Questions and answers about language testing statistics:

Developing rubrics: What steps are needed?

James Dean Brown
brownj@hawaii.edu
University of Hawai'i at Mānoa

Question:

A big question in many Asian countries right now is how to make good quality rubrics for assessing oral and written English. Could you give me some tips on how to do that?

Answer:

This is the second of two columns that I will use to answer your question. In the last one, I talked about the different types of rubrics (analytic and holistic) that can be used for either oral or written language output. In this column, I will explore the stages and steps that you may want to follow in developing a rubric. To do so, I will address three central questions:

1. What steps should you take in developing a rubric?
2. How should you decide on the categories of language behaviors to rate?
3. What should you write in the descriptors inside the cells of the rubric?

What steps should you take in developing a rubric?

Rubric development involves many steps within at least seven stages: planning the rubric, designing the rubric, developing assessment procedures, using the rubric, giving feedback with the rubric, evaluating the rubric, and revising the rubric.

Planning the rubric. The first *stage* in rubric development involves planning, and the first step in planning is to figure out what the goals are for your assessment and rubric. To do that it may help to go back to your source materials (e.g., the syllabuses, teaching materials, other assessments, etc.) and then get together with whatever group of teachers is involved and brainstorm. One key decision that you will need to make in that brainstorming is whether you want to do analytic or holistic scoring (see previous column). If you decide to use an analytic rubric, you should brainstorm which categories of language behaviors you want to use as labels for the columns (as described below in the next main section). If you decide to use a holistic rubric, you will need to brainstorm which categories of language behaviors you want to include in the descriptors for each score level. As a final step in planning, don't forget to decide what range of scores you want to use in your rubric (e.g., 1-3, 1-5, 1-20, other?).

Designing the rubric. Next, you will need to design the actual rubric. This stage will involve three basic tasks, which may seem simple at first, but in reality, will probably take considerable time to accomplish. First, you will need to put the categories on one axis of your rubric. For example, in Table 1 below, the categories are *Fluency, Meaning, Exponents, Register/style,* and *Intonation/stress*. Second, you should put scores on the other axis (e.g., again see Table 1 where the scores are 1, 2, and 3—labeled down the left side). Third, you will need to fill in the descriptors in the cells of the matrix (e.g., again see Table 1 where the descriptors are filled in for *Fluency*). This process is described in the previous column for both analytic and holistic rubrics and further discussed in the third main section of this column.

Developing assessment procedures. If you haven't done so already at this point, you will need to develop your assessment procedures, that is, the processes and methods you will use to gather the language samples from your students. First, you may want to decide on formats for the stimulus material you will use. Will you show students large pictures and ask them to describe what is going on? Will you use a written prompt to get them to write an essay? Will you use a question and answer interview? In short, how will you stimulate the students into producing language samples that you can score and give feedback on? Second, you will need to decide on the response formats that you want to use. Will the students speak into a tape recorder or video recorder? Will they write on paper, or type into a computer file? How will the students actually produce their responses? Third, be sure to write up clear instructions that can easily be understood by those doing the assessing as well as those being assessed. Fourth, make sure that the instructions and stimulus materials are ready at hand when the assessment will take place. Fifth, arrange for the mechanics of assessment (i.e., a scheduled place and clear time, as well as chairs, tape recorders, or any other physical items you may need).

Using the rubric. Using the rubric in practice is one of the most important stages in the assessment process. First, this stage means actually going through the whole process of gathering the language samples from students (i.e., doing the interviews, having them write their emails, or whatever) and compiling all of that language output so that it can be rated using the rubric. Second, if you have a team of raters, you may want to do a training session: to familiarize all raters with the rubric; to show them samples of the language behaviors they will be rating at various levels of proficiency; and to have them practice using the rubric. And third, you will need to have the raters actually use the rubric to score all of the language samples that you gathered. This stage is clearly the heart of the assessment process.

Giving feedback with the rubric. In this stage, you will need to give feedback to the students (and their teachers if that is applicable). Giving them their scores from a holistic rubric or separate for all of the categories in an analytic as well as composite scores is a first step. But because of its pedagogical value, you may also want to make provisions for giving feedback to the whole class or to students individually that explains what the scores mean in terms of the descriptors in the cells of the rubric and their language performances.

Evaluating the rubric. To evaluate the quality of your rubric, you may find it useful to sit down with the raters and get their feedback. They will usually have noticed problems that they had in interpreting the rubric while they were doing the ratings and will probably be more than willing to suggest revisions. It is important to do this while those ideas are fresh in their minds and to carefully listen to what they have to say. It may also prove useful from a validity standpoint to get feedback from other stakeholders (e.g., students, teachers, administrators, etc.). You may also want to evaluate the reliability of your rubric by using two raters to assess each student's language output and checking the consistency of those ratings (either informally, by eyeballing the scores or looking at the percentage of agreement, or more formally, by calculating a correlation coefficient between the two sets of scores; this coefficient will provide a reliability estimate for the scores of either of the raters, ranging from 0.00 for completely unreliable to 1.00 for completely reliable). And, finally don't forget to evaluate the usability of your assessment procedures and rubric by asking yourself if you could make the whole process more efficient and effective. And if so, how?

Revising the rubric. The very last stage involves revising the assessment procedures and rubric to include any observations and insights that surfaced during the previous evaluation stage. The purpose of your revisions should be to make the whole assessment process (including the rubric) work better the next time you use it. If you think of this stage as part of a continuous cycle of revision and improvement, you can't go wrong. In fact, the way I view it, if your assessment procedures and rubrics aren't improving, they are probably deteriorating or getting out of date. At this point, it may be particularly important to pause and

think about the pedagogical implications of what you are doing with your assessment procedures and rubric.

How should you decide on the categories of language behaviors to rate?

In this section, I will focus on: (a) strategies for deciding on the categories of language behaviors you want to include in your rubric and (b) what I hope are some useful ideas for such categories.

Strategies for deciding on categories of language behavior to use in a rubric. In the previous main section, I mentioned *categories of language behaviors* quite a bit. For analytic scoring, the categories will usually serve as the column headings (or sometimes row labels, depending on the orientation of you rubric). However, even in holistic scales, you may want to decide on the various language behaviors that you will describe in your descriptors for each score level. In either case, there are several ways to decide what you will include:

1. You (and perhaps your fellow teachers) could make these decisions based on what it is that you think is important for your students to learn and practice.
2. You can base your decisions on what you want to stress or already stress in your materials and teaching. If a scope-and-sequence chart of those materials is available in the teacher's manual or elsewhere, that may help with these decisions.
3. You may want to send certain messages to your students through your rubric about where they should focus their energies in studying and practicing the language.
4. In addition to deciding on your categories, you may want to put them in order of importance. For example, in developing the analytic rubric in Table 1, a group of teachers thought (and wanted to stress to students) that *Fluency* was most important and then *Meaning*, *Exponents*, etc. Thus, we labeled the rubric columns in that order. In a holistic rubric, you might want to consider ordering the items within your descriptors to serve the same purpose (as in the upper left descriptor in Table 1, *appropriate flow* is most important, followed by *appropriate pauses*, etc.).

Clearly, there are at least four strategies that you can use for making decisions about what language behaviors you want to include in your rubric and in what order. You end up using one, two, three, or all four in making your decisions depending on your pedagogical purpose(s) in using the rubric.

Ideas for categories of language behaviors to consider. Clearly then, both analytic and holistic rubrics are by definition based on the idea of providing scores based on categories of language behaviors. The problem for you in designing your rubric is that there are so many possible categories. For example, in a webinar I did recently on rubrics (see Brown, 2017), I listed the following as possible categories (from Brown, 2012a, p. 20):

1. Pronunciation accuracy or level used
2. Stress timing, rhythm, intonation
3. Grammar accuracy or level used
4. Vocabulary accuracy or level used
5. Collocations
6. Appropriateness of kinesics, proxemics, facial expressions, or gestures
7. Use of down-graders
8. Pragmatics with regard to degree of power difference, social distance, imposition, etc.
9. Fluency
10. Organization
11. Logical development of ideas

12. Topic coverage
13. Getting meaning across
14. Mechanics (capitalization, punctuation, etc.)
15. Coherence
16. Cohesion
17. Register
18. Style
19. Successful task completion
20. Amount of language produced

Given that it originally took me less than ten minutes to come up with this list, imagine the list of ideas that teachers at your institution could generate given more time and brainpower. Notice also that each of the categories listed above could be further divided into separate ratable subcategories.

What should you write in the descriptors inside the cells of the rubric?

Some aspects of creating/wording the descriptors in a rubric were discussed in the previous column. Here, I want to suggest four ways of approaching the creation of such rubric descriptors: *all-or-nothing approaches, target-level approaches*, *matter-of-degrees approaches*, and *multiple-features approaches*. I will use analytic rubrics in my examples here. However, recall that I explained how easy it is to change any analytic rubric into a holistic rubric in the previous column.

Table 1: *Speaking Course Rubric* (adapted from Brown, 2012a, p. 23)

Score	Fluency	Meaning	Exponents	Register/style	Intonation/stress
3	*Almost completely appropriate flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling*				
2	*Somewhat appropriate flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling*				
1	*Mostly inappropriate flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling*				

All-or-nothing approaches. Notice that the top left cell of Table 1 contains a descriptor for the behaviors that one group of teachers decided to use in a rubric for an intermediate level speaking course. That cell describes the characteristics of language behaviors that we were looking for in the fluency of our Chinese students if they were doing well. I call this an *all-or-nothing approach* despite the fact that we were looking for "almost completely appropriate" existence of low, pauses, hesitations, etc. The "almost completely" part of this descriptor simply acknowledged the fact that Chinese speakers of English could be very fluent without being native speakers. Also note that I call this an all-or-nothing approach because (a) all of the characteristics needed to be almost completely present for students to get a 3 for fluency and

(b), if these characteristics were all missing, the score would be 1. Thus, it is all or nothing. However, if the student's speaking performance was somewhere in between (i.e., the characteristics were neither all present, nor all absent), the student could receive a score of 2.

Target-level approaches. Another approach that can be useful if the language behaviors are clear to the students from other sources (like the course objectives, materials, teaching, etc.) is what I call *target-level approaches* because the behaviors are simply judged in terms of whether they are *At Target*, *Approaching Target*, or *Below Target* as shown in Table 2.

Table 2: *Rubric for Group Assessment* (adapted from Jatkowski, 2013)

	At Target (3 out of 3)	Approaching Target (2 out of 3)	Below Target (1 out of 3)
Student maintains topic across three turns			
Student gives appropriate explanations for questions			
Student uses appropriate turn taking			

Matter-of-degrees approaches. I call one other approach the *matter-of-degrees approach* because it depends heavily on adverbs (like *excellent*, *good*, *adequate*, *poor*, and *failing*) that vary in terms of degrees and the raters' abilities to judge language behaviors in terms of those degrees as shown in Table 3 (notice the adjectives at the beginning of each descriptor).

Table 3: *Matter-of-Degrees Variation on the Speaking Course Rubric in Table 1* (adapted from Brown, 2012a, p. 23)

Score	Fluency	Meaning	Exponents	Register/Style	Intonation/Stress
5	Excellent flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling				
4	Good flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling				
3	Adequate flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling				
2	Poor flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling				
1	Little or no flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling				

Multiple-features approaches. Multiple-features approaches can range from descriptors based on simple binary characteristics to more sophisticated descriptions of more characteristics.

Simple binary characteristics descriptors for L2 speech samples might be as simple as the following:

Score of 3 - Plenty of content and that content is intelligible
Score of 2 - Either not very much content or content not intelligible
Score of 1 - Neither much content nor intelligible content

Notice that these descriptors are similar to the all-or-nothing approach described above, but for two characteristics simultaneously.

More sophisticated descriptions for two or more characteristics are also possible if a clear progression of learning can be described. For instance, Table 4 is a brief rubric that provides feedback for *Quality of Information* and *Sources*. Notice that the *Quality of Information* descriptors vary simultaneously in terms of the relationship of the information provided to the main topic and in terms of the amount of supporting details and/or examples, and that *Sources* descriptors vary simultaneously in terms of accuracy of documentation (but only distinguishing scores of 1 from 2-4) and formatting of documentation in terms of degrees.

Table 4: *Example Research Report Rubric* (created using Rubistar at http://rubistar.4teachers.org on April 28, 2018)

Category	4	3	2	1
Quality of Information	Information clearly relates to the main topic. It includes several supporting details and/or examples.	Information clearly relates to the main topic. It provides 1-2 supporting details and/or examples.	Information clearly relates to the main topic. No details and/or examples are given.	Information has little or nothing to do with the main topic.
Sources	All sources (information and graphics) are accurately documented in the desired format.	All sources (information and graphics) are accurately documented, but a few are not in the desired format.	All sources (information and graphics) are accurately documented, but many are not in the desired format.	Some sources are not accurately documented.

Conclusion

In the previous column, I provided you with "tips" on how to think through whether you want to use a holistic rubric or an analytic rubric. In this column, I explained (a) the stages and steps you might need to follow in developing a rubric, (b) how you might go about deciding on the categories of language behaviors you want to rate, and (c) how you can create/word descriptors inside the cells of the rubric. I hope these two columns taken together addressed your question adequately and provided you with the information you need for developing and using rubrics in assessing the oral and written English of your students and, maybe more importantly, for giving them pedagogically useful feedback. [For much more on developing, using, and analyzing rubrics, see Brown, 2012b.]

References

Brown, J. D. (2012a). Developing rubrics for language assessment. In J. D. Brown (Ed.), *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages.* Honolulu, HI: National Foreign Languages Resource Center.

Brown, J. D. (Ed.) (2012b). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages.* Honolulu, HI: National Foreign Languages Resource Center.

Brown, J. D. (2017). *Evaluation criteria and rubrics in online courses.* One-hour invited lesson in the Assessment in Online Language Courses series for the National Foreign Language Resource Center, University of Hawai‘i at Mānoa, Honolulu, HI, 2017. Available from the series website (under Lesson 4) at https://sites.google.com/a/hawaii.edu/assessment-online-language-courses/schedule-1; also available from TED-Ed at https://ed.ted.com/on/7gzI3bES

Jatkowski, K. (2013). Assessing topic maintenance and turn taking. In J. D. Brown (Ed.), *New ways of classroom assessment* (2nd ed.) (pp. 167-168). Alexandria, VA: TESOL.

Where to submit questions:

Your question can remain anonymous if you so desire. Please submit questions for this column to the following e-mail or snail-mail addresses:

brownj@hawaii.edu.

JD Brown
Department of Second Language Studies, University of Hawai‘i at Mānoa
1890 East-West Road
Honolulu, HI 96822 USA