Developing Scoring Rubrics for Speaking and Writing Assessment

Nathan T. Carr
ncarr@fullerton.edu
TESOL Program, Department of Modern Languages & Literatures, California State University, Fullerton

EL Specialist Workshop
Yogyakarta State University
May 23, 2016

Overview

- Terminology notes
- Deciding what we want to do with the results
- Deciding what you want the rubric to tell you about: *Construct definitions*
- Activity 1: Making some plans for a rubric
- Types of scoring rubrics
- Activity 2: Discussion of rubric types
- Writing the descriptors: principles and approaches
- Some things to keep in mind
- Activity 3: Creating a rubric

Terminology Notes

- Two synonyms:
  - Scoring rubric
  - Rating scale
- Bands: The levels on a rating scale
- Descriptors: The descriptions of the levels of a scale

Why Are we Using the Rubric? Some Common Types of Decisions

- Proficiency
- Entrance, selection, screening
- Placement
- Diagnosis
- Progress
- Achievement
  - Courses: Assigning grades; making decisions about passing, promotion, etc.
  - Programs: Graduation, exit, etc.
- Program evaluation

Rubrics: What do you want them to tell you about?

- *Consider the components of your construct definition!*
- A partial list of things that might get included:
  - Grammar
  - Vocabulary
  - Segmental pronunciation
  - Suprasegmental pronunciation
  - Fluency
  - Content
  - Willingness to speak
  - Listener effort necessary
  - Degree of interlocutor sympathy needed
  - Speaker effort
  - Task completion/ performance/ effectiveness
  - Sociolinguistic competence
  - Discourse competence
  - Eye contact with audience

Examples of Common Components of Construct Definitions: Speaking

- Grammar
- Vocabulary
- Segmental pronunciation (vowels and consonants)
- Suprasegmental pronunciation (e.g., stress, rhythm, intonation, prominence, connected speech phenomena)
- Fluency
- Content
- Organization
- Cohesion
- Task performance
- Appropriate use or performance of language functions
- Sociolinguistic appropriacy

Examples of Common Components of Construct Definitions: Pronunciation

- Segmentals (vowels and consonants):
  - Specific vowels, consonants, or sound contrasts
  - Overall level of accuracy
- Suprasegmentals (e.g., word stress, sentence stress, rhythm, intonation, connected speech)
  - Specific suprasegmentals
  - Overall level of accuracy
- Overall pronunciation—including both segmentals and suprasegmentals

Examples of Common Components of Construct Definitions: Fluency

3 definitions (see Bohlke, 2013):

- Hedge (1993): Ability to link units of speech together with facility and without strain or inappropriate slowness or undue hesitation
- Richards & Schmidt (2010): The features which give speech the qualities of being natural and normal, including native-like use of pausing, rhythm, intonation, stress, rate of speaking, and use of interjections and interruptions
- Thornbury (2005): Rate of speech matters, but pausing is more important. Pauses may be long but not frequent, are usually filled, and occur at meaningful transition points. There are long runs of syllables and words between pauses
- **Synthesis:** Smoothness of delivery, without too many pauses or hesitations

Examples of Common Components of Construct Definitions: Writing

- Grammar
- Vocabulary
- Content
- Rhetorical organization
- Cohesion
- Task performance
- Use of appropriate rhetorical mode
- Register

Examples of Common Components of Construct Definitions: Grammar

- Accuracy:
  - Ability to use structures accurately:
  - Ability to comprehend structures
  - Control at the sentence level
  - Control at the discourse or suprasentential level
  - Accuracy of forms
  - Accuracy of meaning (Purpura, 2004)
- Breadth/range/variety
  - Separate from accuracy
  - Avoidance vs. selection of appropriate alternatives
- Specific structures vs. grammatical accuracy in general

Examples of Common Components of Construct Definitions: Vocabulary

- Words and phrases:
  - Ability to recognize and understand
  - Ability to define or explain
  - Ability to use appropriately in context (at the sentential and/or discourse/suprasentential levels)
- Collocations:
  - Ability to recognize and understand
  - Ability to define or explain
  - Ability to use appropriately in context
- Breadth/range/variety
  - Separate from accuracy
  - Avoidance vs. selection of appropriate alternatives
- Specific vocabulary vs. overall vocabulary accuracy

Activity 1: Making Some Plans for a Rubric

See handout:

1. Select a testing context: Who will take the test?
2. Decide what the purpose of your speaking or writing test (or other assessment) will be—that is, what decision(s) will you make using the results?
3. What will you have the students do during the test—what *speaking* or *writing* tasks will they perform?
4. What things are important to you in terms of how you will evaluate their speaking *or* writing performance? Make a list.

Types of Scoring Rubrics: Holistic Rating Scales (1/2)

- Also referred to as *global* or *unitary*
- Rates performance in terms of overall quality or level of ability displayed
- Holistic rating is faster—less for raters to think about—and therefore cheaper

Types of Scoring Rubrics: Holistic Rating Scales (2/2)

- Does not give separate scores for different aspects such as grammar, vocabulary, content, etc.
  - May be satisfactory for native speakers' writing or speaking—various qualities may be closely interrelated
  - For non-native speakers there may be problems, above a certain minimal level of language ability
  - Various aspects of linguistic accuracy are often quite distinct from each other
  - Students with language problems can often generate content and organize it coherently

Types of Scoring Rubrics: Analytic Rating Scales (1/2)

- Also referred to as componential
- Give separate scores to different components of the performance
  - Example: separate scores for grammar, vocabulary, content
  - Another example: vocabulary, grammar, pronunciation, suprasegmentals, and content
- Analytic rating should be more reliable than holistic, all things being equal
  - More scores being used → errors "average out"
  - Analytic scoring can take a little or a lot longer than holistic, so it is more expensive

Types of Scoring Rubrics: Analytic Rating Scales (2/2)

- Components can be averaged together, or weighted, depending on the situation
  - Weighting lets the test designer control how important each component is—not each rater individually
  - In holistic scoring, two raters may not "mentally weight" things the same way.

Types of Scoring Rubrics: Trait Scales

- Trait scales:
  - Include features of the specific task
  - Will not work for another task without revision
- Primary trait: Holistic scale tailored to a particular task
- Multiple trait: Analytic scale tailored to a particular task

Types of Scoring Rubrics: Scale Orientations

- CEFR Fig. 6, types of scales, and their purposes and orientations (after Alderson, 1991)

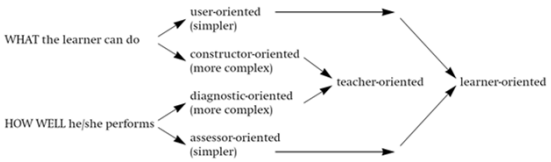

Figure 6

Activity 2: Discussion of Rubric Types

Think of an assessment purpose that you need a rubric for in your context:

1. What would be the advantages and disadvantages of using a holistic rating scale for that purpose?
2. What would be the advantages and disadvantages of using an analytic rating scale for that purpose.
3. For that assessment purpose, what would be the advantages and disadvantages of using a trait-based scale?
4. Do you work with any other assessment purposes where your answers would be different?

Writing the Descriptors: General Principles

- Analytic scales: Number of subscales should reflect the number of components in the construct definition
- Holistic scales: All the components of the construct definition need to be included
- The same things should be mentioned at every level—keep the descriptors parallel across levels
  - Exception 1: "Same as Level _#_"
  - Exception 2: "_X_ does not apply at this level"
- Try to describe what test takers *can* do, not what they can't

Writing the Descriptors: Effects of Features

- Amount of listener effort necessary; amount of listener effort necessary due to _X_
  - Extensive/major listener effort necessary
  - Some/moderate listener effort necessary
  - Little listener effort necessary
  - No listener effort necessary
- _X_ interferes with communication.
- _X_ does not interfere with communication.
- _X_ is distracting.
- _X_ is noticeable.

Writing the Descriptors: Two Approaches

- Rating scales based on theoretical construct definitions:
  - May be derived from a theory of language ability or performance
  - May be based on a set of standards, course syllabus, program goals and objectives, etc.
- Rating scales based on test takers' performance
  - Descriptors based on your familiarity with students' ability at various levels
  - Descriptors based on samples of student performance
    - Group students into levels (HOW???)
    - Describe student performance in each category at that level

Developing Rubrics: Some Points to Keep in Mind

- How many levels can you usefully distinguish?
  - Too few: Can't discriminate between the various levels
  - Too many: Extra levels don't get used, or get overused
- Where do rubrics come from? (parallels Brown, 1995)
  - Adopting existing ones
  - Adapting existing ones
  - Developing new ones
- Make sure the rubric is capturing what you see in *your* students' writing or speaking

Activity 3: Creating a Rubric

See handout:

5. Look at your answers for #4. How many categories do you want to group your criteria/areas of concern into? *Write the names of the categories at the top of p. 3.*
6. In each category, how many levels of ability do you think you can usefully describe?
   - Using the table on p. 3 as a guide, fill in the boxes.
   - Make sure that you include the same criteria at every level.
7. When you finish writing descriptions of all the levels in all of the categories, you are finished. Congratulations: You have now constructed a rating scale from scratch!

References

- Alderson, J. C. (1991). Bands and scales. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy.* London and Basingstoke: MacMillan.
- Bohlke, D. (2013). Fluency-oriented second language teaching. In M. Celce-Murcia, D. M. Brinton, & M. A. Snow (Eds.), *Teaching English as a second or foreign language* (4th Ed.) (pp. 121-135). Boston: National Geographic Learning.
- Brown, J. D. (1995). *The elements of language curriculum: A systematic approach to program development.* Boston: Heinle & Heinle.
- Carr, N. T. (2011). *Designing and analyzing language tests.* Oxford: Oxford University Press.
- Purpura, J. E. (2004). *Assessing grammar.* Cambridge, UK: Cambridge University Press.