions and in North America.

It is important that IELTS employ as wide and appropriate range of content as possible. IELTS has been designed to provide a globally recognized certification of English language proficiency, and it is therefore reasonable that the language represented in the test should reflect varieties of English, that test candidates’ ability to function in the widest range of international contexts, rather than a more restricted local context. At the same time, it is important not to disadvantage any particular candidate group. As part of Cambridge ESOL’s commitment to cultural awareness, question writers are cautioned that material must be generally accessible, avoiding bias for, or against, any candidates whatever their field of specialization, country of origin or country of destination.

It follows that processing the text should not require understanding of terms which are narrowly specific to a particular discipline that any very low frequency lexical items should be glossed, and references to historical events or personalities should be glossed or removed as they assume background knowledge on the part of the reader that may be unrelated to the intended construct. The thematic link between the Reading and Writing modules, which was employed in the original IELTS test, was removed in 1995 in part because candidates’ cultural and educational background could have an impact upon the extent to which they exploited the reading input (Cambridge ESOL, 2004).

In the next section we discuss how these different aspects of the standard setting study – panelist selection, candidate performance representation, test content representation – apply to the ITC’s (2000) Test Adaptation Guidelines.

Application to the ITC Guidelines

immigration was their purpose for taking the test.

There are a number of the ITC Test Adaptation Guidelines (2000) from each of the strands that were applied in this study. Table 1 summarizes how the ITC Guidelines were considered in this standard setting study.

[Insert Table 1 Here]

First, within the *Context* strand, Guideline C.1 suggests that the effects of cultural differences that are not relevant to the intended purpose should be minimized. In this standard setting study, characteristics of the panel, illustrative candidate performance, and content representation all required consideration of this recommendation. In selecting the panelists, it was important that their experience with students developing English skills represented a broad range of the potential pool of candidates for immigration to Canada. As noted above the panel members were experienced with a fairly wide range of geographic regions and cultures that speak or are learning to speak English.

A second consideration related to this guideline is the selection of illustrative performances for modules that relied on the Analytical Judgment Method (Plake & Hambleton, 2000) to communicate the panelists' recommended cut scores. These performances were selected to reflect, so far as possible within the inevitable practical constraints, the variability in score performance and representation of the language by different cultures. Third, the items and tasks represented a variety of stimuli that may be encountered in English-speaking cultures. This is an important consideration as culturally-specific items or tasks may advantage or disadvantage candidates who may not be as familiar with the context.

In the *Test Development and Adaptation* strand, there are two guidelines that have applications to standard setting studies. Guidelines D.1 and D.2 both speak to developers' responsibilities to ensure that cultural and linguistic differences among populations are included both in how the constructs are represented, but also in the operational

characteristics of the test (e.g., items/tasks, directions, scoring rubrics). For standard setting, these are characteristics that panelists' would need to be able to consider in their judgments. For example, if the scoring strategies for items and tasks are specific to a particular culture, the panelists could mitigate their judgments knowing the parameters or restrictions in the scoring criteria.

Within the *Administration* strand, Guideline A.2 suggests that administrators should be sensitive to factors that may be related to testing materials, the administration protocols, and scoring approaches. This is particularly important as variations in these areas may threaten the validity of the inferences made about the scores. In a standard setting study, panelists should be informed of these test characteristics to be better able to predict how target candidates will perform through consideration of potentially mitigating factors. For example, acceptable variations in scoring or interpreting candidates' responses should be included in the administration and scoring manual to allow panelists to know the range of acceptable responses and consider those in their judgments. Sharing this information with panelists in advance provides another source of information that allows them to contextualize their judgments.

One guideline found in the *Documentation/Score Interpretations* strand can be applied in a standard setting study. Guideline I.4 recommends that the developer document how socio-cultural contexts may impact candidates' performance on the test. More importantly, the developer should provide strategies for users to interpret scores that may consider these variations. In this illustration, the developer noted how different variations in English were considered in the development and scoring of their test modules. Just as a Reading and Listening source text may exhibit the features characteristic of North American or Australian English, so it is recognized that the candidates may produce English which has been shaped by the varieties of English to

which they have been exposed pedagogically. Thus, British and American spellings or grammatical conventions are equally acceptable. A key concept here is communicative competence: candidates are assessed on their ability to make themselves understood, rather than how far short they might fall from a notional native speaker standard. Errors are penalized in so far as they hinder the process of communication.

The face-to-face format of the IELTS speaking test is unique in this context: there is direct interaction between the participants; and assessment in the speaking test will take into account multiple factors including range, accuracy and appropriateness of grammar and vocabulary, coherence, extent and relevance of a speaker's contribution (discourse management); ability to produce comprehensible utterances, in terms of stress and rhythm intonation and individual sounds; and ability to use interactive strategies to achieve meaningful communication. The Speaking test was extensively revised between 1999 and 2001. An overview of the validation process undertaken during this redesign has been published by Cambridge ESOL (Taylor, 2001).

Assessment of writing performance involves considering a number of criteria. Accuracy, including spelling and punctuation remains relevant, but content, organization and cohesion, range of structures and vocabulary, register and format, and effect on the target reader are equally salient features of performance to be taken into account. The writing assessment criteria were extensively revised in 2002 (Bridges and Shaw, 2004).

Recommendations

As the population of examinees increases for testing programs that are international in scope, decisions made about examinees' performance may be subject to greater scrutiny about factors that are not construct relevant. Reliance on professional standards such as the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) and the ITC (2000) *Guidelines* can serve as a framework for

practitioners as they consider factors to make their tests more universally interpretable. Because many testing programs use cut scores to assist with decisions about admissions, licensure, or certification, the methods used to recommend these cut scores need to be aware of factors that can influence the results.

We offer three recommendations to practitioners when conducting standard setting studies for testing programs that expect some variation in their representation of the construct when applied to an international population.

- Select panel members who are representative of the variation in the linguistic and/or cultural context and have familiarity with the variation in the anticipated examinee population.
- Inform panel members of test adaptation approaches in the development, administration, and scoring so that they can appropriately consider these in their judgments about expected candidate performance.
- For standard setting methods that rely on anchor performances or exemplars, the performances should reflect the anticipated range of score points and also the range of anticipated variation in the construct.

By considering these recommendations, researchers may be better able to represent some of the core elements observed in standard setting literature. Specifically, selecting panelists who are both familiar with the content measured on the test and the intended population of examinees insures that those people making the recommendations are most qualified to do so. Second, the standard setting methodology should consider the interaction of the abilities of target examinees (e.g., minimally competent candidate) and the difficulty of the items or tasks whether panelists' judgments are made on test characteristics or examinee characteristics. Finally, any examinee performances that are

used to illustrate different score points or variations within or across score points consider the representation of those score points in the anticipated population of examinees.

Conclusions

This paper described factors that were considered in a standard setting study that included selecting anchor performances on a test of English literacy that is being used as a component of immigration eligibility. The sampling of the performances selected and other considerations were discussed in the context of the *Guidelines for Test Adaptation and Translation* (International Test Commission, 2000). These considerations included factors that may influence selection of panelists, the representation of candidate performances, and representation of the content. Additional studies are needed in this area to determine the extent to which adaptation methodologies should be considered within the context of a language as well as continued research in this topic across languages.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). Standards for Educational and Psychological Testing. Washington, DC: American Educational Research Association.
- Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement, 2nd Edition, Washington, DC: American Council on Education.
- Bridges, G. & Shaw, S. (2004). IELTS Writing: revising assessment criteria and scales (Phase 4). Cambridge ESOL Research Notes, 18, 8-12.
- Cambridge ESOL, (2004). IELTS – some frequently asked questions. Cambridge ESOL Research Notes, 18, 14-17.
- Impara, J. C. & Plake, B. S. (1997). An alternative approach to standard setting. Journal of Educational Measurement, 34(4), 355-368.
- International English Language Testing System (September, 2003). Handbook. Cambridge, U.K.: University of Cambridge ESOL.
- International Test Commission (2000). ITC Test Adaptation Guidelines. Author. Retrieved March 17, 2006 from <http://www.intestcom.org>.
- Plake, B. S., & Hambleton, R. K. (2000). A standard-setting method designed for complex performance assessments: Categorical assignments of student work. Educational Assessment, 6(3), 197-215.
- Smith, J. (2004). IELTS Impact: a study on the accessibility of IELTS GT Modules to 16-17 year old candidates. Cambridge ESOL Research Notes, 18, 5-8.

Taylor, L. (2001). Revising the IELTS Speaking Test: developments in test format and task design. Cambridge ESOL Research Notes, 5, 3-5.

Taylor, L. (2006). The changing landscape of English: implications for language assessment. ELT Journal, 60(1), 51-60.

Table 1. Application of ITC Test Adaptation Guidelines (2000) to Standard Setting Studies.

Selected ITC Guidelines	Considerations for Standard Setting
C.1 Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible.	Panelist selection, illustrative candidate performances, and content representation.
D.1 Test developers/publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the instrument are intended.	Inform panelists of the strategies that the developer has used to attend to differences among intended populations in representing the construct.
D.2 Test developers/publishers should provide evidence that the language use in the directions, rubrics, and items themselves as well as in the handbook are appropriate for all cultural and language populations for whom the instrument is intended.	Inform panelists of the directions, rubrics, scoring guidelines, and items/tasks for awareness of cultural or linguistic sensitivity.
A.2 Test administrators should be sensitive to a number of factors related to the stimulus materials, administration procedures, and response modes that can moderate the validity of the inferences drawn from the scores.	Inform panelists of the administration directions and procedures that may mitigate their judgments about expected performance among target candidates.
I.4 The test developer should provide specific information on the ways in which the socio-cultural and ecological contexts of the populations might affect performance on the test, and should suggest procedures to account for these effects in the interpretation of results.	Provide documentation for panelists to describe how population characteristics may impact score interpretation and allow them to consider these in their judgments.