Combining language testing and second language acquisition research – insights from Project CEFLING

Riikka Alanen, Ari Huhta, Scott Jarvis, Maisa Martin & Mirja Tarnanen

Centre for Applied Language Studies & Department of Languages
University of Jyväskylä, Finland
Outline

• SLA and language testing
• CEFLING project as an example of combining SLA and language testing expertise
  • Design of tasks
  • Design of assessment procedures
  • Data analyses
• Discussion of some issues, dilemmas, contradictions and misunderstandings in SLA and language testing
SLA and language testing

• Traditionally, SLA and language testing as separate fields of research: history, functions, goals, constructs, methods (see e.g. Bachman & Cohen 1998)

• Search for ways of combining the research perspectives of language testing and second language acquisition around common interests
  • Byrnes 1987, Bachman & Cohen (eds.) 1998

• → L2 proficiency: development / emergence and variability
  • Across individuals
  • Across tasks
CEFLING Project – an example of combining SLA and language testing

- CEFLING (http://www.jyu.fi/cefling), funded by the Academy of Finland (2007-2009)
- Part of the SLATE network → Integrating SLA and language testing research perspectives
- Aim of the project is to find out which linguistic features, or combinations of features, characterise the Common European Framework (CEFR) levels in L2 Finnish and English
- Focus on L2 writing of both adult and young L2 learners of English and Finnish:
  - Adult language learners: test takers from the National Certificate examination system
  - Young language learners: pupils in grades 7-9 (ages 12-16), L2 English (n=250) and L2 Finnish (n=226)
Design of tasks and assessment procedures in CEFLING

In an SLA project, decisions need to be made, for example, about:

1) Instruments or tasks to be used in data gathering: what kind of tasks, how many, …?
   → task design / development / trialling

2) Assessment procedures and data analyses
   → in CEFLING, use was made of good language testing practice
Joint contribution of SLA and testing in CEFLING: Task design

- Operational definition adopted by CEFLING:
  - Task is “an activity which requires learners to use language, with emphasis on meaning, to attain an objective” (Bygate, Skehan & Swain, 2001, p.11).

- A set of communicative writing tasks designed to elicit L2 data from learners for both SLA research and assessment purposes
Tasks in CEFLING

- Variability in L2 performance across individuals: learners’ developing L2 proficiency ↔ rating scales
  - Task-based assessment:
    - “the process of evaluating, in relation to a set of explicitly stated criteria, the quality of the communicative performances elicited from learners as part of goal-directed, meaning-focused language use requiring the integration of skills and knowledge”. (Brindley, 1994/2009, p. 437)
    → Communicative adequacy of learner performance measured by using qualitative rating scales (Pallotti 2009)

- Variability in L2 performance across tasks:
  - L2 variation ↔ task difficulty (see Alanen, Huhta & Tarnanen, forthcoming)
Tasks in CEFLING 2

- Task design
  1. Authenticity: functions, text types and register taken into account and matched with already existing tasks used for assessing adult language learners’ proficiency (from the NC examination system data base)
    - From informal and formal email messages (a complaint to an Internet company) to argumentative and narrative texts
  2. Targeted at specific proficiency levels: from A1 - B2
- Tasks piloted and raters trained in the use of scales by using pilot data and existing sets of descriptors, benchmarks; new benchmarks created (see Alanen, Huhta, Tarnanen, forthcoming)
- Final set of tasks
Tasks in CEFLING 3

- Task 1: Informal email message to a friend
- Task 2: Informal email message to the teacher
- Task 3: Formal message to an Internet store
- Task 4: Opinion
- Task 5: Story
Task 5 Narrative


- Tell what happened (what, where, when, and so on).
- Tell why the experience was scary / funny / great.

Write in English in clear characters in the space below (continues on the reverse side).
Designing assessment procedures – This involves deciding on:

- How many raters? Their qualifications?
- What kind of rating scales?
- How much and what kind of rater training?
- Are benchmark performances available and how might they be used?
- What exactly should the rating process be like?
- Should ratings be monitored when they are in progress?
- How should the rating data be analysed? What kind of quality standards should the rating meet? (e.g. reliability)

= decisions that also language testing systems / projects need to make (cf. good language testing practice)
Assessment (and task design) procedures in CEFLING

NC test tasks
Draft tasks
Curriculum, textbooks
NC & CoE benchmarks
Piloting
Rater training
Rating
Scale design
NC & CoE benchmarks

Data gathering 1
Final tasks

Rating
Additional benchmarks

Rating
Rater self-training

Rating
Rater self-training

Analyses of ratings
Assigning writings to CEFR levels (approach 1, 2, 3, ...)

1 rater removed
3-4 raters / writing (10 in total)

SLA analysis of X, Y, Z, ... (1, 2, 3, ...)

Assigning writings to CEFR levels (approach 1, 2, 3, ...)

Data gathering 2
Analysis of ratings: from raw rating data to CEFR levels

Level 1 decisions: We need to decide how to convert the raw ratings (3-4 per each piece of writing) to CEFR level; when we have 2 or more raters and they are not in perfect agreement, what are the options?:
- calculate the mean of the 4 ratings?
- use the median (middlemost) rating?
- include only cases where all raters agree? something else?

Level 2 decisions: We can place learners on the CEFR levels (based on their performance on ALL writing tasks)
→ focus of study: learners and their writing skills
OR We can place individual pieces of writing, relating to a particular task on the CEFR levels
→ focus of study: tasks
From ratings to CEFR levels – our present approach

In CEFLING, we have so far focused on tasks (how learners perform on particular tasks), not on learners’ overall performance across tasks.

- So far, we have placed writings on CEFR levels on the basis of rater agreement.
  - English: **3 out of 4 raters had to agree** on the level.
Our present approach (3 of the 4 rater agree) means that we …

… have to reject some of our data:

- **Example 1**: ratings received by a piece of writing:
  A2 A2 B1 A2
  → included in our data

- **Example 2**: ratings received:
  A2 A2 B1 B1
  → excluded from our data!

→ for English, we were left with less than 70% of the data
The way we assign learners to CEFR levels – does it matter?

• Some empirical evidence from CEFLING on whether changing the way we place students on CEFR levels changes our findings about language development.

• An example:
  • Analyses of *vocabulary* in English learners’ writing at different CEFR levels.
  • Task 5.
The way we assign learners to CEFR levels – does it matter? Task 5

<table>
<thead>
<tr>
<th>Learners assigned to CEFR levels on the basis of ...</th>
<th>CEFR LEVEL</th>
<th>% of learners at this level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rater agreement</strong> in Task 5 (3 out of 4 agreed)</td>
<td>A1</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>4</td>
</tr>
<tr>
<td>n = 104</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Raters’ median</strong> rating in Task 5</td>
<td>A1</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>6</td>
</tr>
<tr>
<td>n = 183</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Facets analysis</strong> of learners’ writing skill in ALL four tasks</td>
<td>A1</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>2</td>
</tr>
<tr>
<td>n = 183</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### The way we assign learners to CEFR levels – does it matter? Task 5

<table>
<thead>
<tr>
<th>Learners assigned to CEFR levels on the basis of ...</th>
<th>CEFR LEVEL</th>
<th>% of learners placed at this level</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rater agreement</strong> in Task 5 (3 out of 4 agreed)</td>
<td>A1</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>4</td>
</tr>
<tr>
<td><strong>Raters’ median rating in Task 5</strong></td>
<td>A1</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>27</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>6</td>
</tr>
<tr>
<td><strong>Facets analysis</strong> of learners’ writing skill in ALL four tasks</td>
<td>A1</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>2</td>
</tr>
</tbody>
</table>
The way we assign learners to CEFR levels – does it matter? Vocabulary in Task 5

<table>
<thead>
<tr>
<th>Learners assigned to CEFR levels on the basis of ...</th>
<th>CEFR LEVEL</th>
<th>TOKENS (average)</th>
<th>TYPES (average)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater agreement in Task 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>37</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>102</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>157</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>181</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Raters’ median rating in Task 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>37</td>
<td>26</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>90</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>141</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>170</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Facets analysis / ALL four tasks</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>41</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>96</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>142</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>B2</td>
<td>207</td>
<td>105</td>
<td></td>
</tr>
</tbody>
</table>
More on vocabulary in Task 5: indices of lexical rarity and density – correlation with CEFR level?

<table>
<thead>
<tr>
<th>CEFR level (according to rater agreement in Task 5)</th>
<th>Index of vocab. rarity</th>
<th>rho = .143</th>
<th>p = .147</th>
</tr>
</thead>
<tbody>
<tr>
<td>Index of vocab. density</td>
<td>rho = .434</td>
<td>p = .000</td>
<td></td>
</tr>
</tbody>
</table>

| CEFR level (according to raters’ median rating in Task 5) | rho = .144 | p = .052 | rho = .443 | p = .000 |

| CEFR level (according to Facets analysis / ALL four tasks) | rho = .168 | p = .023 | rho = .381 | p = .000 |
Contributions of SLA and language testing in CEFLING

- Language testing $\rightarrow$ SLA, immediate impact
  - e.g. in the design of tasks and rating procedures
  - e.g. showing how the way we convert ratings to CEFR levels can change the substantive SLA-related findings
- SLA $\rightarrow$ language testing, more in the future when it will
  - Improve our understanding of the constructs assessed $\rightarrow$ Diversifying the concept of L2 proficiency
  - Render a more precise description of linguistic development needed in e.g. rating scales for diagnostic purposes
• Do scales such as the CEFR scales imply linear progress in language learning / acquisition?

• Development in SLA non-linear
Does a scale such as CEFR imply that learning is linear?

Whereas in reality, learning is usually non-linear = view of learning in SLA

Assessment is often a snapshot of somebody’s proficiency at a certain point in time. It does not often contain any information about how the learner got where he/she is now.
You can describe changes / development in proficiency with the help of scales – but only if you take longitudinal view.

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
<th>Year X</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>C2</td>
<td>C2</td>
<td>C2</td>
<td>C2</td>
</tr>
<tr>
<td>C1</td>
<td>C1</td>
<td>C1</td>
<td>C1</td>
<td>C1</td>
</tr>
<tr>
<td>B1</td>
<td>B1</td>
<td>B1</td>
<td>B1</td>
<td>B1</td>
</tr>
<tr>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
<td>A2</td>
</tr>
<tr>
<td>A1</td>
<td>A1</td>
<td>A1</td>
<td>A1</td>
<td>A1</td>
</tr>
</tbody>
</table>

Series of snapshots
Issues, dilemmas, contradictions, misunderstandings, etc in SLA & language testing

- Different kinds of scales present different challenges and problems for assessment (for whatever purposes – certification vs. use of scales in SLA research as data gathering instruments)
  - Overall scales vs. very specific scales → the more of the language proficiency a scale attempts to capture, the more difficult it is to place any learner at a particular level
Very comprehensive / wide (overall) scales are problematic both in LT and SLA

<table>
<thead>
<tr>
<th>Writing</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
</tr>
<tr>
<td>C1</td>
</tr>
<tr>
<td>B2</td>
</tr>
<tr>
<td>B1</td>
</tr>
<tr>
<td>A2</td>
</tr>
<tr>
<td>A1</td>
</tr>
</tbody>
</table>

At which level should we place this learner or piece of writing?

Different features of grammar, vocabulary, mechanics, cohesion, sociolinguistic appropriateness, content. …
Issues, dilemmas, contradictions, misunderstandings, … 3

• Share an interest in L2 data collection but methods and aims differ, both between SLA & LT and within them (tradition & purpose):
  • Naturally occurring communication vs. communicative tasks vs. experimental tasks / specific elicitation procedures / discrete point / ’diagnostic’ tests
  • Data gathering (= use of tests etc) in SLA research is done for one purpose only
  • Whereas testing has many different purposes (→ choice of instruments)

→ Task-based SLA research and LT?
  • Communicative authenticity a new source of variability (Norris, Bygate & Van Den Branden, 2009)
Issues, dilemmas, contradictions, misunderstandings, …4

- Constructs of competence – performance – language use?
  - Absolutely necessary for conducting LT
  - For SLA: by collecting and analyzing L2 data, SLA research aims to construct and/or verify theories or models of SLA
- Role of native speaker?
  - Different connotations, significance for LT and SLA research
Thank you!

Riikka Alanen   riikka.a.alanen@jyu.fi
Ari Huhta       ari.huhta@jyu.fi