# Establishing assessment categories for rating scales: business as usual?

Gergely A. Dávid and Katalin Piniel-Brózik
School of English and American Studies
Eötvös Loránd University, Budapest
2011

# Research focus

- How rating scale categories may be/ ought to be identified, esp. in the case of complex constructs
  - language proficiency,
  - or even any of its subconstructs such as speaking, writing, etc. proficiencies
  - other complex psychological constructs

# Rating scale design: The case of speaking

- Broad approaches (Fulcher, 2003)
  - Intuitive
  - Data-based
    - Discourse-based rating scale design (Fulcher 1993, 1996)
    - Empirically-derived, binary choice, boundary definition (EBB) scales (Upshur and Turner 1995, 1999)
    - Scaling descriptors, associated primarily with North, 1996/2000)

# Threats to intuitive methods of rating scale development

- A handful of professionals (even a single person) involved rather than a broader professional community
- Copying, „lifting” from other existing scales rather than them being developed on the basis of empirical data (even writers’ experience, samples, etc.)
  - Jeopardises validity

# Ways of scale development that raise ethical questions

- Editing process affected by variable attendance as well as useful process features (cycles of drafting and redrafting)
- Outcomes affected by the group dynamics of the meetings and status of participants rather than theory, practice and research
- Flies in the face of proclaimed democracy
- Demands heavy training as scales have been invented by the chosen few for the many.

# Measurement considerations

- Representing complex constructs
  - assumes a good number of potential categories (spoken and written proficiency, even „MA thesis writing proficiency”)
- Consolidation of categories is desirable in most research/educational contexts,
  - incl. assessor-oriented rating scales (Alderson 1991)
- Potential conflict with feasibility (practicality)
  - Are the many categories feasible as an assessment tool?  Can raters pay attention to a large number of categories?
  - Are all potential categories really necessary?
- Intuitive approach, coupled withy weighting considerations, may result in combinations of categories that are suspect.

# Research so far: How are scale categories determined?

- Categories cannot be combined in a voluntary or haphazard way, but in a way that represents the central tendency of insights from a professional community.
- Not much guidance from research.
  - Fulcher (1996) only developed a single scale.
  - Upshur and Turner (1995, 1999) only designed two.
  - North (2000) descriptors on a single scale after qualitatively identifying a hierarchy of categories, which were determined on a qualitative basis:
    - Consideration of the content of the scales
    - Reference to theory
    - Reference to categories under discussion in the CEFR authoring group (pp.182-183).
- Attention focussed on determining the vertical placement of descriptors on a scale rather than on determining rating categories.

# From vertical to horizontal

- Similar venture: Chalhoub-Deville (1995)
  - Hierarchical set of generic and task specific categoires were replaced with 3 generic categories
- Inspiration from the law of comparative judgment by Thurstone (1927, discussed in Edwards 1957).

# This project

- To develop a set of scales for MA theses in the MA in ELT programme at Eötvös Loránd University, Budapest
  - A very complex construct
  - Useful opportunity because participants' judgement was not going to be overshadowed by a tradition of „this is how we have always done it".

# An overview of the approach

- 3 phases of the project
  - Collecting 100-word definitions of what makes a good thesis.
  - Consolidating the assessment points of view (of what counts in a thesis).
  - Scaling of descriptors for the criteria (how good the theses need to be according to each criterion).
- Only the first two phases will be dealt with in this presentation.

# Phase 1

- Ask for 100-word definitions
  - Looking for the "what" (nouns and NPs) rather than the "how„
  - 13 useful responses by staff
- Leading to 21 potential/ preliminary categories.

# The list of 21 desirables

| | |
|---|---|
| analytical framework | implications |
| argumentation | (quality and number of) sources |
| [enhanced] awareness | quality of research |
| citation conventions | quality of writing |
| contribution to the field | reporting (of research) |
| focus | research methods and procedures |
| familiarity with literature | structure of writing |
| formal requirements | synthesis of knowledge and skills |
| independence | theoretical and experiential basis |
| interpretation (of findings) | |
| layout | |
| originality | |

# Questions for phase 2

- How can the number of categories (21) be reduced in a way that the categories are still inforrmative?
- How many categories would result?
  - 4, 5 or 6, but hardly more.
- Which concepts are „close enough” to be brought together?
  - Some are fairly specific (narrow)
  - Some are more general or inclusive (broad).
- Which specific concepts belong to or under which broad ones?

# Design of Phase 2

- Assumed relationships
  - Identical (no distance)
  - Part/whole (almost no distance)
  - Near synonymous (some distance)
  - Different (considerable distance)
- The measurement of psychological distance
  - Inspiration: Thurstone's (1927a,b,c) law of comparative judgements
  - Coded on a scale 0-1-2-3
- The 21 concepts to be compared to each other
  - (n x (n-1))/2 = 210 comparisons
- A long questionnaire had to be designed

## Paired contrasts questionnaire

Please fill in the questionnaire below. Respond to each paired contrast with a single mark (an X or a √).

Do not leave out contrasts and do not mark two or more.

| | Paired contrasts | | Identi-cal | Part/ whole | Near syno-nym | Diffe-rent |
|---|---|---|---|---|---|---|
| 1. | argumentation | analytical framework | | | | |
| 2. | (enhanced) awareness | argumentation | | | | |
| 3. | citation conventions | (enhanced) awareness | | | | |
| 4. | contribution to the field | citation conventions | | | | |
| 5. | focus | contribution to the field | | | | |

| Concepts and terms | |
|---|---|
| **Concepts** | **Co-text** |
| analytical framework | …creates a framework for the analysis, and applies [knowledge and skills] in an analytical framework |
| argumentation | … academic argumentation<br>[Argumentation]: coherent line of argument, aims and objectives should be clearly defined at the beginning of the thesis.<br>[Argumentation]. …well-structured (cohesive and coherent) |
| [enhanced] awareness | of the link btw theory and practice,<br>of a close connection to the realities of classrooms,<br>a deeper, more sophisticated understanding of the chosen area |
| citation conventions | well written (i.e.. adheres to linguistic, stylistic standards and the citation conventions)<br>Formal requirements: length, layout and presentation, citation conventions |
| contribution to the field | findings that will be of value to other academics; evidence that the writer is (capable of) proceeding towards ….PhD-level research |
| focus | has a clear aim and focus<br>aim and focus related to the field of teaching English<br>topic and focus should be related to teaching and learning of English |
| familiarity with literature | a good understanding and critical appraisal of the lit.<br>reviewing the relevant literature |

# Phase 2: Data analysis

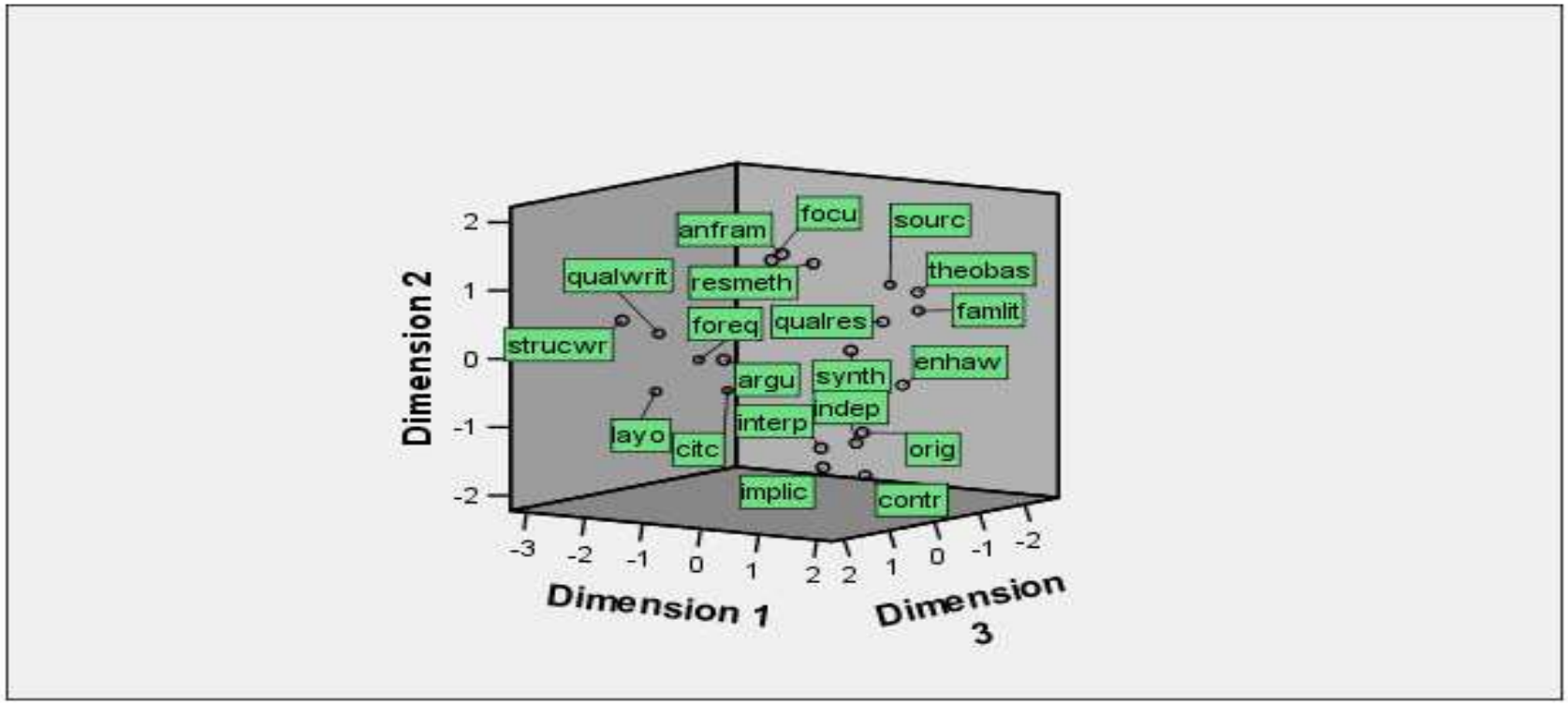
- Statistical analysis of answers
  - Reliability of the qnnaire and scale (SPSS and Facets)
  - Uncertainties about the scale
  - Series of non-parametric tests (One-way $Chi^2$)
  - distributions
  - Multidimensional scaling (MDS)
    - The qualitative analysis of a three-dimensional solution,  stress=0.14, $R^2$=0.86
    - Warning message for 4 dimensions

# Phase 2 results in brief

- 40 colleagues from two relevant departments asked
  - 33 responded (83%), matrix of 33 by 210
- 8 problematic contrasts (qnnaire items) by SPSS
- 3 problematic (misfitting) respondents and 1 of the list of 21 eliminated by Facets
  - Uninformative and inconsistent respondents
  - reporting of results
- 6 consolidated scale categories (+1 quality language)
- Further reduction to 4+1 is possible
  - Initial staff reaqction: „Perhaps we should not.”

| | | | |
|---|---|---|---|
| 15 | quality of research | 1 | |
| 1 | analytical framework | | 1 |
| 18 | research methods and procedures | | |
| 7 | focus | | |
| 14 | quality and number of sources | 2 | 2 |
| 21 | theoretical and experiential basis | | |
| 6 | familiarity w/ the literature | | |
| 20 | synthesis of knowledge and skills | | |
| 5 | enhanced awareness | 3 | |
| 9 | implications | | 3 |
| 11 | interpretation of findings | | |
| 4 | contribution to the field | 4 | |
| 13 | originality | | |
| 10 | independence | | |
| 8 | formal requirements | 5 | 4 |
| 3 | citation conventions | | |
| 12 | layout | | |
| 16 | quality of writing | 6 | |
| 2 | argumentation | | |
| 19 | structure of writing | | |

# Graphical representation of the three-dimensional solution



# Categories as a result of Phase 2

**1. Research methods and procedures**

- Analytical framework
- Focus
- (Quality of research)

**2. Theoretical and expreiential basis**

- Quality and number of sources
- Familiarity with the literature
- (Synthesis of knowledge and skills)

# Phase 2: Results, further consolidated as Interpretation

3. Interpretation of findings
- Implications
- (Enhanced awareness, i.e. evidence of professional development)
- Independence
- Contribution to the field
- Originality

# Phase 2: Results, further consolidated as Formal requirements

5. **Formal requirements**
- Layout
- Citation conventions

6. **Quality of writing**
- Argumentation
- Structure of writing

# Discussion and conclusion

- Example of the MA thesis scales useful for foreign language testing
  - Most language testing operations have scales, the danger of well-established/ engrained categories interfering with the redistribution of construct elements into new categories
- Encouragement from Chalhoub-Deville's research (1995)
  - A case of redistributing of existing construct elements into new, data-based categories
- More accountable, democratic and ethical to arrive at categories in a data-based way.
- Less intensive training demanded?

- Thank you.

# References


- Alderson, J. C. (1991). Bands and scores. In Alderson, J. C. and North, B. (Eds.). *Language testing in the 1990s: The Communicative Legacy*. London: Macmillan.
- Chalhoub-Deville, Micheline (1995). Deriving oral assessment scales across different tests and rateer groups. *Language testing*. 12(1). 16-33.
- Edwards, Allen L. (1957). *Techniques of Attitude Scale Construction.* New York: Appleton-Century-Crofts, Inc.
- Fulcher, G. (1993). The Construction and Validation of Rating Scales for Oral Tests in English as a Foreign Language. Unpublished PhD thesis. University of Lancaster.
- Fulcher, G. (1996b). Does thick description lead to smart tests? *Language Testing* 13/2. 208-238.
- Fulcher, G. (2003). *Testing Second Language Speaking*. Harlow: Pearson Education Limited.
- North, B. (2000). *The Development of a Common Framework Scale of Language Proficiency*. In Belasco, Simon (General Editor). Theoretical Studies in Second Language Acquisition Vol. 8, New York: Peter Lang Publishing.
- Thurstone, L.L. (1927a). A law of comparative judgement. *Psychological Review*. 34, 273-286.
- Thurstone, L.L. (1927b). Psychophysical analysis. *American Journal of Psychology*. 38, 368-389.
- Thurstone, L.L. (1927c). The method of Paired comparisons for social values. *Journal of Abnormal Social Psychology.* 21, 384-400.
- Upshur, J. A. and C. E. Turner (1995). Constructing rating scales for second language tests. *English Language Teaching Journal*. 49(1). 3-12.
- Upshur, J. A. and C. E. Turner (1999). Systematic effects in the rating of secons language speaking ability. test method and learner discourse. *Language Testing*. 16(1). 82-111.