

SHIKEN

Volume 22 • Number 1 • June 2018

Contents

1. Test-taking strategy instruction for Part 3 of the TOEIC Bridge
Stephen M. Paton, Mark W. Howarth, and Alexander Cameron
7. Statistics Corner: Developing rubrics: What steps are needed?
James Dean Brown



Testing and Evaluation SIG Newsletter

ISSN 1881-5537

Shiken

Volume 22 No. 1
June 2018

Editor

Trevor Holster
Fukuoka University

Reviewers

Jeffrey Durand
Rikkyo University

Trevor Holster
Fukuoka University

J. W. Lake
Fukuoka Jogakuin University

Edward Schaefer
Ochanomizu University

Jim Sick
New York University, Tokyo Center

Jonathan Trace
Keio University

Column Editors

James Dean Brown
University of Hawai'i at Mānoa

Jeffrey Durand
Rikkyo University

Website Editor

William Pellowe
Kinki University Fukuoka

Editorial Board

Jeffrey Durand
Rikkyo University

Trevor Holster
Fukuoka University

Jeff Hubbell
Hosei University

J. W. Lake
Fukuoka Jogakuin University

Edward Schaefer
Ochanomizu University

Jim Sick
New York University, Tokyo Center

Test-taking strategy instruction for Part III of the TOEIC Bridge

Stephen M. Paton¹, Mark W. Howarth², and Alexander Cameron²

home@stevepaton.com

1. Fukuoka University

2. Kyushu Sangyo University

Abstract

Part III of the TOEIC Bridge requires candidates to listen to a short conversation and then answer a written multiple-choice question about it. A commonly-taught test-taking strategy for this part of the TOEIC is for test-takers to read the question and answer choices before listening to the conversation so as to know what to listen for, rather than trying to recall what had been heard and attempting to answer the question post-listening (Trew, 2008). The current study aims to examine the effectiveness of this approach to Part III of the TOEIC Bridge in two ways: firstly by comparing the gains of students who had and hadn't received this particular strategy instruction, and also examining the effect of forcing the strategy to be used. The study was carried out with 148 Japanese university students of low English proficiency with TOEIC Bridge scores between 90 and 140. Results seem to indicate that teaching this strategy to students at this level does not significantly improve scores, and may, in fact, hinder their performance.

Keywords: Test-wiseness, strategy instruction, TOEIC Bridge, testing

Studies have shown that high-proficiency learners and low-proficiency learners often use similar strategies but with varying degrees of success, especially when taking tests (Purpura, 1998). High proficiency learners tend to be product-oriented test-takers, paying more attention to what is read or heard by retrieving language from their long-term memory. Low proficiency learners, on the other hand, are more process-oriented test-takers, meaning they have to devote more attention to what the task is asking of them than reading or listening to the actual contents of the test. Purpura (1999) claims that process-oriented test-takers are highly disadvantaged in timed testing environments.

One way that process-oriented learners can improve their performance on tests is by improving their test-taking skills, often referred to as "test-wiseness". Millman, Bishop, & Ebel (1965) defined this as "a subject's capacity to utilize test characteristics and formats of the test-taking situations to receive a higher score" (p. 707). While knowledge of the content which is being taught is necessary to perform well on a test, evidence shows that test performance can be enhanced by improving test-wiseness and by acquiring test-taking skills (Rogers & Bateson, 1991; Sarnacki, 1979). A study by Yang (2000) found that takers of the TOEFL who were deemed to be "test-wise" were found to have a more thoughtful and less random approach to answering questions than "test-naïve" learners, leading to higher scores. It appears, therefore, that explicitly teaching test-taking skills and improving "test-wiseness" can enhance learners' performance on language tests, especially for process-oriented learners.

The object of the current study is test-taking strategies for Part III of the TOEIC Bridge, which despite being part of the listening section of the TOEIC Bridge, draws heavily on the students' reading ability. Hoover and Gough (1990) argued in their Simple View of Reading (SVR) that reading consists of only two components, decoding written text, and linguistic comprehension. Decoding is the ability to derive lexical meaning from written forms of words, whereas linguistic comprehension is the ability to derive sentence and discourse-level meaning from what has been decoded. These two components are independent, and one can be more skilled in one than the other, but reading proficiency depends on the ability to do both. Part III of TOEIC Bridge places considerable demands on the test-taker by combining a listening component with the already challenging task of reading the questions.

Outline of TOEIC Bridge Part III: Short Conversations and Short Talks

Like parts III and IV of the regular TOEIC test, Part III of TOEIC Bridge requires that candidates answer written multiple-choice questions, printed in their test book, pertaining to short audio conversations or talks that contain (or imply) the answer. Unlike the regular TOEIC, the Bridge test asks just one question for each short audio passage. There are fifteen items, and the audio passages are separated by approximately ten seconds of silence. In that short amount of time, candidates need to both select an answer for the passage they've just heard, and read ahead in order to focus their listening and search for answers during the upcoming audio. Good time-management skills are essential, and a common pitfall often warned against in TOEIC preparation textbooks (Trew, 2008) is for candidates to spend too much time thinking back over what they've just heard, attempting to remember what was said as they re-read a question, rather than utilizing the time available by reading the text of the upcoming question.

Current Experiment

This experiment was initially intended to test the following hypothesis: teaching candidates the strategy of utilizing the silent period between audio passages to read the upcoming question in preparation for the upcoming audio, rather than spending the time reflecting on the audio passage that had just passed, will lead to an improvement in their scores.

Materials by which to teach this strategy were created using audio editing software and presentation software, and are discussed at length below. The process of creating such materials, though, gave rise to an additional hypothesis that also warranted investigation. When using PowerPoint or Keynote for instructional purposes, for example to display sample questions on the screen, the teacher obviously controls what the class can and can't see on the screen at any time. That is, the teacher can limit and control what the student can place their attention on at any given time during the presentation. Perhaps, then, a teacher could use presentation software during a mock test to 'enforce' the strategy of utilizing the silent period for reading the upcoming question, by removing the preceding question from the students' view and effectively prohibit them from seeing it. Rather than giving the students a printed page containing all the questions, the teacher could display and hide the question text strictly according to what stage of the audio they were at. Controlling which question items the students could see, it was hypothesized, would likely focus students' attention even more firmly on reading for the upcoming question, and result in even stronger gains than only teaching the strategy but giving students the freedom to take advantage of it or not.

Research Questions

The research questions were clarified thus:

- 1) Will candidates perform better in a practice test of Part III of TOEIC Bridge after being taught to utilize each period of silence between audio passages to read only the question for the upcoming audio passage, rather than remaining focused for too long on reading and answering the question for the preceding audio passage?
- 2) Having been taught the strategy, will candidates perform better if, during the silent period, they are actually prevented from reading the question for the preceding audio passage, and are instead only allowed to see the question for the upcoming audio passage? In other words, will 'enforcing' the strategy have a beneficial effect?

Method

Participants

The sample was made up of 148 first-year university students who were preparing to take the TOEIC Bridge test. The students varied in levels of English proficiency and were members of ten separate classes being taught by two teachers.

Design

Materials

Presentation materials were produced using Keynote presentation software for the purpose of introducing students to Part III. It consisted of around 25 slides, with audio passages embedded, and took about 15 minutes to present. For all students, in experimental and control groups, this was comprised of a demonstration of the part's structure and format, the length and speed of the audio passages, the multiple-choice nature of the question items, and its timing and pace. During the reading time, students were encouraged to identify and select 'key' words in the questions, to be specifically listened for later. A number of 'key' words were suggested by way of underlines appearing on the slides. Sample items from Barron's TOEIC Bridge Test (Lougheed, 2010) were used in the presentation.

The two experimental groups, in addition to the demonstration mentioned above, were also introduced to the test-taking strategy. In this version of the presentation the students were implored to "Read the question first. Find the keywords, THEN listen." One slide sequence asked students to mentally re-frame the test items, and to think of each one as starting at the end of the preceding audio, rather than at the beginning when each test question is announced with its question number. Another slide implored students several times: "When the speaking stops, start reading the next question."

At one point, the value of reading the questions in advance of the audio passage was demonstrated by actually disallowing any reading until after the audio had stopped, and challenging students to answer the question with reference only to their immediate memory of the audio—something which even a native speaker would not find easy. This was done in the hope that the students themselves would recognize the value of the strategy being taught.

In another section of the experimental groups' presentation, by displaying the script text after listening to audio passages, the students were shown, visually, that the answer to the question may come at the beginning, middle, or end of the talk; that it may be surrounded by convincing distractors relating to incorrect answer options; and, once again, that by reading the question before listening, one can confidently discard irrelevant information when something specific is being sought out. Students were also instructed to begin reading the next question as soon as they were confident they'd identified the correct answer, without waiting for the audio to end. The strategy of maximizing reading time, then, was emphasized in several ways.

Care was taken to make sure that the students in each group were given a comparable exposure to practice items.

Administration

The experiment called for three groups, one control group and two experimental groups.

Group A (n=51) was the control group, and would receive the explanatory introduction to the section, but not receive instruction in the test strategy. Groups B (n=52) and C (n=45) were to receive the explanatory introduction and the strategy instruction, but be post-tested differently.

Before any instruction had taken place, to establish a baseline score and to control for rehearsal benefit, the members of all three groups were given a full 15-item Part III practice test. All of the students in all three groups did an identical test. It was carried out in much the same way as a regular TOEIC Bridge test, i.e., with each student being given a printed copy of the questions, and an answer sheet on which they selected a, b, c, or d for each question. The audio was played directly from a CD as it would be in a regular test, with no interference from the teacher. The answer sheets were collected and scores were recorded. What happened in the subsequent lesson varied for each group.

Following the presentation that each group was shown, a second test was administered to check for gain. This test was taken from the same test preparation textbook as the pre-test and was deemed to be of equal difficulty and therefore appropriate. For group “A”, the control group, and group “B”, the test was administered once again in line with usual TOEIC Bridge circumstances; i.e., with questions printed on paper and given to each candidate, and the audio CD played from start to finish with no interference by the teacher.

Group “C” was given a test comprising the exact same items as the other groups but administered differently. The candidates were given a paper answer sheet only and did not receive a printed copy of the questions. Rather, Keynote slides were prepared that showed each question text and the four possible answers. This allowed the teacher to fully control what the candidates could see and read, and thus to ‘enforce’ the strategy of only reading the question for the upcoming audio once each audio passage had ended. During the test, within a second or two of each audio passage ending, the teacher would advance to the next slide, which featured the question text for the upcoming question. The silent period was therefore spent with only the question for the upcoming audio visible, and candidates were prevented from reading for the preceding audio. Once again, for all groups, answer sheets were collected and scores were recorded. Pre-test and post-test scores were compared to measure gain.

Results

Each group’s mean pre-test and post-test scores are shown. Group A saw the biggest gains, Group B scores decreased, and Group C scores increased very slightly. One-way ANOVA did not find statistical significance, $F(2, 145) = 2.10, p = .126$. Table 2 shows the results of multiple *t*-tests, with Bonferroni adjustment. Group A saw an average gain of about 1.05 points out of 15 (7.0%) when compared to Group B, and about 0.41 points (2.73%) when compared to Group C. Group B, however, scored 0.64 points (4.26%) lower than group C. In terms of raw scores, a clear pattern of group A outperforming the others is evident.

Table 1

Pre-test/post-test mean scores

	Pre-test Mean Score	Post-test Mean Score	Gain
Group A	6.20	7.00	0.80
Group B	8.14	7.76	-0.38
Group C	7.75	8.06	0.31

Table 2
Multiple group comparisons

		Mean Difference	S.E.	p
Group A	Group B	1.05	0.52	.13
	Group C	0.41	0.54	1.00
Group B	Group A	-1.05	0.52	.13
	Group C	-0.64	0.54	.70
Group C	Group A	-0.41	0.54	1.00
	Group B	0.64	0.54	.70

Discussion and Conclusion

In terms of research question number one, in which we asked whether candidates would perform better after being taught the test taking strategy, the results indicate that this may not be the case. Group A, the control group that was not taught the test-taking strategy, showed the biggest gain. This essentially means that students who received no instruction on the specific test-taking strategy out-performed students who did.

There are a number of possible explanations for this. One is that the students who received specific instructions were perhaps not given enough time to practice the strategy between the pre- and post-tests. Given more time to internalize the strategy through extensive practice may have produced different results. Another explanation might be that the students, who were at a very beginner-level of language learning, may have been overwhelmed and confused by the test-taking strategy instruction. One last possible explanation, as mentioned above with regards to the Simple View of Reading, is that the students may have lacked efficient decoding skills, which would have hindered their ability to process the combination of written and spoken English that characterizes this part of the test. Future experiments of this nature with mid or high-level learners might produce different results.

In terms of research question two, in which we asked whether enforcing the strategy would have a beneficial effect, it appears that having taught the strategy, enforcing it produced a better result than not enforcing it. However, it should be noted that students who were not taught the strategy outperformed those who were taught and enforced, albeit not to a statistically significant degree.

Definitive conclusions are not possible due to the small sample size and relatively simple design of the experiment, especially with respect to the small number of test items, but the results did run counter to what was expected. At the very least, therefore, we cannot conclude that teaching a test-taking strategy will make a positive difference. This is not to say that teaching test-taking strategies is ill-advised. Rather, that students at this level might actually perform worse on tests when they are asked to utilize test-wisness strategies such as the one taught here and might be better served by being given practice which will develop their decoding skills. Concerning them with test-taking strategies while their reading skills are insufficient might actually hinder their ability to perform well on the test.

References

- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, 2(2),127-160. doi: 10.1007/bf00401799
- Lougheed, L. (2010). *Barron's TOEIC Bridge Test*. New York: Barron's.
- Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wisness. *Educational and Psychological Measurement*, 25(3), 707-726. doi: 10.1177/001316446502500304

- Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test-takers: A structural equation modeling approach. *Language Testing*, 15(3), 333-379. doi: 10.1177/026553229801500303
- Purpura, J. E. (1999). *Learner strategy use and performance on language tests: A structural equation modeling approach*. Cambridge, UK: Cambridge University Press.
- Rogers, W. T., & Bateson, D. J. (1991). Verification of a model of test-taking behavior of high school seniors. *Journal of Experimental Education*, 59(4), 331-349. doi: 10.1080/00220973.1991.10806571
- Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive domain. *Review of Educational Research*, 49(2), 252-279. doi: 10.3102/00346543049002252
- Trew, G. (2008). *Tactics for the TOEIC listening and reading tests*. Oxford: Oxford University Press.
- Yang, P. (2000). *Effects of test-wiseness upon performance on the Test of English as a Foreign Language*. (Doctoral dissertation, University of Alberta). Retrieved from http://www.nlc-bnc.ca/obj/s4/f2/dsk1/tape3/PQDD_0009/NQ59700.pdf

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 your 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	<i>Almost completely appropriate flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling</i>				
2	<i>Somewhat appropriate flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling</i>				
1	<i>Mostly inappropriate flow, pauses, hesitations, fillers, speed, connectedness, and back-channeling</i>				

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

Call for Papers

Shiken is seeking submissions for publication in the December 2018 issue. Submissions received by 1 September, 2018 will be considered, although earlier submission is strongly encouraged to allow time for review and revision. *Shiken* aims to publish articles concerning language assessment issues relevant to classroom practitioners and language program administrators. This includes, but is not limited to, research papers, replication studies, review articles, informed opinion pieces, technical advice articles, and qualitative descriptions of classroom testing issues. Article length should reflect the purpose of the article. Short, focused articles that are accessible to non-specialists are preferred and we reserve the right to edit submissions for relevance and length. Research papers should range from 4000 to 8000 words, but longer articles are acceptable provided they are clearly focused and relevant. Novice researchers are encouraged to submit, but should aim for short papers that address a single research question. Longer articles will generally only be accepted from established researchers with publication experience. Opinion pieces should be of 3000 words or less and focus on a single main issue. Many aspects of language testing draw justified criticism and we welcome articles critical of existing practices, but authors must provide evidence to support any empirical claims made. Isolated anecdotes or claims based on "commonsense" are not a sufficient evidential basis for publication.

Submissions should be formatted as a Microsoft Word (.doc or .docx format) using 12 point Times New Roman font, although plain text files (.txt format) without formatting are also acceptable. The page size should be set to A4, with a 2.5 cm margin. Separate sections for tables and figures should be appended to the end of the document following any appendices, using the section headings "Tables" and "Figures". Tables and figures should be numbered and titled following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition*. Within the body of the text, indicate approximately where each table or figure should appear by typing "Insert Table x" or "Insert Figure x" centered on a new line, with "x" replaced by the number of the table or figure.

The body text should be left justified, with single spacing for the text within a paragraph. Each paragraph should be separated by a double line space, either by specifying a double line space from the Microsoft Office paragraph formatting menu, or by manually typing two carriage returns in a plain text file. Do not manually type a carriage return at the end of each line of text within a paragraph.

Each section of the paper should have a section heading, following the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition*. Each section heading should be preceded by a double line space as for a regular paragraph, but followed by a single line space.

The reference section should begin on a new page immediately after the end of the body text (i.e. before any appendices, tables, and figures), with the heading "References". Referencing should strictly follow the guidelines of the *Publication Manual of the American Psychological Association, Sixth Edition*.

