Cambridge English

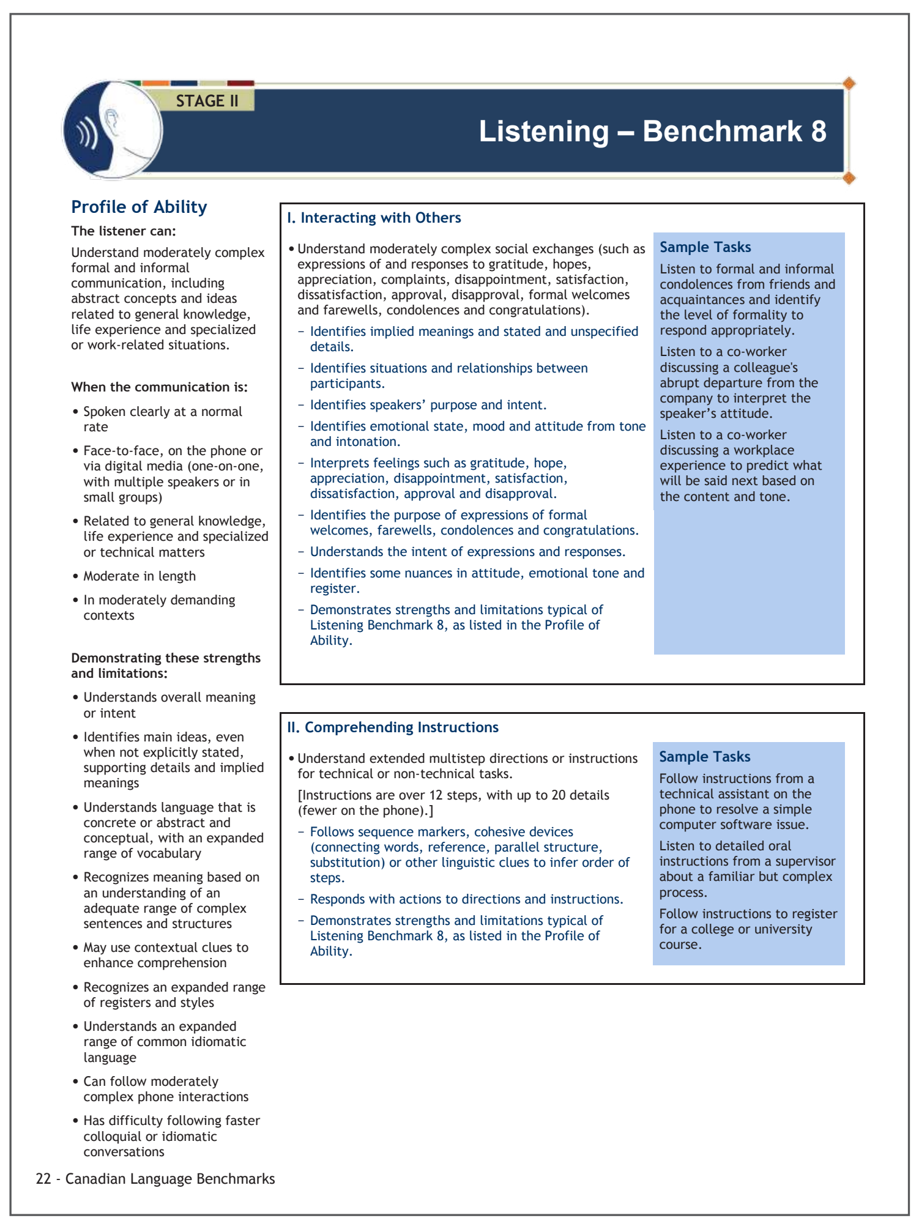
Multiple Standard Setting Outcomes: Empirical Exercises Informing Theory

How Should We Think About Divergent Standard Setting Outcomes?

It has been observed that standard setting sometimes results in different outcomes. How should we think about such results?

Standard setting theory argues that standard setting is not an exercise in homing in on a "true", objectively-existing cut score, but a values-driven enterprise arrived at for some policy-related end (Camilli, Cizek & Lugg, 2001; Cizek, 1993; Kane, 1998; Zieky, 2001), so disparate outcomes should be acceptable.

However, standard setting is sometimes concerned with finding levels on standards, frameworks and benchmarks with objective criteria, e.g. the Common European Framework of Reference, the Canadian Language Benchmarks (Figures 1 and 2). Under these circumstances, standard setting should be seen as expert informed but data constrained exercises where convergent outcomes should be expected.



STAGE II

Listening – Benchmark 8

Profile of Ability

The listener can:

Understand moderately complex formal and informal communication, including abstract concepts and ideas related to general knowledge, life experience and specialized or work-related situations.

When the communication is:

- Spoken clearly at a normal rate
- Face-to-face, on the phone or via digital media (one-on-one, with multiple speakers or in small groups)
- Related to general knowledge, life experience and specialized or technical matters
- Moderate in length
- In moderately demanding contexts

Demonstrating these strengths and limitations:

- Understands overall meaning or intent
- Identifies main ideas, even when not explicitly stated, supporting details and implied meanings
- Understands language that is concrete or abstract and conceptual, with an expanded range of vocabulary
- Recognizes meaning based on an understanding of an adequate range of complex sentences and structures
- May use contextual clues to enhance comprehension
- Recognizes an expanded range of registers and styles
- Understands an expanded range of common idiomatic language
- Can follow moderately complex phone interactions
- Has difficulty following faster colloquial or idiomatic conversations

I. Interacting with Others

- Understand moderately complex social exchanges (such as expressions of and responses to gratitude, hopes, appreciation, complaints, disappointment, satisfaction, dissatisfaction, approval, disapproval, formal welcomes and farewells, condolences and congratulations).
  - – Identifies implied meanings and stated and unspecified details.
  - – Identifies situations and relationships between participants.
  - – Identifies speakers' purpose and intent.
  - – Identifies emotional state, mood and attitude from tone and intonation.
  - – Interprets feelings such as gratitude, hope, appreciation, disappointment, satisfaction, dissatisfaction, approval and disapproval.
  - – Identifies the purpose of expressions of formal welcomes, farewells, condolences and congratulations.
  - – Understands the intent of expressions and responses.
  - – Identifies some nuances in attitude, emotional tone and register.
  - – Demonstrates strengths and limitations typical of Listening Benchmark 8, as listed in the Profile of Ability.

Sample Tasks

Listen to formal and informal condolences from friends and acquaintances and identify the level of formality to respond appropriately.

Listen to a co-worker discussing a colleague's abrupt departure from the company to interpret the speaker's attitude.

Listen to a co-worker discussing a workplace experience to predict what will be said next based on the content and tone.

II. Comprehending Instructions

- Understand extended multistep directions or instructions for technical or non-technical tasks.

  [Instructions are over 12 steps, with up to 20 details (fewer on the phone).]
  - – Follows sequence markers, cohesive devices (connecting words, reference, parallel structure, substitution) or other linguistic clues to infer order of steps.
  - – Responds with actions to directions and instructions.
  - – Demonstrates strengths and limitations typical of Listening Benchmark 8, as listed in the Profile of Ability.

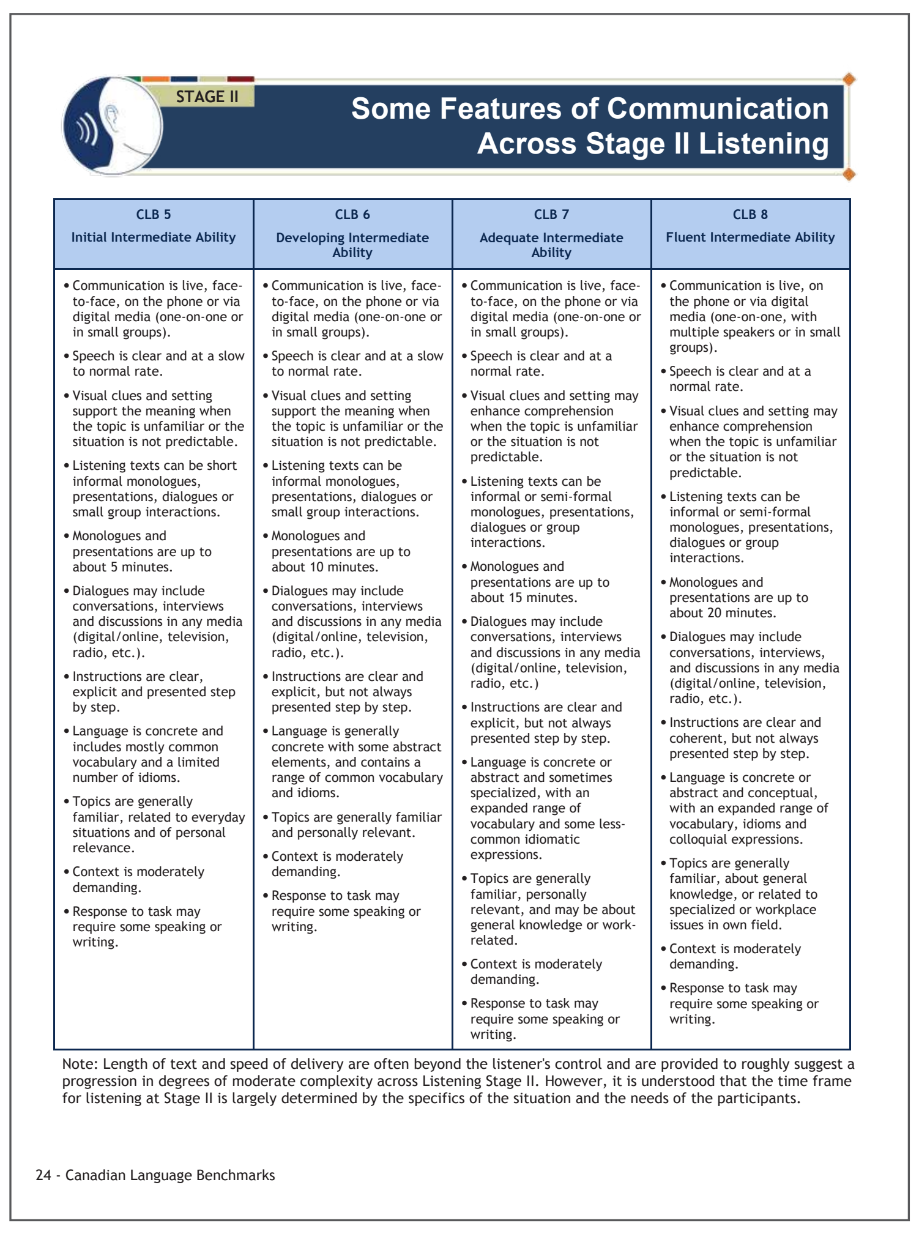
Sample Tasks

Follow instructions from a technical assistant on the phone to resolve a simple computer software issue.

Listen to detailed oral instructions from a supervisor about a familiar but complex process.

Follow instructions to register for a college or university course.

22 - Canadian Language Benchmarks



STAGE II

Some Features of Communication Across Stage II Listening

CLB 5 Initial Intermediate Ability	CLB 6 Developing Intermediate Ability	CLB 7 Adequate Intermediate Ability	CLB 8 Fluent Intermediate Ability
• Communication is live, face-to-face, on the phone or via digital media (one-on-one or in small groups). • Speech is clear and at a slow to normal rate. • Visual clues and setting support the meaning when the topic is unfamiliar or the situation is not predictable. • Listening texts can be short informal monologues, presentations, dialogues or small group interactions. • Monologues and presentations are up to about 5 minutes. • Dialogues may include conversations, interviews and discussions in any media (digital/online, television, radio, etc.). • Instructions are clear, explicit and presented step by step. • Language is concrete and includes mostly common vocabulary and a limited number of idioms. • Topics are generally familiar, related to everyday situations and of personal relevance. • Context is moderately demanding. • Response to task may require some speaking or writing.	• Communication is live, face-to-face, on the phone or via digital media (one-on-one or in small groups). • Speech is clear and at a slow to normal rate. • Visual clues and setting support the meaning when the topic is unfamiliar or the situation is not predictable. • Listening texts can be informal monologues, presentations, dialogues or small group interactions. • Monologues and presentations are up to about 10 minutes. • Dialogues may include conversations, interviews and discussions in any media (digital/online, television, radio, etc.). • Instructions are clear and explicit, but not always presented step by step. • Language is generally concrete with some abstract elements, and contains a range of common vocabulary and idioms. • Topics are generally familiar and personally relevant. • Context is moderately demanding. • Response to task may require some speaking or writing.	• Communication is live, face-to-face, on the phone or via digital media (one-on-one or in small groups). • Speech is clear and at a normal rate. • Visual clues and setting may enhance comprehension when the topic is unfamiliar or the situation is not predictable. • Listening texts can be informal or semi-formal monologues, presentations, dialogues or group interactions. • Monologues and presentations are up to about 15 minutes. • Dialogues may include conversations, interviews and discussions in any media (digital/online, television, radio, etc.) • Instructions are clear and explicit, but not always presented step by step. • Language is concrete or abstract and sometimes specialized, with an expanded range of vocabulary and some less-common idiomatic expressions. • Topics are generally familiar, personally relevant, and may be about general knowledge or work-related. • Context is moderately demanding. • Response to task may require some speaking or writing.	• Communication is live, on the phone or via digital media (one-on-one, with multiple speakers or in small groups). • Speech is clear and at a normal rate. • Visual clues and setting may enhance comprehension when the topic is unfamiliar or the situation is not predictable. • Listening texts can be informal or semi-formal monologues, presentations, dialogues or group interactions. • Monologues and presentations are up to about 20 minutes. • Dialogues may include conversations, interviews, and discussions in any media (digital/online, television, radio, etc.). • Instructions are clear and coherent, but not always presented step by step. • Language is concrete or abstract and conceptual, with an expanded range of vocabulary, idioms and colloquial expressions. • Topics are generally familiar, about general knowledge, or related to specialized or workplace issues in own field. • Context is moderately demanding. • Response to task may require some speaking or writing.

Note: Length of text and speed of delivery are often beyond the listener's control and are provided to roughly suggest a progression in degrees of moderate complexity across Listening Stage II. However, it is understood that the time frame for listening at Stage II is largely determined by the specifics of the situation and the needs of the participants.

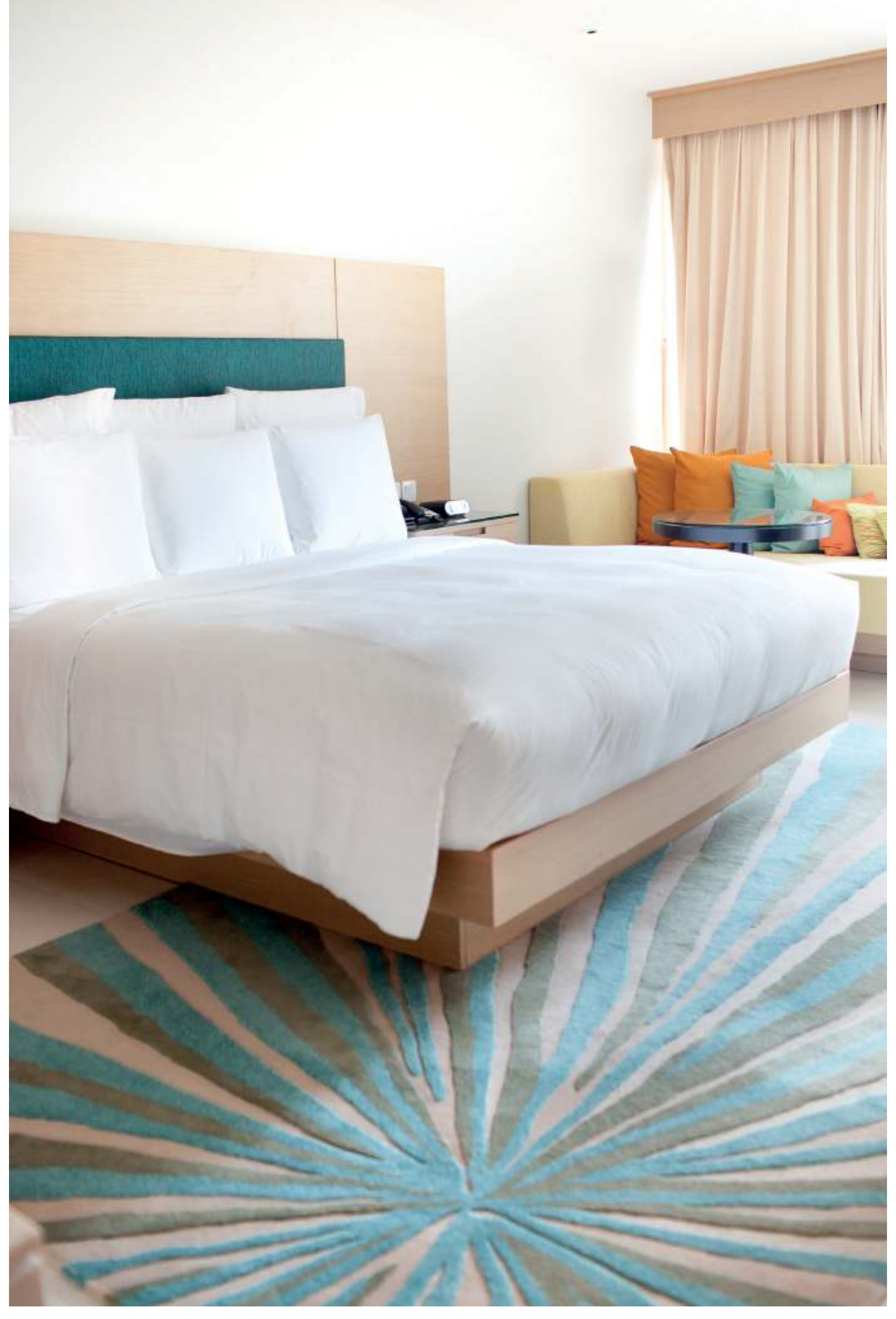
24 - Canadian Language Benchmarks

An Analogy

If the question about hotels is about "good enough", people can be expected to disagree and that is all right. On the other hand, if the question is about the star rating for a hotel, the objective criteria limit the correct answer that the informant can arrive at.

Is this hotel good enough for us to stay in on our trip?

What is the class / star rating for this hotel?



Stars	Class	Excerpt of Criteria
★	**Tourist**	• All rooms with table and chair • All rooms with colour TV and remote control • All rooms with shower/WC or bath tub/WC
★★	**Standard**	**In addition to ★:** • Linen shelves • Bath essence of shower gel • Bath towels • Credit cards accepted
★★★	**Comfort**	**In addition to ★★:** • Place to put luggage/suitcase • Dressing mirror • Telephone in room
★★★★	**First Class**	**In addition to ★★★:** • Upholstered chair/couch with side table • Minibar or 24 hour beverages via room service • Cosmetic products, bath robe and slippers on demand
★★★★★	**Luxury**	**In addition to ★★★★:** • Fresh flowers or a present in the room • Internet PC in the room • Safe in the room • Turndown service

The Argument

When multiple standard setting exercises are conducted in relation to criterion standards, convergent outcomes should be expected and divergent outcomes investigated. This makes standard setting outcomes subject to validation/falsification, making them more useful, and leading to better supported validity arguments. They should be seen as repeated measures (with a certain measurement error attached to them) which together leads to a more accurate outcome.

Empirical Illustration 1

Standard setting exercises were conducted in 2005 and 2013 relating IELTS, a test of English language proficiency, to the Canadian Language Benchmarks. The representative excerpt in Table 1 shows that results generally converge, and where they do not, the standard error informs what the correct cut score should be.

Table 1. *IELTS Band Scores Relating to CLB Benchmark 7*

		-1 SE	Ave	+1 SE
Listening	2005	6.0	6.0	6.5
	2013	5.5	6.0	6.0
Reading	2005	5.5	6.0	6.5
	2013	5.5	5.5	6.0

		-1 SE	Ave	+1 SE
Speaking	2005	5.5	6.0	6.5
	2013	6.0	6.0	6.5
Writing	2005	5.5	6.0	6.5
	2013	5.5	6.0	6.0

Empirical Illustration 2

Table 2 shows four studies (A: Pearson, 2009; B: Tannenbaum & Wiley, 2004, ETS, 2006, 2010; C: Buckendahl & Khalifa, 2009; D: Lim, Geranpayeh, Khalifa & Buckendahl, 2013) relating IELTS to level C1 on the Common European Framework of Reference appear to be divergent, making them unhelpful.

However, further thought about the studies reveal the following:

- Study A related IELTS to the CEFR via another test, asking test takers to sit both exams. But data did not account for test-takers' motivation, (participants took IELTS test for high-stakes purpose, took non-bearing field test of other exam), providing an inaccurate picture of the relationship between the two exams.
- Study C was influenced by standard setting method effects (Jaeger, 1989). The Analytic Judgment method (Plake & Hambleton, 2000) was used for the Writing and Speaking tests, where panellists select a performance deemed at a level, and the score for that is the cut score. IELTS reports wide band scores with half bands (4.0, 4.5, 5.0, 5.5…). Thus, if the true cut point is, say, 3.13, panellists would logically choose a 3.5 response, inflating the cut score. The Yes-No Tucker-Angoff (Angoff, 1971; Impara & Plake, 1997) was used for the discrete-point Reading and Listening tests, which produces overestimates at higher levels where items are easy for the target test taker (Cizek & Bunch, 2007; Council of Europe, 2009).
- After divergent outcomes are investigated and cut scores validated, a clearer picture emerges (Table 3), and results become much more useful.

Table 2. *IELTS Band Scores Relating to CEFR Level C1, Before Validation*

Study	C1 Cut Score
A	7.5
B	~6.6
C	~7.3
D	~6.6

Table 3. *IELTS Band Scores Relating to CEFR Level C1, After Validation*

Study	C1 Cut Score
A	~~7.5~~
B	~6.6
C	~6.8
D	~6.6

Conclusion

Situations exist where multiple standard setting exercises should produce convergent outcomes. Standard setting theory needs to account for these situations involving objective criteria, seeing them as expert informed but data constrained exercises subject to validation and falsification, leading to better supported validity arguments. Examples from empirical studies illustrated approaches for validating cut scores.

Authors

Gad S Lim, Cambridge English Language Assessment

Chad W Buckendahl, Alpine Testing Solutions

CAMBRIDGE ENGLISH
Language Assessment
Part of the University of Cambridge

Alpine
Testing Solutions