Designing, evaluating and improving language tests

Henry Emery
Designing, evaluating and improving language tests
Designing, evaluating and improving language tests

Diagram:

1. Design
2. Pilot
3. Investigate
4. Report

Design → Pilot → Investigate → Report → Design
Design

➢ Who are we testing?
➢ Do our tasks together represent:
  ▪ On-the-job language use?
  ▪ The six criteria?
  ▪ The six levels?
➢ How is the test assessed?
- ground, approach, tower and en-route ATC
- flight crew, cabin crew and passengers
- maintenance personnel
- operations and dispatch
- tug drivers, de-icing crews etc
- taxi drivers, hotel staff...
- simulator instructors and examiners
- *language testers?*
Designing, evaluating and improving language tests

On-the-job language use tasks

Test tasks
Subject matter knowledge  

Language knowledge

---

*Designing, evaluating and improving language tests*
Subject matter knowledge

Scenario:
EFOT, AC climbing to 5,000 ft. Intend to return to airfield.

Task:
Dump fuel
C: 321 heavy. And will you be dumping fuel?
P: **321 heavy we will have... yes... maybe three minutes’ dumping**
C: OK, do you want to do that over the ocean right now that would be probably the best bet
P: *That is affirmative. I’ll call you back when starting if it is possible on that heading*
C: 321 heavy. If possible fly heading one eight zero and I’ll just keep you over the water while you do the dump
P: **OK, maintain ... heading one eight zero call you back ending... uh... dumping**
Designing, evaluating and improving language tests

Subject matter knowledge

Language knowledge

Test tasks
<table>
<thead>
<tr>
<th>Level</th>
<th>Pronunciation</th>
<th>Structure</th>
<th>Vocabulary</th>
<th>Fluency</th>
<th>Comprehension</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 5</td>
<td>Pronunciation, stress, rhythm, and intonation, though possibly influenced by the first language, are characterized by a variety of familiar and unfamiliar topics, and speech is usually of moderate length and controlled.</td>
<td>Both basic and complex grammatical structures and sentence patterns are present in the text.</td>
<td>Vocabulary range and accuracy are sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics.</td>
<td>Able to speak at length with a natural, effortless flow.</td>
<td>Comprehension is consistently accurate in nearly all situations. Interacts with ease in personal and non-social contexts and responds to them appropriately.</td>
<td></td>
</tr>
<tr>
<td>Extended 4</td>
<td>Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, are characterized by a variety of familiar and unfamiliar topics, and speech is usually of moderate length and controlled.</td>
<td>Basic grammatical structures and sentence patterns are present in the text.</td>
<td>Vocabulary range and accuracy are usually sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics.</td>
<td>Able to speak at length with relative ease on familiar topics, and may not vary speech flow for a stylistic effect.</td>
<td>Comprehension is usually immediate, appropriate, and interactive. Can initiate and maintain exchanges with minimal personalization.</td>
<td></td>
</tr>
<tr>
<td>Operational 3</td>
<td>Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, are characterized by a variety of familiar and unfamiliar topics, and speech is usually of moderate length and controlled.</td>
<td>Basic grammatical structures and sentence patterns are present in the text.</td>
<td>Vocabulary range and accuracy are usually sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics.</td>
<td>Able to speak at length with relative ease on familiar topics, and may not vary speech flow for a stylistic effect.</td>
<td>Comprehension is usually immediate, appropriate, and interactive. Can initiate and maintain exchanges with minimal personalization.</td>
<td></td>
</tr>
<tr>
<td>Pre-Operational 2</td>
<td>Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, are characterized by a variety of familiar and unfamiliar topics, and speech is usually of moderate length and controlled.</td>
<td>Basic grammatical structures and sentence patterns are present in the text.</td>
<td>Vocabulary range and accuracy are usually sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics.</td>
<td>Able to speak at length with relative ease on familiar topics, and may not vary speech flow for a stylistic effect.</td>
<td>Comprehension is usually immediate, appropriate, and interactive. Can initiate and maintain exchanges with minimal personalization.</td>
<td></td>
</tr>
<tr>
<td>Primary 1</td>
<td>Pronunciation, stress, rhythm, and intonation, though influenced by the first language or regional variation, are characterized by a variety of familiar and unfamiliar topics, and speech is usually of moderate length and controlled.</td>
<td>Basic grammatical structures and sentence patterns are present in the text.</td>
<td>Vocabulary range and accuracy are usually sufficient to communicate effectively on a wide variety of familiar and unfamiliar topics.</td>
<td>Able to speak at length with relative ease on familiar topics, and may not vary speech flow for a stylistic effect.</td>
<td>Comprehension is usually immediate, appropriate, and interactive. Can initiate and maintain exchanges with minimal personalization.</td>
<td></td>
</tr>
</tbody>
</table>

...a wide variety of familiar and unfamiliar topics...

...transition from rehearsed or formulaic speech to spontaneous interaction...

...comprehension of linguistic and cultural subtleties...

...is sensitive to non-verbal cues...

Designing, evaluating and improving language tests
### Designing, evaluating and improving language tests

<table>
<thead>
<tr>
<th>Level</th>
<th>Pronunciation</th>
<th>Structure</th>
<th>Vocabulary</th>
<th>Fluency</th>
<th>Comprehension</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert 6</td>
<td>Pronunciation, stress, rhythm, and intonation,</td>
<td>Both basic and complex grammatical structures and</td>
<td>Vocabulary range and accuracy are</td>
<td>Able to speak at length with a natural,</td>
<td>Comprehension is consistently accurate in</td>
<td>Interacts with ease in nearly all situations. Is</td>
</tr>
<tr>
<td></td>
<td>possibilities influenced by the first language or</td>
<td>sentence patterns are consistently well</td>
<td>sufficient to communicate effectively</td>
<td>effortless flow. Varies speech flow for</td>
<td>nearly all contexts and includes</td>
<td>sensitive to verbal and non-verbal cues, and</td>
</tr>
<tr>
<td></td>
<td>or regional variation, almost never interfere</td>
<td>controlled.</td>
<td>on a wide variety of familiar and</td>
<td>stylistic effect, e.g., to emphasize a</td>
<td>comprehension of linguistic and cultural</td>
<td>responds to them appropriately.</td>
</tr>
<tr>
<td></td>
<td>with ease of understanding.</td>
<td></td>
<td>unfamiliar topics. Vocabulary is</td>
<td>point. Uses appropriate discourse markers</td>
<td>subtleties.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>idiomatic, nuanced, and sensitive</td>
<td>and connectors spontaneously.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extended 5</td>
<td>Pronunciation, stress, rhythm, and intonation,</td>
<td>Basic grammatical structures and sentence</td>
<td>Vocabulary range and accuracy are</td>
<td>Able to speak at length with relative ease</td>
<td>Comprehension is accurate on common, concrete,</td>
<td>Responses are immediate, appropriate, and</td>
</tr>
<tr>
<td></td>
<td>possibilities influenced by the first language or</td>
<td>patterns are consistently well controlled.</td>
<td>sufficient to communicate effectively</td>
<td>on familiar topics, but may not vary</td>
<td>and work related topics and mostly accurate</td>
<td>informative. Manages the speaker/listener</td>
</tr>
<tr>
<td></td>
<td>or regional variation, rarely interfere with ease</td>
<td>Complex structures are attempted but with errors</td>
<td>on common, concrete, and work related</td>
<td>speech flow as a stylistic device. Can</td>
<td>when the speaker is confronted with a</td>
<td>relationship effectively.</td>
</tr>
<tr>
<td></td>
<td>of understanding.</td>
<td>which sometimes interfere with meaning.</td>
<td>topics. Paraphrases consistently</td>
<td>make use of appropriate discourse markers</td>
<td>linguistic or situational complication or an</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>and successfully. Vocabulary is</td>
<td>and connectors.</td>
<td>unexpected turn of events. Is able to</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>Pronunciation, stress, rhythm, and intonation are</td>
<td>Basic grammatical structures and sentence</td>
<td>Vocabulary range and accuracy are</td>
<td>Able to speak at length with</td>
<td>comprehend a range of speech varieties (dialect</td>
<td>Responses are usually immediate, appropriate,</td>
</tr>
<tr>
<td>4</td>
<td>influenced by the first language or regional</td>
<td>patterns are used creatively and are usually</td>
<td>sufficient to communicate effectively</td>
<td>relative ease on familiar or unexpected</td>
<td>and/or accent) or registers.</td>
<td>and informative. Initiates and maintains</td>
</tr>
<tr>
<td></td>
<td>variation but only sometimes interfere with ease</td>
<td>well controlled. Errors may occur, particularly</td>
<td>on common, concrete, and work related</td>
<td>turn of events, comprehension may be</td>
<td></td>
<td>exchanges even when dealing with an unexpected</td>
</tr>
<tr>
<td></td>
<td>of understanding.</td>
<td>in unusual or unexpected circumstances, but rarely</td>
<td>topics. Can often paraphrase</td>
<td>slower or require clarification strategies.</td>
<td></td>
<td>turn of events. Deals adequately with</td>
</tr>
<tr>
<td>Pre-</td>
<td>Pronunciation, stress, rhythm, and intonation are</td>
<td>Basic grammatical structures and sentence</td>
<td>Vocabulary range and accuracy are</td>
<td>Produces stretches of language at an</td>
<td>Comprehension is mostly accurate on common,</td>
<td></td>
</tr>
<tr>
<td>Operational</td>
<td>influenced by the first language or regional</td>
<td>patterns are used with predictable situations are</td>
<td>sufficient to communicate effectively</td>
<td>appropriate tempo. There may be occasional</td>
<td>concrete, and work related topics when the</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>variation and frequently interfere with ease of</td>
<td>not always well controlled. Errors frequently</td>
<td>on common, concrete, or work related</td>
<td>loss of fluency on transition from</td>
<td>accent or variety used is sufficiently</td>
<td></td>
</tr>
<tr>
<td></td>
<td>understanding.</td>
<td>interfere with meaning.</td>
<td>topics but range is limited and the</td>
<td>rehearsed or formal speech to spontaneous</td>
<td>intelligible for an international community of</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>word choice often inappropriate.</td>
<td>interaction, but this does not prevent</td>
<td>users. When the speaker is confronted with a</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>is often unable to paraphrase</td>
<td>effective communication. Can make</td>
<td>linguistic or situational complication or an</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>successfully when lacking</td>
<td>limited use of discourse markers or</td>
<td>unexpected turn of events. Is able to</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>vocabulary.</td>
<td>connectors.</td>
<td>comprehend a range of speech varieties (dialect</td>
<td></td>
</tr>
<tr>
<td>Elementary</td>
<td>Pronunciation, stress, rhythm, and intonation are</td>
<td>Basic grammatical structures and sentence</td>
<td>Vocabulary range and accuracy are</td>
<td>Produces stretches of language, but phrasing</td>
<td>and/or accent) or registers.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>heavily influenced by the first language or</td>
<td>patterns associated with predictable situations</td>
<td>sufficient to communicate</td>
<td>and pausing are often inappropriate.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>regional variation and usually interfere with</td>
<td>are not always well controlled. Errors</td>
<td>effectively on common, concrete, or</td>
<td>Habitations and slowness in language</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ease of understanding.</td>
<td>frequently interfere with meaning.</td>
<td>work related topics but range is</td>
<td>processing may prevent effective</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>limited and the word choice often</td>
<td>communication. Fillers are not</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pre</td>
<td>Performs at a level below the Elementary level.</td>
<td>Shows only limited control of a few simple</td>
<td>Limited vocabulary range consisting</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Elementary</td>
<td></td>
<td>memorized grammatical structures and sentence</td>
<td>only of isolated words and memorized</td>
<td>Comprehension is limited to isolated,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>patterns.</td>
<td>phrases.</td>
<td>memorized phrases when they are</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>carefully and slowly articulated.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note: The table above outlines the characteristics of language proficiency across different levels, including pronunciation, structure, vocabulary, fluency, and comprehension.*
Pilot

- Do the rubrics / instructions work?
- Do the tasks work?
- Does the timing work?
- Do the procedures work?
- Does the test flow?
Investigate

- How difficult are the tasks?
- Are the items (questions) / tasks working well?
- Are versions of the test equally difficult?
- Do the tasks / raters produce reliable scores?
- What do independent experts think?
- What do the candidates think?

Designing, evaluating and improving language tests
What improvements can we make?
Report

- Information for candidates
- Information for other stakeholders
- Evidence for test quality
- Plans for improvement

Designing, evaluating and improving language tests
Designing, evaluating and improving language tests
Are the systems and procedures working well?

Is the test producing consistent results?

Is the test discriminating at the various levels?

Are the versions equally difficult?

Is the test biased towards one group or another?
Designing, evaluating and improving language tests
Designing, evaluating and improving language tests
What cognitive processes do the candidates apply?

What cognitive processes do the raters apply?

What effect does the test have on language learning?

Does the test produce scores consistent with other validated tests?

Does the test predict on-the-job language performance?
Designing, evaluating and improving language tests
Of all the concepts in testing ... validity is the most basic and far-reaching, for without validity, a test ... and any inferences made from it are meaningless

Hubley & Zumbo 1996

Validity is not a characteristic of a test, but a feature of the inferences made on the basis of test scores and the uses to which a test is put

Alderson, 2002
Does it fly?

IS IT SAFE?
References:
