Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies

Spiros Papageorgiou
Donald E. Powers
Mary Schedl

Presented at the 14th EALTA Conference, Paris, June 2, 2017

Numeric scores v. performance levels

- Language tests are used to facilitate various decisions
- Results should be communicated in a transparent and meaningful way (AERA, APA, NCME, 2014)
- Numeric scores alone may convey limited information about what test takers know and can do
- Researchers in educational measurement advocate an "achievement performance-level narrative approach" (Ryan, 2006)

External v. internal levels and descriptors

- Linking to external levels and descriptors (e.g., through score mapping)
  - Examples: CEFR, ACTFL, CLB
  - Advantages: scores useful when these levels are meaningful to the community and relevant to the test construct
  - Disadvantages: information not always relevant to the test; they are typically generic; they suffer from "descriptional inadequacy" (Fulcher, Davidson & Kemp, 2011)
- Develop internal levels and descriptors (e.g., through scale anchoring)
  - Examples: *TOEFL ITP®* Test Score Descriptors
  - Advantages: offer performance information directly relevant to the test construct
  - Disadvantages: performance information is primarily meaningful to those familiar with test construct and content
- *Could a combination of both types of levels provide more transparent and useful information?*

Context of this study

- *TOEFL ITP®* test
  - Member of the TOEFL Family of Assessments (along with *TOEFL® Primary™* , *TOEFL Junior®*, *TOEFL iBT®*)
  - Paper-delivered
  - Primarily used for institutional purposes
  - 3 test sections: Listening comprehension (50 items), Structure and written expression (40 items), Reading comprehension (50 items)
  - Intended for intermediate/high intermediate students using English in an academic context

Combining two approaches to facilitate score interpretation

- Previous score mapping study linked *TOEFL ITP* scores to four CEFR levels (Tannenbaum & Baron, 2011)
- This study aimed to enhance score interpretation through scale anchoring
  - test items ("scale anchors") that characterize different score levels were identified
  - score levels were defined based on the results of the previous score mapping study
  - descriptors developed based on the scale anchors for each level

Overall procedure for scale anchoring

- Select test forms
- Define number of reported levels
- Select anchor items
- Characterize the skills and knowledge measured by potential anchor items
- Create, review and edit performance descriptors

Test forms

- 5 equated test forms
- Administered between 2012-2014
- 41,606 test takers (6,694-11,542 per test form)
- Multiple test forms selected to explore the generalizability of the descriptors
- Total mean scores for each test form were very similar
- Reliability estimates for the total test ranged between .94 and .95

How many performance levels?

- Procedures for determining the number of reported levels varies across scale anchoring studies (e.g., Garcia Gomez et al., 2007)
- 4 reporting levels selected for this study
  - associated with CEFR levels A2-C1 from the previous score mapping study
  - number of levels seemed sufficient for the score scale
  - each level was wide enough to enable identification of a sufficient number of anchor items

The performance levels of the study

Level	Listening Comprehension (31-68)	Structure & Written Expression (31-68)	Reading Comprehension (31-67)
Level 4 (C1)	64-68	64-68	63-67
Level 3 (B2)	54-63	53-63	56-62
Level 2 (B1)	47-53	43-52	48-55
Level 1 (A2)	38-46	32-42	31-47

Test takers

Level	Listening Comprehension	Structure & Written Expression	Reading Comprehension
Level 4 (C1)	911	1,309	1,139
Level 3 (B2)	11,827	8,628	5,935
Level 2 (B1)	16,179	18,887	16,201
Level 1 (A2)	12,689	12,782	18,331

Criteria for anchor items

- Criteria in scale anchoring studies tend to vary (Beaton & Allen, 1992; Garcia Gomez et al., 2007)
- General criteria in the literature
  - proportion of test takers within the band level who correctly answered the item is above a certain minimum
  - proportion of test takers at adjacent next lower band level is substantially less
- This study
  - At least 80% of test takers at the level answered the item correctly
  - Fewer than 60% of test takers at the next lower level answered the same item correctly.
  - For the lowest level items answered correctly by a simple majority were selected
  - 15 items were selected as anchors at each level

Summary statistics for anchor items

Level	Median (Range)	Listening Comprehension	Structure & Written Expression	Reading Comprehension
Level 4 (C1)	Proportion correct	.95 (.81 to .98)	.90 (.80 to .98)	.87 (.78 to .96)
	Difference from next lower level	.38 (.34 to .53)	.40 (.35 to .49)	.29 (.26 to .37)
Level 3 (B2)	Proportion correct	.85 (.80 to .90)	.84 (.80 to .93)	.83 (.77 to .91)
	Difference from next lower level	.34 (.32 to .39)	.37 (.35 to .43)	.33 (.31 to .38)
Level 2 (B1)	Proportion correct	.85 (.77 to .93)	.85 (.79 to .94)	.85 (.78 to .91)
	Difference from next lower level	.38 (.30 to .46)	.36 (.32 to .47)	.38 (.36 to .41)
Level 1 (A2)	Proportion correct	.57 (.52 to .72)	.58 (.53 to .77)	.61 (.54 to .76)
	Difference from next lower level	NA	NA	NA

Crafting the performance descriptors

- Three-person team of experienced assessment developers for each of the three test sections
- Teams examined each anchor item and associated reading and listening stimuli
- Statistical information about the items was available to the teams
- Teams considered the language skills and abilities test takers most likely needed to answer each anchor item correctly
- Repeated instances of skills and abilities for a particular level were noted and included in the descriptors of each level

Example of low level item

Passage: "during the early days of the reproductive cycle"

According to the passage, which of the following activities is characteristic of the early part of the reproductive cycle of birds?

(A) Selecting a mate

(B) Collecting nest-building materials

(C) Playing with nest-building materials

(D) Building a nest

Passage: "to play with the building materials"

All of the lowest level items had item-stem-to-passage-text matches, making it relatively easy to locate the information requested

From anchor item to descriptor for low level

Test takers at this level are usually able to:

- understand the main idea of some texts in which the idea is reinforced by the repetition of important vocabulary across many sentences
- follow simple sentence references (e.g., “it,” “they”) to determine the grammatical referent of a pronoun
- locate requested information in some sentences if pointed directly to the part of the passage containing the information (e.g., “in line x,” “in paragraph y”)

Example of high level item

Passage: "Because standard music notation makes no provision for many of these innovations, recent music scores may contain graph-like diagrams, new note shapes and symbols, and novel ways of arranging notation on the page."

In the third paragraph, the author mentions **diagrams** as an example of a new way to

(A) chart the history of innovation in musical notation
(B) explain the logic of standard musical notation
(C) design and develop electronic instruments
(D) indicate how particular sounds should be produced

- ***Test takers must understand from the discussion in the previous paragraph why the diagrams are required.***
- ***The sophisticated vocabulary and the density of information in the second paragraph make understanding this information challenging***.

- **no text match for the answer**
- **answer not in the immediate area of the stem**

From anchor item to descriptor for high level

Test takers at this level are usually able to:

- Follow discourse at the idea level to understand detailed information and major ideas, both explicitly stated and implied, even when
  - texts contain an accumulation of low-frequency academic vocabulary
  - comparisons and contrasts, causal relationships, illustrations, etc. are not explicitly stated or indicated by discourse markers
  - texts are on abstract topics, such as music composition and computer animation

External v. internal descriptors (Reading)

Level	External descriptor (CEFR)	Internal descriptors (item anchors)
Level 3 (B2)	Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.	Can process information across typical academic texts to understand detailed information and major ideas both explicitly stated and implied, when texts contain high-frequency academic vocabulary and typical academic discourse markers.

Discussion

- Empirically-derived, test-specific levels and descriptors were intended to complement information provided in the more generic CEFR descriptors
- The score mapping study helped decide on the number of reporting levels for the subsequent scale anchoring study
- The combination of such methodologies might facilitate interpretations of test scores
- Involvement of teachers, students and assessment developers might offer additional insights into the meaningfulness and usefulness of these descriptors

More details

- Descriptors available at https://www.ets.org/toefl_itp/research

TOEFL ITP® Test Score Descriptors

Reading Comprehension

TOEFL ITP® Section Scores	CEFR Level	Proficiency Descriptors
63–67	C1	**Test takers at this level are usually able to:** • Follow discourse at the idea level to understand detailed information and major ideas, both explicitly stated and implied, even when: – texts contain an accumulation of low-frequency academic vocabulary – comparisons and contrasts, causal relationships, illustrations, etc. are not explicitly stated or indicated by discourse markers – texts are on abstract topics, such as music composition and computer animation
56–62	B2	**Test takers at this level are usually able to:** • Process information across typical academic texts to understand detailed information and major ideas, both explicitly stated and implied, when texts: – contain high-frequency academic vocabulary and typical academic discourse markers – are on concrete topics that discuss the physical and social sciences (e.g., glacier formation, moon terrain, theories of child development)
48–55	B1	**Test takers at this level are usually able to:** • understand descriptions of relatively simple processes and narration in well-marked academic texts • understand high-frequency vocabulary and recognize paraphrased information • follow sentence-level comparisons and contrasts and understand meaning conveyed by the most common conjunctions, such as "and," "or" and "but" • connect meaning across some simple sentences that contain high-frequency vocabulary
31–47	A2	**Test takers at this level are sometimes able to:** • understand the general idea of some sentences that use simple, everyday vocabulary • understand the main idea of some texts in which the idea is reinforced by the repetition of important vocabulary across many sentences • follow simple sentence references (e.g., "it," "they") to determine the grammatical referent of a pronoun • locate requested information in some sentences if pointed directly to the part of the passage containing the information (e.g., "in line x," "in paragraph y")

More details

- Powers, D., Schedl, M., & Papageorgiou, S. (2017). Facilitating the interpretation of English language proficiency scores: Combining scale anchoring and test score mapping methodologies. *Language Testing, 34*(2), 175-195.

Questions?

spapageorgiou@ets.org