Using CEFR-scales for assessing young learners‘ oral interactional FL-skills in different testsettings

Astrid Jurecka, Judith Buendgens-Kosten, Ilonca Hardy & Daniela Elsner

Goethe-University Frankfurt
Background

- To meet the European council’s (e.g. 1998; 2002) demands with regard to early foreign language acquisition, foreign language instruction (mostly in English) in German primary schools (1st-4th grade) has become compulsory from 3rd grade on.

- Correspondingly, educational standards for foreign language skills in primary school have been developed, based on the CEFR (European Council, 2001; e.g. Ministry of Education Hessen, 2010):
  - Instruction and educational standards both focus on oral foreign language skills.
  - Competence level at the end of primary school (e.g. Dausend, 2014):
    - level A1 (after two years instruction),
    - or A2 (after four years of instruction)
  - Valid assessment of oral FL-Skills at primary school important (also for teachers)
Construct „oral skills“ in the CEFR:

1) Listening
(Reception; e.g. understanding spoken language)

2) Speaking
(Production of language; e.g. giving presentations)

3) Interaction:
- „At least two individuals participate in an oral exchange in which production and reception alternate and may in fact overlap in oral communication. (...) Learning to interact thus involves more than learning to receive and to produce utterances.
- High importance is generally attributed to interaction in language use and learning in view of its central role of communication“ (CEFR, p. 14).

= primary processes, required for
In Germany, a number of foreign language tests for primary school age include the assessment of reception/listening comprehension and production of spoken language:

**Assessment of FL-listening comprehension:**
- EVENING-Study (Groot-Wilkens et al., 2007)
- Cambridge-Young Learners English (YLE; UCLES, 2013; Toefl Primary
- KESS-Study (Students’ competencies and beliefs; Bos et al. 2006)

**Assessment of spoken language:**
- EVENING-Study (Groot-Wilkens et al., 2007)
- Cambridge-Young Learners English (YLE; UCLES, 2013; Toefl Primary
EVENING: very short interview with test administrator (n=120; 6 minutes, 7 questions: e.g. time until reaction to interview-question, complexity of answer)

   Results/Conclusion (Keßler, 2009, p 168): young learners are able to take part in dialogic communications

TAPS (Testing and Assessing Spoken English in Primary school; Diehr & Frisch, 2008): Assessment of discourse competence, e.g. flexibility and coherence (learned mini-dialogue, learned role-play, presentation, story telling): students exceed standards/requirement

ELLIE (Enever, 2011): longitudinal international study: primary school childrens‘ FL-Skills in different European countries; e.g. Interactive components: learned role play, guessing game

Students in primary school age (6-10) are able to produce spoken foreign language and to communicate in interactive situations to a certain amount, even after one year of instruction.

A few studies and tests include interactive aspects, but:
- interaction is only partly assessed (e.g. discourse competence),
- within very structured test settings (usually interview settings) with limited topics, limited range of interactive activities (e.g. answering (and asking) questions)

- Not all interactive activities are assessed
- Probably leads to restriction of construct validity
Research and Assessment of oral interactional skills at primary school: Summary

- FL-Tests for adults/older children: often include less structured and restricted test settings (e.g. peer-to-peer-settings, even on a low level, e.g. Cambridge-KET; Fit in Deutsch).

- Better overall test performance, more complex interactions and negotiations in peer-to-peer-settings (e.g. Brooks, 2009; Taylor & Wigglesworth, 2009)

- Can less structured peer-to-peer-settings also be used for primary school children?

- Is it possible to assess a broader range of interactive activities by adding peer-to-peer-settings (improvement of construct validity)?

- Aim: Systematic comparison of interview and peer-to-peer-settings with regard to amount of language and interactive activities
Comparing interview and peer-to-peer-settings: Research questions and hypotheses

- **RQ: mount of language production:** How many words are produced and how many interactions (take floor, turns) take place within structured interview- and less structured peer-to-peer-settings?
  - Assumption: more words are spoken in peer-to-peer-settings (topic is less restricted; talking about meta content is possible)

- **RQ: Language:** Which language (English/German) is produced to what amount within the two settings?
  - Assumption: children also use German language in both settings to answer questions or reach communicative goal (low competence levels; social reasons)
  - Assumption: Children use English language more often in interview-settings:
  - possibility to use chunks and learned dialogues even on low competence level
  - social reasons: not used to speak to peers in foreign language;
  - more used to talk in FL with teacher than with peers (e.g. Elsner et al.: group work during FL-instruction often takes place in German)
Comparing interview and peer-to-peer-settings: interactive activities and improvement of construct validity

- **RQ:** Which interactive activities are performed within the two settings?
  - Assumption: Main interactive activity within interviews: answering questions (e.g. CEFR-scale „interview“; EVENING-Study (Keßler, 2009))
  - Broader range of different interactive activities are performed within peer-to-peer-setting (topic less restricted, setting less structured; more different interactive activities possible)
    - Different parts of the construct can be assessed within the two settings

- **(How) Can CEFR scales be used to describe primary school childrens‘ interactive foreign language activities within different settings?**
  - Aim: Development of CEFR-based coding scheme
Method: Sample

- Data collected in Frankfurt/Main: primary school students after about one year of English as FL-instruction (beginning of 4th grade)
  - N=38 (m(age) = 9;4 years; male=18, female=20; German as second language = 12)
  - Students worked in pairs on communicative task
  - Test-administrator: Instruction and interview-questions (English language)

- Task: computer-based storybook „MuViT“ (Multilingual Virtual Talking Book; Elsner, 2011)
  - 16 pages, visual and auditive Input
  - General instruction: Children were repeatedly asked to use English language as much as possible, but were allowed to switch input language, or to use German or Turkish language in case they didn't understand the content or didn't know the vocabulary
  - Test sessions were videotaped and transcribed by trained students
Multilingual Virtual Talking Book MuViT (Elsner, 2011)

1. Early in the morning the alarm clock rings. Ruben doesn’t want to go to school.

2. At breakfast he asks his mum, “Do I really have to go to school today?” “Yes,” his mother replies, “just like any other day.”

3. Ruben walks to school. On his way he finds a red stone.
Interview-Setting & Peer-to-Peer-Setting

**Interview-Setting(s):**
- 4 standardized personal questions (each student)
- 5-6 standardized questions during instruction
- 5 standardized prompts during story

**Peer-to-Peer-Setting(s):**
- Administered in turns
- 5 Peer-to-Peer-Sequences:
  - standardized instruction:
  - students were asked to work together on the story, helping each other to understand the content, discuss incidents

Same amount of standardized speaking opportunities; both settings contain story-based content & elements ▶ Comparability of test settings
Method: Developing a CEFR-based coding scheme

Illustrative CEFR-scales for interactive activities are provided for:

- Overall spoken interaction
- Understanding a native speaker
- Conversation
- Formal discussion and meetings
- Goal-oriented co-operation
- Transactions to obtain goods and services
- Information exchange
- Interviewing and being interviewed
Method: Developing a CEFR-based coding scheme

- Described interactive activities not always consistent (not available for all levels, not available for all contexts = interaction context/level/interactive activity)
- Only few descriptors available for levels A1/A2
- Not all illustrative scales and descriptors seemed adequate for young language learners (see for example also Hasselgreen, 2005; Diehr & Frisch, 2008), (e.g. „formal discussion and meetings“)
- Generalizability of descriptors restricted
- To make them generalizable for use in different settings, we decided to make a „collection“ of all language activities described within the scales, independent of level or illustrative scale
Example: goal-oriented cooperation (Coding)

<table>
<thead>
<tr>
<th>Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>B2</strong></td>
<td>Can understand detailed instructions reliably. Can help along the progress of the work by inviting others to join in, say what they think, etc. Can outline an issue or a problem clearly, speculating about causes or consequences, and weighing advantages and disadvantages of different approaches.</td>
</tr>
<tr>
<td><strong>B1</strong></td>
<td>Can follow what is said, though he/she may occasionally have to ask for repetition or clarification if the other people’s talk is rapid or extended. Can explain why something is a problem, discuss what to do next, compare and contrast alternatives. Can give brief comments on the views of others.</td>
</tr>
<tr>
<td></td>
<td>Can generally follow what is said and, when necessary, can repeat back part of what someone has said to confirm mutual understanding. Can make his/her opinions and reactions understood as regards possible solutions or the question of what to do next, giving brief reasons and explanations. Can invite others to give their views on how to proceed.</td>
</tr>
<tr>
<td><strong>A2</strong></td>
<td>Can understand enough to manage simple, routine tasks without undue effort, asking very simply for repetition when he/she does not understand. Can discuss what to do next, making and responding to suggestions, asking for and giving directions.</td>
</tr>
<tr>
<td></td>
<td>Can indicate when he/she is following and can be made to understand what is necessary, if the speaker takes the trouble. Can communicate in simple and routine tasks using simple phrases to ask for and provide things, to get simple information and to discuss what to do next.</td>
</tr>
<tr>
<td><strong>A1</strong></td>
<td>Can understand questions and instructions addressed carefully and slowly to him/her and follow short, simple directions. Can ask people for things, and give people things.</td>
</tr>
</tbody>
</table>
Method: Development of coding scheme

- Very similar interactive activities were combined
- Resulted in 35 interactive activities, used for coding of videos by two trained raters
- 22 interactive activities coded at least once:

<table>
<thead>
<tr>
<th>Interactive Activities</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ask question</td>
<td>Showing lack of understanding/indecision</td>
</tr>
<tr>
<td>Answer/react to question</td>
<td>Make decision</td>
</tr>
<tr>
<td>Tell/recap/translate content</td>
<td>Make suggestion</td>
</tr>
<tr>
<td>Confirm understanding</td>
<td>Determine/comment</td>
</tr>
<tr>
<td>Make assumption</td>
<td>Give instruction/make request</td>
</tr>
<tr>
<td>Ask back/for clarification</td>
<td>Negotiate content</td>
</tr>
<tr>
<td>Agree</td>
<td>Explain/give reason</td>
</tr>
<tr>
<td>Follow instruction</td>
<td>Discuss procedure</td>
</tr>
<tr>
<td>Continue (task/conversation)</td>
<td>Verify information</td>
</tr>
<tr>
<td>Complete sentences/add content/information</td>
<td>Make consideration</td>
</tr>
<tr>
<td>Endorse proposal</td>
<td>Decline/contradict</td>
</tr>
</tbody>
</table>

- Frequencies of interactive activities were calculated for interview- and peer-to-peer-setting
- Average inter-rater-reliability ICC: .65/.76; Average ratings used for analyses
**Results:**
amount of spoken language and interactions

### Descriptive statistics/ Paired-Samples T-Test

<table>
<thead>
<tr>
<th></th>
<th>mean*</th>
<th>sd</th>
<th>T</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words English overall</td>
<td>39.42</td>
<td>39.49</td>
<td>-11.12</td>
<td>(.000)</td>
</tr>
<tr>
<td>Words German overall</td>
<td>535.84</td>
<td>190.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words English Interview</td>
<td>27.84</td>
<td>20.54</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words English Peer-to-Peer</td>
<td>11.58</td>
<td>23.26</td>
<td>3.70</td>
<td>(.002)</td>
</tr>
<tr>
<td>Words German Interview</td>
<td>118.16</td>
<td>50.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words German Peer-to-Peer</td>
<td>417.68</td>
<td>158.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactions Interview</td>
<td>54.05</td>
<td>12.24</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interactions Peer-to-Peer</td>
<td>58.42</td>
<td>21.94</td>
<td>-.89</td>
<td>(.385)</td>
</tr>
<tr>
<td>Words interview overall</td>
<td>146.00</td>
<td>48.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Words Peer-to-Peer overall</td>
<td>429.26</td>
<td>164.77</td>
<td>-8.430</td>
<td>(.000)</td>
</tr>
<tr>
<td>Words/interaction Interview</td>
<td>2.78</td>
<td>0.95</td>
<td>-7.590</td>
<td>(.000)</td>
</tr>
<tr>
<td>Words/interaction Peer-to-Peer</td>
<td>7.75</td>
<td>2.95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Significantly less English than German words were produced by the children (average: 39.4/video)
  - Possible reasons: low competence level/social reasons?
- As assumed, significantly more English words produced in interview-setting
  - Possible reasons: possibility to use learned phrases & chunks; familiarity of situation

- As assumed, significantly more words (overall; English+German) produced within peer-to-peer-setting
- No difference with regard to amount of interactions
  - Both settings evoke interactive language activities (54/58 per video)
  - Sig. more words per interaction within peer-to-peer-setting
  - Amount of words spoken = effect of test setting?
  - peer-to-peer-setting: more activating/motivating than interview-setting?
## Results: interactive activities within testsettings cross-language analysis (German included)

- Analysis of Variance for Repeated Measurement
- Not all variables could be included (some were only coded once or twice)

<table>
<thead>
<tr>
<th>Measurement</th>
<th>F</th>
<th>Sig.*</th>
<th>Eta²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asking Questions/Clarify</td>
<td>11.729</td>
<td>.003</td>
<td>.395</td>
</tr>
<tr>
<td>Answer/react to Questions</td>
<td>90.983</td>
<td>.000*</td>
<td>.835</td>
</tr>
<tr>
<td>Tell/recap/translacte content</td>
<td>58.900</td>
<td>.000</td>
<td>.766</td>
</tr>
<tr>
<td>Continue task</td>
<td>61.455</td>
<td>.000</td>
<td>.773</td>
</tr>
<tr>
<td>confirm understanding</td>
<td>1.703</td>
<td>.208</td>
<td>.086</td>
</tr>
<tr>
<td>Make assumption</td>
<td>0.486</td>
<td>.494</td>
<td>.026</td>
</tr>
<tr>
<td>Follow instruction</td>
<td>45.521</td>
<td>.000</td>
<td>.717</td>
</tr>
<tr>
<td>Complete sentences/add content</td>
<td>15.657</td>
<td>.001</td>
<td>.465</td>
</tr>
<tr>
<td>agree</td>
<td>10.419</td>
<td>.005</td>
<td>.367</td>
</tr>
<tr>
<td>decide</td>
<td>10.349</td>
<td>.005</td>
<td>.365</td>
</tr>
<tr>
<td>determine/comment</td>
<td>21.411</td>
<td>.000</td>
<td>.543</td>
</tr>
<tr>
<td>Give instruction/make a request</td>
<td>7.966</td>
<td>.011</td>
<td>.307</td>
</tr>
<tr>
<td>decline/contradict</td>
<td>6.761</td>
<td>.018</td>
<td>.273</td>
</tr>
</tbody>
</table>

Alltogether, results confirm assumptions:

**Significant main effect for repeated measurement factor („Testsetting“):**

\[ F=20,892 \ (\cdot001), \ \eta^2=\cdot97 \]

- Univariate tests:
  - Significantly more questions answered within interview-setting
  - 10 of 13 interactive activities significantly more often performed within peer-to-peer-setting
  - Different parts of the construct assessed by different test settings

➤ * = in favor for interview-setting
Discussion

- Peer-to-peer-setting adequate for primary school age?

- Why didn’t they use more English language: due to competence or social reasons?
  - Empirical studies: children in primary school are able to produce foreign language to a certain amount;
  - Possibility to learn/use English vocabulary and sentences at the beginning of the story, and by switching input-language
    - speaks against low competence as only reason
    - Familiarity of situation? How can children become more familiar with free/unstructured peer-to-peer-settings?
- More words and broader range of interactive activities produced.
Discussion

- **Improvement of construct validity by use of peer-to-peer-settings?**
  - Different parts of construct „oral interaction“ assessed by different settings
    - Inclusion of peer-to-peer-settings improves construct validity
  - But: both test settings still do not cover all interactive activities described in the CEFR (only 22 interactive activities coded, 9 less than once per video)
    - More communicative settings should be included (e.g. problem-solving-task for „make suggestions“, „explain“, „negotiate“, etc.)
  - Children did not use English language very much,
    - Interaction between interactive activity and language?
  - Results generalizable for foreign languages/other languages?
    - Feasibility study; next step: conduction of additional (larger) empirical study
Thank you for your attention!