5th Annual EALTA conference, Athens, 9-11 May 2008

Exploring the Theoretical Basis for Developing Measurement Instruments on the CEFR

Gary Buck, University of Michigan
Spiros Papageorgiou, University of Michigan
Forest Platzek, Lidget Green, Inc.

LG LIDGET GREEN, INC.
Providing Assessment Services for Education

Overview

- Rationale
- Aims
- Design
- Results
- Implications

Rationale

- We need a set of guidelines for writing tests to measure ability at different CEFR levels.
- Some work has been done. E.g.
  - Dutch CEFR construct project (Alderson et al., 2006)
  - Proficiency scales (Generalitat de Catalunya, 2006)
  - CoE Manual (2003)
  - DIALANG (Alderson and Huhta, 2005)
- But none of these give explicit guidelines on how to write items to each CEFR level.
- This is what we address in this study.

Aims of the Study

- To develop a draft set of guidelines for writing items at each CEFR level.
- To explore how well these guidelines relate to empirical item difficulty.
- To compare the results of the CEFR item writing guidelines to the results of a set of independently developed set of ILR item writing guidelines, on the same set of items.

Design: Instrument Used

The Test of College Academic English (CAE)

- A p&p test developed by Lidget Green, Inc.
- Two forms of the test
- 135 Reading Items (N = 522)
- 124 Listening Items (N = 522)
- Designed to cover ILR levels 0+ to 3+
- Construct is General English set in a US college environment.

ILR Item Writing Guidelines for CAE

0+	**Minimal Functional Proficiency**	**Texts** are very simple; at the word level; often lists, or isolated words. **Tasks** require recognizing the most high-frequency words, usually nouns. Tasks do not require the user to relate that noun to a verb. **Topics** are very basic, immediate survival needs.
1	**Elementary Social Proficiency**	**Texts** are sentence level, with very basic grammar, and high-frequency vocabulary. Texts may consist of more than one sentence, but with only the loosest discourse structure. Sentences are joined by conjunctions, with no embedding or subordination—often the sentences can be reordered without affecting the message. **Tasks** require understanding simple sentences, and tasks, the user must understand how the subject and object relate to the verb. **Topics** tend to relate to immediate survival needs.
1+	**Developing Social Proficiency**	**Texts** are still at the sentence level, with little grammatical complexity, but the sentences tend to be more organized—structured paragraphs are starting to emerge. **Tasks** still largely consist of understanding simple and compound sentences, but may require understanding and connecting information from more than one sentence. **Topics** still tend to relate to basic needs, and are very concrete, and usually “in the moment.”
2	**Elementary Academic Proficiency**	**Texts** are longer, with complex sentences, and grammatical subordination. Vocabulary is mainly concrete but not necessarily simple. Texts may require understanding of temporal complexity, using past, present or future. Paragraph use is fully developed, and texts have a coherent structure that ‘tells a story’ or ‘makes a point.’ **Tasks** require users to understand coherent extended discourse, but the information is fairly concrete and is generally explicit and given in the text. **Topics** become much broader at this level, but are still rather concrete.

Some Assumptions

When relating person ability to task difficulty:

- A deterministic model is unreasonable
- We need a probabilistic model
- Therefore, we assume, by definition, that a person at any CEFR level would have a:
  - 75% probability of getting a reading item correct at that same level
  - 70% probability of getting a listening item correct at that level
  - Since readers can refer back to the text, but listeners can’t.
- This was the basis for judging items at each level

Design: Methodology

Part One: Developing Guidelines

Reading

- Step 1: Familiarization task
- Step 2: Independent rating of Form A by 3 researchers
- Step 3: Group discussion and consensus on Form A
- Step 4: Item writing guidelines drafted
- Step 5: Independent rating of Form B by 3 researchers
- Step 6: Group discussion and consensus on Form B
- Step 7: Item writing Guidelines revised, finalized.

Repeated for Listening

Methodology (cont.)

Part Two: Validation of Guidelines

Reading

- Step 1: Correlation with empirical difficulty
- Step 2: Comparison to ILR item writer guidelines

Repeated for Listening

Familiarization task: Percentage of On-target Ratings

CEFR / Raters	Exact level	± 1 level	± 2 levels	± 3 levels
Reading (k=56; n=3)	48.2%	46.4%	4.8%	0.6%
Listening (k=71; n=3)	57.7%	36.6%	5.6%	0%

Classification Agreement on Form A: Percentage of On-target Agreement

	A1	A2	B1	B2	C1
Reading (k=65)	87.5%	40%	82.1%	53.1%	66.7%
Listening (k=60)	58.1%	74.3%	87.5	50%	0%

Classification Agreement on Form B: Percentage of On-target Agreement

	A1	A2	B1	B2	C1
Reading (k=62)	81.8%	35.5%	67.3%	69.1%	30%
Listening (k=62)	83.8%	78.6%	80.1%	81.5%	71.4%

Mean Rasch Measure for Each CEFR Level

		A1	A2	B1	B2	C1
Reading	Mean S.D.	-1.89 1.29	-0.70 1.41	0.83 1.02	1.81 0.93	2.14 0.65
Listening	Mean S.D.	-1.71 1.35	-0.56 0.97	0.46 1.13	1.05 0.89	1.87 0.61

Correlations of Item Classifications with Rasch difficulty estimates

	Spearman's rho
Reading (k=127)	.790
Listening(k=122)	.636

Correlation of Item Classifications on the CEFR with ILR Scales

	Spearman's rho
Reading (k=127)	.903
Listening (k=122)	.863

Comparison of Reading Items on CEFR and ILR

ILR \ CEFR	A1	A2	B1	B2	C1
0+	41.9%	0.0%	0.0%	0.0%	0.0%
1	58.1%	47.4%	0.0%	0.0%	0.0%
1+	0.0%	42.1%	61.4%	3.8%	0.0%
2	0.0%	10.5%	29.5%	23.1%	0.0%
2+	0.0%	0.0%	9.1%	42.3%	28.6%
3	0.0%	0.0%	0.0%	30.8%	71.4%

Comparison of Listening Items on CEFR and ILR

ILR \ CEFR	A1	A2	B1	B2	C1
0+	50.0%	13.8%	0.0%	0.0%	0.0%
1	50.0%	62.1%	0.0%	0.0%	0.0%
1+	0.0%	24.1%	59.6%	4.3%	0.0%
2	0.0%	0.0%	21.3%	52.2%	0.0%
2+	0.0%	0.0%	19.1%	43.5%	100.0%

Sample Guidelines for Reading:

A1

Topics: survival level; immediate needs;

Texts: short; simple; clauses or short sentences; usually lacking any discourse structure; fixed expressions;

Tasks: Need to process only one word, or a short or fixed phrase; in picture recognition tasks, matching an explicitly stated word, or short or fixed phrase, with a clear picture; Simple word matching between text and stem/options;

General Advice: the simplest of written texts, or very simple realia would be suitable for this level;

Sample Guidelines for Listening:

B1

Topics: Familiar topics, such as work, school, leisure, family; Topics covered in beginning content classes; Mainly concrete; Some simpler abstract ideas;

Texts: Have a set of connected idea units; some embedding; at fairly normal speed or perhaps a little less, but not noticeably slow; more complex texts will have some accommodations; standard accent; Factual; Discourse features become important; idiomatic usage;

Tasks: Some main idea items may require synthesis of information scattered in different parts of the text; Scanning may be required; Recognizing a summary of information from different part of the text; Pragmatic inferencing skills may be required;

General: A large proportion of the texts, tasks and topics found at the beginning of an undergraduate degree course in a English speaking university would be acceptable material for these items, including lectures, seminars and spoken interaction typically encountered by undergraduates in their first year. Routine interpersonal skills are fairly well established;

Summary of Results

1. A set of draft guidelines for writing items to each CEFR level, in Reading and Listening.
2. A draft methodology for developing and validating such guidelines.
3. Some evidence of how the two leading sets of L2 proficiency descriptors relate to each other.

Implications and Limitations

- Clearly we expanded the CEFR, adding new concepts and new ideas to what it means to be at a certain level.
- The guidelines are still a work in progress. More research is necessary.

References

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly, 3*(1), 3–30.

Generalitat de Catalunya. (2006). *Proficiency scales: The Common European Framework of Reference for Languages in the Escoles Oficials d'Idiomes in Catalunya.* Madrid: Cambridge University Press.

Council of Europe. (2003). *Manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment. Preliminary pilot version.* Strasbourg: Council of Europe.

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment.* London: Continuum.

For DIALANG see

Alderson, J. C. (2007). The challenge of (diagnostic) testing: Do we know what we are measuring? In J. Fox & M. Wesche (Eds.), *Language testing reconsidered: Proceedings of the 27th Language Testing Research Colloquium (LTRC)* (pp. 21-39). Ottawa: University of Ottawa Press.

Alderson, J. C., & Huhta, A. (2005). The development of a suite of computer-based diagnostic tests based on the Common European Framework. *Language Testing, 22*(3), 301-320.

Huhta, A., & Figueras, N. (2004). Using the CEF to promote language learning through diagnostic testing. In K. Morrow (Ed.), *Insights from the Common European Framework* (pp. 65-76). Oxford: Oxford University Press.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S., & Teasdale, A. (2002). DIALANG: A diagnostic language assessment system for adult learners. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment-Case Studies* (pp. 130-146). Strasbourg: Council of Europe.

Contact Details

Gary Buck garybuck@umich.edu

Spiros Papageorgiou spapag@umich.edu

Forest Platzek forest@lidgetgreen.org