TESL-EJ Teaching English as a Second or Foreign Language

The Electronic Journal for English as a Second Language

Constructing Analytic Rubrics for Assessing Open-Ended Tasks in the Language Classroom

February 2021 – Volume 24, Number 4

Mary Lou Vercellotti
Ball State University
<mlvercellott@bsu.edu>

Dawn E. McCormick
University of Pittsburgh
<mccormic@pitt.edu>

Abstract

With the increased emphasis in communicative language learning and task-based language teaching, classroom assessment increasingly includes performance-based assessments (e.g., essays, speeches, projects, presentations), which require careful planning and can be challenging to implement and assess. The ability to design and use well-designed assessments for language performance tasks is a critical skill for classroom instructors. A rubric is a tool to more objectively measure the quality of the language performance when assessing language use in open-ended tasks. Analytic rubrics, with multiple categories and descriptions reflecting various levels of performance, help the instructor evaluate the effectiveness of instruction, document evidence of learner progress, and give feedback to learners. This article synthesizes theoretical and empirical scholarship in order to describe how to construct well-designed analytic rubrics for classroom language assessment. Four main steps are described: 1) establishing categories, 2) describing levels of performance, 3) reviewing the components of the rubric before implementation, and 4) evaluating the effectiveness of the rubric after implementation.

Keywords: *rubrics, scoring scales, assessment, grading, performance-based, task-based*

Classroom Assessment

Classroom assessment documents evidence of learner progress, gives feedback to learners, and allows teachers to evaluate the effectiveness of instruction. Instructors should be able to design and implement "varied and valid assessments…to support student learning" (TESOL International Association, 2018, p. 60). Classroom instructors may choose a variety of assessments, from objective tests, such as form-focused multiple choice tests, to open-ended performance-based tasks, with many tasks in between these two sides of the assessment continuum. Assessing what a learner knows *about* a language can be relatively straightforward. For instance, the learner supplies answers on an objective test (e.g., vocabulary matching quiz) whose questions have expected answers, and the instructor can score the learners' work reliably with an answer key. For classroom assessment with specific form-focused objectives, other assessment tools, such as detailed grading criteria, checklists, and tallies may be useful. Learner knowledge about the language, though, does not reflect what the learner can *do* with that knowledge (Van Gorp & Deygers, 2014), and the results of objective tests do not reflect the learner's language skills, such as the ability to speak or write the language. With increased emphasis on communicative competence and task-based language teaching, language instructors have begun to assign more open-ended language tasks as pedagogical activities and as assessments (Purpura, 2016; Van Gorp & Deygers, 2014). Given the variety of assessment types and assessment tools, a crucial issue in assessment is the alignment between the content to be assessed and the assessment procedure. In a performance-based (or task-based) approach to language assessment, the learner's performance in open-ended tasks is seen as evidence of the learner's knowledge, skills, and ability (Purpura, 2016). With an open-ended language production task (e.g., essays, speeches, projects, presentations), the learner has the opportunity to choose the vocabulary and grammatical structures in response to the prompt or assignment. For instance, one learner may choose to use more complex grammar to express an idea while another may choose to create simpler but accurate structures. These open-ended performance-based assessments are vital in classroom language assessment because they more directly measure productive skills (e.g., speaking, writing). Performance-based assessments, however, require careful planning and can be challenging to implement. Instructors may not yet have the requisite knowledge to assess these language performances. Because of these challenges and others, instructors may borrow an existing assessment from another context (e.g., another course, another proficiency level), but using a rubric designed for another context is a potential misuse of the assessment tool (Purpura, 2016). Accordingly, it is important for instructors to create well-designed assessments for their classroom. In fact, designing valid assessments is a key teaching principle (TESOL International Association, 2018).

This article describes how to construct an analytic rubric for classroom language assessment, informed by theoretical and empirical scholarship. After introducing rubrics as a tool for assessing student performance of open-ended tasks, we then describe how to construct an analytic rubric by identifying categories and describing levels of performance for each category. The third and fourth sections provide recommendations for pre-use and post-use reviews, critical steps in the assessment process. Last, we provide a checklist that summarizes steps in constructing analytic rubrics for open-ended tasks and provides practical suggestions.

Performance-based Assessments and Rubrics

Classroom instructors choose from a range of assessments, with consideration of the link between course objectives and the assessments themselves. Open-ended tasks focus on course-level communication objectives, rather than on specific grammar or vocabulary of any particular lesson, thus requiring an assessment tool that measures the quality of the language produced during these tasks. Analytic rubrics are one tool for open-ended tasks in an instructor's assessment tool kit, thus instructors must determine whether a rubric is the right tool for the assessment. The concepts discussed below provide guidance to determine if a rubric is an appropriate assessment tool.

Open-ended tasks often are evaluated through the rater's impressions or judgements (Green & Hawkey, 2012), which are inherently subjective, even when learners are to be evaluated by how well they meet stated assessment objectives. The ratings could differ because different raters can have different impressions of the performance (Green & Hawkey, 2012), but even a single rater may have difficulty scoring consistently. For instance, scoring may become more conservative (or more generous) as fatigue increases (Van Moere, 2014), or a rater might start assessing quite strictly and lower expectations after viewing multiple performances. Raters can also be influenced by the previous language performance during the grading process (Upshur & Turner, 1995). After scoring a really strong essay, for instance, the rater may score the next essay lower than if it had followed a weaker essay. Having clearly ranked descriptions of language performance will help increase the consistency when assessing, and so rubrics have been increasingly used to assess the quality of the performance in open-ended tasks, such as speaking and writing.

The term "rubric" has sometimes been used to refer to any grading criteria. In this article, the term rubric refers more narrowly to a specific assessment tool with descriptions of various levels of the quality of the performance. Goldberg (2014) has defined a rubric as "a scoring guide that outlines features of work at different levels of performance" (p. 1). Holistic rubrics and analytic rubrics are two types of common tools for language performance. With holistic rubrics, the descriptions of the levels of performance include multiple traits, resulting in a single score. Holistic rubrics can be more practical when only an overall score is needed (e.g., a quick placement test, a proficiency test) (Brown & Abeywickrama, 2019) as a measure of learning that has happened (Brookhart, 2018). Since holistic rubrics provide less specific information, they are less useful for classroom assessment. Analytic rubrics, on the other hand, are more useful during the learning process (Brookhart, 2018). Analytic rubrics list multiple traits or categories separately for the rater to rank (e.g., good, better, best) based on descriptions of levels of performance. Analytic rubrics strengthen the reliability of the assessment for language produced during such tasks (Green & Hawkey, 2012). Since the instructor can give separate ratings for each category, the learner receives specific feedback about what they can do and what skills need more improvement (Brown, 2018). Additionally, the results of the assessment with an analytic rubric provides the instructor information about the strengths and weaknesses of the learners.

Table 1 shows an author-created analytic rubric for an individual informational speech in a language class for beginners to serve as an example. It is designed to assess three language skills: vocabulary, grammar, and pronunciation at three possible levels: meets standard, approaching standard, and needs improvement. In Table 1's format, the lowest performance is at the bottom and each higher row has descriptions of higher level of performance.

Table 1. Analytic Rubric for Informational Speech- City; High Beginner Learners

Criterion	**Vocabulary**	**Grammar**	**Pronunciation**
Meets Standard	Speaker used specific, descriptive words to create a visual picture.	Speaker created simple sentences using present tense.	Speaker was clear, comprehensible and listener could understand meaning.
Approaching Standard	Speaker used appropriate but general descriptive words.	Speaker created simple and incomplete sentences using present tense.	Speaker was comprehensible but listener had to focus to understand.
Needs Improvement	Speaker used incorrect (or missing) descriptive words.	Speaker created incomplete sentence with base form verbs.	Speaker was generally not comprehensible even with listener effort.

In summary, language instructors often assign communicative language tasks, but these open-ended assignments are challenging to assess because the resulting language performances are all different. Analytic rubrics are an appropriate choice for instructors when assessing open-ended extended language performances because they provide a structure for consistent assessment and feedback.

In the following sections, we review how to construct an analytic rubric for teacher-implemented classroom assessment. Author-created rubrics are provided to model the concepts and steps presented. These are illustrative and not intended to be applied to specific classroom contexts. The imagined context for the following rubric is an adult, mixed-gender, integrated-skills English as an additional language (EAL) class in an English-speaking country. The students' levels range from high-intermediate to low advanced [1[i]] with a minimum of 52 on the TOEFL iBT or 5.5 on the IELTS to meet the entrance requirement. The students identify their motivation for language learning as academic, professional, and/or personal. The learning goals focus on speaking and listening, and reading, writing, grammar, and pronunciation skills are developed explicitly when they support speaking and listening. We focus on a rubric for an individual speaking activity. Students engage in language practice that facilitates transfer to language use in everyday, work, and academic settings. Although a specific class situates the example rubric, the concepts and steps are applicable across multiple teaching contexts.

Constructing Analytic Rubrics

Analytic rubrics have two main parts: categories to evaluate and ranked descriptions of performance (Brookhart, 2018), usually set up like a table, with the categories either listed in the columns (shown in Table 1) or the categories set up in the rows (shown in Table 2). Both the categories in the analytic rubric and levels of student performance are designed for a specific context, a specific course with a specific student population (Moskal & Leydens, 2000). The following two sections describe how to identify categories and determine performance descriptors for specific classroom assessments.

Categories

Each category (i.e., evaluation criteria) in the analytic rubric assesses a different aspect of the learner’s skill (e.g., Popham, 1997; Van Moere, 2014). In other words, the categories of an analytic rubric are independent. For classroom assessments, each learning objective to be assessed can be represented as a category in the rubric (Brookhart, 2018). The rubric in Table 2, for instance, has categories focused on linguistic skills (i.e., vocabulary, grammar structures, pronunciation, fluency). These categories reflect the skills required for effective communication. The rubric also includes a category on content to represent a learning objective of presenting factual information with supporting statements. The inclusion of discourse-level categories, such as content and fluency, reflects the understanding that effective communication is more than adherence to sentence-level forms that is captured by the other categories. In Table 2’s format, the categories are along the leftmost column.

The categories in the rubric must focus on observable and measurable skills of the performance and be directly linked to the assessment’s learning objectives. For instance, perhaps the students are asked to complete the task, “Present a ≤ 4-minute prepared speech,” in order to meet objectives that include:

1. Use topic-appropriate vocabulary

2. Use a variety of grammatical structures, including compound and complex sentences, and transitions that support meaning

3. Produce segmentals and suprasegmentals that are comprehensible to listeners

4. Use appropriate pacing and pausing

5. Present original or given information with relevant supporting ideas

The five categories in Table 2 relate to the students’ language learning goals of the unit and reflect the course learning objectives, and the instructor can hear the evidence (i.e., observable) and evaluate performance (i.e., measurable) during the speech.

A rubric with many categories may become impractical because it is difficult to assess all of the language skills listed (Schreiber et al., 2012). Generally, five categories may be the maximum a rater can attend to (Green, 2014; Popham, 1997). No assessment can assess every skill; a rubric’s categories must be carefully selected from the learning objectives for that particular assessment.

Table 2. Analytic Rubric Categories: Formal Informational Speech; High-Intermediate Learners

Category				
Vocabulary				
Grammar Structures				
Pronunciation				
Fluency				
Content				

Since productive language skills tap language knowledge, analytic rubrics to assess writing and speaking may have the same linguistic categories of vocabulary and grammar. Despite these similarities, the grammar patterns and vocabulary use differ in written text and in speech because written language can be revised (Vasylets et al., 2017; Vercellotti, 2018).

In some rubrics for assessing speaking, instructors may include nonverbal behaviors (e.g., eye-contact, gestures) to evaluate that aspect of oral communication. Raters have been found to attend to nonverbal behavior in dialogues, even when that category is not included in the rubric (May, 2009). Nevertheless, research with native English speakers has shown that nonverbal behaviors do not correlate with speech grades (Schreiber et al., 2012), which indicates that those skills do not seem to be integral to effective monologic speaking. Accordingly, nonverbal communication could be included when the course (or unit) objectives specifically include development of such skills, particularly in interactive communication, but a category for nonverbal behavior during speaking performances may not be necessary to assess speaking skill.

Levels of Performance

After identifying each category in the analytic rubric, the next step is to create descriptions for different levels of performance in each category. The description of each level of performance should convey the criteria for that category to help the rater identify which description most aligns with the learner's performance (Goldberg, 2014). In other words, the descriptions should help the rater place the performance at a particular level for each category. It is often helpful to first describe the expected performance at the "Meets Standard" level because the expectations of the stated standard is most evident. After describing the expectation for the standard, a description of the level of performance in that category at one level higher or one level lower can be developed. The differences between each level of performance should be distinct

(Moskal, 2002). Labeling each level can help focus the rating process around the standard (Suskie, 2009). Such labels are also useful to anchor the descriptions around the standards when drafting the ranked descriptions (e.g., Exceeds Standard, Meets Standard, Approaching Standard). The example rubric in Table 3 includes descriptive labels as well as the corresponding letter grade (used in schools in the United States). It is important to note that the rubric's top level of performance should not expect learners to have a native-like performance; learners should not be expected to have a perfect performance to receive full points.

Importantly, although the labels for the levels of performance may be evaluative terms (such as "excellent" or "exceeds standards), the descriptions differentiating the levels of performance should not use evaluative terms. Only "descriptive language helps students envision where they are in their learning and where they should go next" (Brookhart, 2018, p. 2) and "provide a clear description of the learning goal" (Brookhart, 2018, p 5). In addition, the descriptions should predominantly focus on what the learner demonstrates rather than what is absent. This recommendation does not limit the descriptions to "can do" statements. Some negative descriptions might be used, but without describing what is missing. For instance, the phrase "*word choice may sometimes be too specific or vague*" (rather than "*did not use a variety of words"*) focuses on what the learner produced.

Since rubrics are designed to evaluate the quality of the language performance, the descriptions must focus on distinguishing quality, rather than measuring quantity. Descriptions which differentiate levels of performance primarily with quantifiers (e.g., few, some, many) should be avoided also because it is difficult to consistently measure "few" of any aspect of a language performance. For instance, rather than a focus on the number of errors, the descriptions for a category for grammatical accuracy in a language classroom can include phrases about whether the errors interfere with understanding the speaker's message. Likewise, descriptions should not attempt to differentiate levels of performance with relativistic or comparison terms (e.g., exceptional, stronger) because such terms may encourage scoring in comparison to others rather than to the stated objective. Further, since the categories in the rubric focus on the language skills (e.g., grammar, vocabulary) that contribute to a successful performance, the categories should not include requirements of the assessment itself, such as a word count or time fulfillment. In fact, student learning can be hindered by rubrics that focus on the assignment requirements (or directions) rather than describing the quality of the work (Brookhart, 2018). Failure to meet the requirements of the assessment can be penalized, ideally outside of the rubric's categories because rubrics are a tool to assess the quality of the language performance.

It is best practice to have only "as many scale points as can be well defined and that adequately cover the range" of performances (Perlman, 2002, p. 8). Generally, rubrics have between three and five levels of performance (Brookhart, 2018; Suskie, 2009). With too many ranked levels, the task of matching a learner's performance to the most similar level becomes difficult and more time-consuming. Additionally, the number of categories in the rubric may influence the number of levels. For instance, an analytic rubric with a higher number of categories may, in turn, have fewer levels of performance, in order to keep the rubric a manageable size. The example rubric in Table 3 has four levels of performance arranged with the lowest performance in the column next to the categories and each higher level of performance builds up to the highest performance in the rightmost column.

Table 3. Analytic Rubric: Formal Informational Speech

Category	**In Progress (below D)** *3 points*	**Approaching Standard (C-)** *4 points*	**Meets Standard (B+)** *4.5 points*	**Exceeds Expectations (A)** *5 points*
Vocabulary	Words often unsuitable for task; errors interfere with meaning	Word choice is limited or reduces effective expression of meaning	Variety of words; word choice may sometimes be too specific or vague but meaning is clear	Variety of words; word choice supports meaning & appropriate for task
Grammar Structures	Limited variety of structures & transitions; errors interfere with meaning	Variety of useful structures & transition words; some choices reduce effective expression of meaning	Variety of structures; useful transition words; grammar choices may be generic but appropriate for meaning	Variety of complex structures & effective transition words; grammar choices support meaning
Pronunciation	Substitution of sounds, suprasegmental patterns interfere with meaning; difficult to understand by all audiences	Sounds substitutions and suprasegmentals occasionally slow down comprehension but overall meaning is understandable to people familiar w/ENL speech	Minor, predictable sound substitutions which are understandable to people unfamiliar with ENL speech	Segmentals and suprasegmental patterns are understandable to all audiences
Fluency	Pacing, pausing, and/or fillers interferes with comprehen-sibility	Pacing, pausing and/or fillers slows down comprehensibil-ity	Appropriate pacing and pausing	Pacing and pausing enhance message
Content	Topic not clear; information may not be factual or relevant to topic	Speech is focused on a single unstated topic; information was factual but incomplete, general, or less relevant	Topic was explicitly stated; information was factual and relevant to the topic and generally supported	Topic was explicitly stated; information was factual and relevant to the topic; specific details enhance message

The descriptions of the levels of performance could be created based on previously scored learner performances, for example, by sorting the graded performances and identifying the features that differentiate the quality at each level for each category. Often, however, the instructor must create the descriptions at each performance level based only on the course expectations and previous experience with the teaching context.

Points and Weighting

When designing classroom assessment, considerations for assessing student work and assigning a score to the assessment must be addressed. In Table 1, we presented a rubric that described student performance without assigning a point value, but instructors and students may benefit from assigning points to performance because assigning points allows the instructor to grade the performance and the learner has concrete feedback of the distance between their performance and the expected standards of the activity. In other words, the levels in a rubric serve to assign a score and provide feedback to the learner (Brown, 2018; Goldberg, 2014). In this section, we explain setting the point value of each performance level in relation to other levels and the weight of each category in relation to other categories. At the end of this section, we explore the question of using a range of scores for each performance level.

When assigning point values, the difference between point values should represent the difference in quality between the levels. For instance, in Table 2, the point values in the "Meets Standard (B+)" level is only slightly lower than the "Exceeds Expectation (A)" level because the quality of the performances described is only slightly less successful. It is also important to consider how the resulting grade percentage and/or letter grade align with the descriptions of performance. For instance, if the second-best level of performance equates to a "B" performance, the numbers assigned at that level should equal a percentage in the B range. Note also that according to the rubric in Table 2, meeting the standard is not sufficient to earn full points, which may not be appropriate in some teaching contexts. The point values also make sense mathematically and do not penalize students unintentionally. For instance, assigning point values of 1, 2, 3, and 4 when there are four levels of performance may be incongruous for grading, unless the descriptions reflect vastly different performances which should earn 25%, 50%, 75%, and 100%.

Additionally, analytic rubrics allow the instructor to weight the categories (Green, 2014; Popham, 1997), which means that a specific category can be worth more points in the analytic rubric, based on the course focus, unit focus, or learner level. For instance, the rubric shown in Table 4 lists five total categories with the categories of fluency and content worth half the value of the other categories, indicating the relative importance of those categories. Ballard (2019) found that the arrangement of the categories may inadvertently influence raters' perception of which categories are most important, in that the leftmost column may be perceived as most important while the rightmost column may be perceived as the least important. This finding suggests that weighted categories should be placed left-most or topmost to align with rater's expectations, to avoid causing a conflict in expectations. Finally, when weighting categories, it is more efficient to list the point values in the individual cells, which can be circled, rather than having a multiplication notation (e.g., x2 for a category) because the notation requires another calculation for the rater and is often confusing to (or overlooked by) learners.

Table 4. Weighted Analytic Rubric: Formal Informational Speech

Category	In Progress (below D)	Approaching Standard (C-)	Meets Standard (B+)	Exceeds Expectatioı (A)
Vocabulary	Words often unsuitable for task; errors interfere with meaning (6)	Word choice is limited or hinders effective expression of meaning (8)	Variety of words; word choice may sometimes be too specific or vague but meaning is clear (9)	Variety of words; word choice supports meaning & appropriate for task (10)
Grammar Structures	Limited variety of structures & transitions; errors interfere with meaning (6)	Variety of useful structures & transition words; some choices hinder effective expression of meaning (8)	Variety of structures; useful transition words; grammar choices may be generic but appropriate for meaning (9)	Variety of complex structures & effective transition words; grammar choices support meaning (10)
Pronunciation	Substitution of sounds, supra-segmental patterns interfere with meaning; difficult to understand by all audiences (6)	Sounds substitutions and suprasegmentals occasionally hinder but overall meaning is understandable to people familiar w/ENL speech (8)	Minor, predictable sounds substitutions and suprasegmentals which do not interfere with meaning; understandable to people unfamiliar with ENL speech (9)	Segmentals and suprasegmental patterns are understandable to all audiences (10)
Fluency	Pacing, pausing, and/or fillers interferes with comprehensibility (3)	Pacing, pausing and/or fillers hinders comprehensibility (4)	Appropriate pacing and pausing (4.5)	Pacing and pausing enhance message (5)
Content	Topic not clear; information may not be factual or irrelevant to topic (3)	Speech is focused on a single unstated topic. Information was factual but incomplete, general or less relevant (4)	Topic was explicitly stated; information was factual and relevant to the topic and generally supported (4.5)	Topic was explicitly stated; information was factual and relevant to the topic; specific details enhance message (5)

Some rubrics include a range of points within each level of performance (e.g., 18-19-20). This option adds flexibility to the scoring, but it counters one of the main purposes of the analytic rubric, namely to describe the features of each level of performance for the rater and for the learner. As Perlman (2002, p. 8) has stated "each point on the scale should be clearly labeled and defined". Therefore, a range of scores within a level lowers the reliability because the rubric does not include guidelines for awarding the points within each level; thus, the instructor may not consistently score student work as an 18 vs. 19 vs. 20 when each value is connected to the same description of performance. Additionally, such rubrics are less helpful for the student to know how to improve or for the instructor to know what to teach. Accordingly, a range of available points within levels of performance should be used sparingly in classroom assessment. An alternative solution is to describe the levels of performances for specific point values (e.g., 20, 18, 16, and 14) and leave scores in-between (e.g., 19, 17, 15) unspecified. The rubric can note that the scores in-between the described levels of performance would represent a performance slightly stronger than the lower score but not quite fulfilling the description of the higher score. (See Appendix B for an example.) This formatting allows performances to be placed in a single point in the scoring according to how it maps onto the descriptions or in between the stated descriptions. This option may be the best balance between reliability and practicality (e.g., time, effort) when assessing tasks where the performances will vary greatly and when differentiating slight differences in quality is desired.

Pre-use Review

As with all materials development, rubrics require a careful review and an iterative revision process, particularly the descriptions of the levels of performance. This revision process, labeled as pre-use review, involves 1) ensuring that the expectations in the descriptions match the teaching context, 2) checking for consistency in the performance descriptors across levels, 3) considering the practicality of the rubric, and 4) improving the beneficial consequences of the rubric on student assessment and language development. These pre-use review components are described below.

First, the descriptions of level of performance should describe expectations appropriate to the context. Recall that descriptors must focus on the quality of performance, allow the rater to identify the appropriate level of the student performance, and provide the student with insight into their current and potential learning. The descriptions can sometimes unintentionally muddle the underlying construct represented in the learning objective that the category is meant to evaluate. For instance, Isaacs (2014) warned about including allusions to accentedness because learners of a language will commonly have an accent, but the accent itself does not prevent listeners from comprehending the speaker. In fact, the descriptions should not expect native-speaker-like performance (Green, 2014) because in most contexts learners should not be expected to have native-like performance, for instance in pronunciation (Isaacs, 2014; Purpura, 2016).

Second, the descriptions across the levels of performance should display parallelism, including consistency in language and syntax (Goldberg, 2014) so that the learners' work can be consistently assessed across the levels. Further, the descriptions should be concise without "dysfunctional detail," which includes long descriptors that try to describe every possible nuance within each level of performance (Popham, 1997, p. 74).

Third, the resulting rubric should be practical. Well-crafted descriptions of levels of performance will make the grading process more efficient (Ambrose et al., 2010) when the levels are easily compared. Practicality is further improved when the descriptions help the rater focus on the specific features which differentiate levels. Font enhancements, such as bolding, may be useful, when used judiciously (Goldberg, 2014). Then, extraneous and redundant wording perhaps can be eliminated. Depending on the context, phrases (rather than sentences) may be sufficient and efficient to describe the performances. When fewer words are used in the descriptions, the grading can be quicker because the instructor has less to compare to identify the most appropriate level. It is also important to review if the rubric's format is easy to use.

Analytic rubrics can be visually crowded, so spacing and formatting should be carefully considered. At the same time, a one-page rubric is likely easier to use than having categories or descriptions of levels of performance across pages.

Fourth, a well-constructed rubric also improves the beneficial consequences of the assessment because classroom assessments should be in service to learning. An analytic rubric can also clarify the instructor's expectations, serving as a blueprint for success for the learners, a benefit of the assessment process known as positive washback (Green, 2014). Accordingly, the categories and the descriptions of the levels of performance should be understandable to all users, the instructors who will assess and the learners who will be assessed (Moskal, 2002; Pui et al., 2020), so that the grading process is transparent and the scores provide feedback to the learner (Brown, 2018; Goldberg, 2014). While reviewing the rubric for transparency, think about how to explain the rubric's components and scoring procedure to the students. In addition to giving a score for each of the categories, instructors may give individual feedback on the language performance. The rubric formatting can facilitate feedback by including a place for comments, either for each category, or general comments, perhaps below the rubric.

Post-use Evaluation

After using any assessment tool, an evaluation and reflection of the assessment process is necessary so that the instructor can reflect on the teaching implications (Ambrose et al., 2010) and the opportunities for improving the assessment. Despite careful design, no rubric will be perfect. Through a review process, both the rubric and the assessment process can be improved. The following content presents specific, feasible suggestions for reviewing a rubric after using it for classroom assessment.

The post-use review should include an investigation of the resulting scores. The class mean (average) and the range (the lowest and highest scores) for the assessment should be calculated to consider whether the scores are meaningful and useful for understanding the learners' skills. It is also useful to review the mean and the range of scores for each section of the assessment. If the learners' performances were organized in rank order, lowest to highest score, they should reveal a corresponding increase in quality. Any discrepancies can be reviewed to improve the assessment's scoring. Additionally, any two or more performances which were given the same score in the stated scoring system should have equivalent quality. If the post-use review suggests that they are not, the scoring system may be revised to reflect the difference in quality. Of course, any changes should align with the purpose of the assessment and the assessment's stated learning objectives.

Both the categories and the descriptions of the levels of performance of analytic rubrics should be carefully reviewed after using the rubric. A key principle of an analytic rubric is that the categories are distinct with each giving valuable information about the learner's skills (e.g., Youn, 2015). The independence of the rubric categories can be reviewed by checking if any two categories pattern together. For instance, speakers who are rated as "meets expectations" in category X have also been rated "meets expectations" in category Y; and speakers rated "insufficient" in X have also been rated as "insufficient" in Y. If this is the case, those categories may be confounded or intertwined based on how the descriptions are written. With confounded categories, learners who are deficient (or superior) in one category will not be able to score higher (or lower) in another category; a rubric with confounded categories is incongruent with a main advantage of analytic rubrics, which is assessing various skills separately. To remedy this issue, the descriptions can be revised so that the categories are independent, ensuring that the skill level in one category does not limit the speaker's opportunity to demonstrate skill in another category. Another possibility is that the categories are inherently connected. For instance, productive vocabulary use and collocation use both increase with proficiency as part of vocabulary development (Bonk, 2011). If the two categories are inherently related, those categories could be combined in the rubric, or one category could be eliminated, knowing that the other category has been shown to effectively measure the construct.

Research has shown that raters sometimes evaluate language performances using their own ideas of what is relevant (Fulcher, 2003), perhaps implicit, unstated expectations (Moskal & Leydens, 2000). Consequently, it is important to honestly consider whether the scoring followed the descriptions in the rubric. Consider possible sources for any inconsistency and how the assessment could help the rater evaluate more consistently. Further, the best learners' performances can be reviewed for shared features; such a review may suggest features that better differentiate quality of performances. If it was difficult to place student work into one of two levels of performance in a certain category, those descriptions of level of performance can be clarified or more finely separated to capture the differences among the learners' performance in that particular context.

A post-use review of an analytic rubric should also look carefully at the number of students placed at each level of performance. If the majority of learners have not met the stated level of performance, further instruction/practice is needed so that the learners are meeting the objective or the stated expectations in the rubric may need to be adjusted to the teaching context. Conversely, if most learner performances earned the same score in one category in the rubric, that category fails to differentiate between levels of performance. That result may, in fact, be appropriate, such as when all learners have met the stated standard. With a standards-based assessment, each learner's work is assessed to the stated standard (not to the other learners), which means potentially every student can pass or even get full points (Green, 2014).

Alternatively, if all students appear to perform at the same level, this result may indicate that the descriptions of levels of performance for that category are too broad. If so, an adjustment should be made to better reflect the context, specifically differences in proficiency for those learners. Typically, the full range in the rubric should be used, sorting learners into each of the rubric's levels of performance *because* the classroom assessment is designed for those specific learners. If, for instance, a post-use evaluation reveals that no performance was rated as "inadequate" for any category, the instructor should consider deleting that level when none of the performances of your particular learners fit the description for that level. On the other hand, an unused lowest level of performance can be used to shield learners from negative affect. Even though the ratings will accurately describe the learner's performance, the rubric format can allow every learner to be above the "worst" description. Likewise, the same inquiry can be considered within each category, where superfluous descriptions can be deleted. Generally, when the rubric can be simplified by deleting unnecessary descriptions, practicality is strengthened. The practicality can also be evaluated with a review of how easy it was to grade the student work with the rubric. If the rubric (or a section of the rubric) was difficult to use, consider how the assessment could be improved. Often, simply adjusting the order of the evaluation categories (to match the order during the assessment process) or adjusting the physical layout of the rubric improves practicality.

Summary

An analytic rubric is a tool to more objectively and reliably measure the *quality* of the language performance; it is a tool to assess language production in open-ended tasks where the learner has freedom to decide what vocabulary and grammar constructions to use. Despite the recognition of the importance of assessing productive language skills, the assessment of language performance has several challenges. No rubric is perfect; therefore, revision before and after using it is always necessary. The steps listed below provide a summary of the process for constructing strong analytic rubrics for classroom use. (Note: A more detailed list in provided in Appendix A):

- **Step 1: Categories:** Identify categories that reflect separate skills of the stated learning objectives.

- **Step 2: Levels of performance:** Describe the expected levels of performance in each category appropriate for the context.

- **Step 3: Pre-Use Review:** Review the rubric for validity (e.g., categories are aligned with the assessment's stated learning objectives), reliability (e.g., performances can be consistently scored with the descriptions), practicality (e.g., rubric is easy to use), and the beneficial consequences of using the rubric.

- **Step 4: Post-Use Evaluation:** Check that the scores are meaningful and based on the descriptions, the categories are independent, the descriptions are level-appropriate, and the rubric is easy to use.

The effort to carefully design a rubric for a language assessment, while challenging, improves grading consistency and transparency during the grading. Analytic rubrics are a powerful tool for open-ended tasks to help the instructor evaluate their own teaching, document evidence of learner progress, and give feedback to learners.

About the Authors

Mary Lou Vercellotti is an Associate Professor of English at Ball State University where she teaches courses in linguistics and TESOL, including an assessment course. Dr. Vercellotti earned her PhD in Applied Linguistics at the University of Pittsburgh. Her research interests include investigating language development of adult instructed learners of English, often using speech data.

Dawn E. McCormick is a Senior Lecturer and the Director of the English Language Institute in the Department of Linguistics at the University of Pittsburgh. She is interested in IEP program administration, professional development for language teachers, and student self-assessment and self-correction.

Acknowledgements

The authors thank the participants of the "Creating Rubrics for Assessing Language Performance in the TESOL Classroom" workshop at the 2019 TESOL International Pre-conference Institute. The authors also thank Taylor Lutz, Virginia Robson, and M. Christine O'Neill, and the journal's reviewers for giving feedback on a draft of this article.

To cite this article:

Vercellotti, M. L. & McCormick. D. E.(2021). Constructing Analytic Rubrics for Assessing Open-Ended Tasks in the Language Classroom. *Teaching English as a Second Language Electronic Journal (TESL-EJ), 24*(4). https://tesl-ej.org/pdf/ej96/a2.pdf

References

Ambrose, S. A., Bridges, M. W., DiPietro, M., Lovett, M. C., & Norman, M. K. (2010). *How learning works: Seven research-based principles for smart teaching*. John Wiley & Sons, Inc.

Ballard, L. (2019) Analytic rubric format: How category position affects raters' mental rubric. In S. Papageorgiou & K. M. Bailey (Eds.) *Global perspectives on language assessment: Research, theory, and practice* (pp. 3-17). Taylor & Francis Group.

Bonk, W. J. (2001). Testing ESL learners' knowledge of collocations. In T. Hudson & J. D. Brown (Eds.), *A focus on language test development: Expanding the language proficiency construct across a variety of tests* (pp. 114-132). University of Hawai'i Press.

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education 3*(22). doi: 10.3389/feduc.2018.00022.

Brown, J. D. (2018). Rubrics. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 1436 – 1440). SAGE Publications.

Brown, H. D., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices*. Pearson Longman.

Fulcher, G. (2003). *Testing second language speaking*. Pearson Longman.

Goldberg, G. L. (2014). Revising an Engineering Design Rubric: A case study illustrating principles and practices to ensure technical quality of rubrics. *Practical Assessment, Research & Evaluation, 19*(8). https://doi.org/10.7275/vq7m-e490

Green, A. (2014). *Exploring language assessment and testing*. Routledge.

Green, A., & Hawkey, R. (2012). Marking assessments: Rating scales and rubrics. In C. Coombe, P. Davidson, B. O'Sullivan & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 299-306). Cambridge University Press.

Isaacs, T. (2014). Assessing pronunciation. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 1, pp. 140-155). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118411360.wbcla012

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*(3), 397-421. https://doi.org/10.1177/0265532209104668

Moskal, B. M. (2002). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research, & Evaluation, 8*(14). https://doi.org/10.7275/jz85-rj16

Moskal, B. M. & Leydens, J. A. (2000). Scoring rubric development: Validity and Reliability. *Practical Assessment, Research & Evaluation, 7*(10).

Perlman, C. (2002). An introduction to performance assessment scoring rubrics. *Understanding scoring rubrics: A guide for teachers* (pp. 5-13). ERIC Clearinghouse on Assessment and Evaluation.

Popham, W. J. (1997). What's wrong–and what's right–with rubrics. *Educational Leadership 55*(2), 72–75.

Pui, P., Yuen, B., & Goh, H. (2020). Using a criterion-referenced rubric to enhance student learning: A case study in a critical thinking and writing module. *Higher Education Research & Development.* https://doi.org/10.1080/07294360.2020.1795811

Purpura, J. E. (2016). Second and foreign language assessment. *Modern Language Journal, 100*(S1)*,* 190-208. https://doi.org/10.1111/modl.12308

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of a public speaking competence rubric. *Communication Education, 61*(3), 205-233. https://doi.org/10.1080/03634523.2012.670709

Suskie, L. (2009). *Assessing student learning: A common sense guide.* Jossey-Bass.

TESOL International Association. (2018). *The 6 principles for exemplary teaching of English learners, grades K-12*. TESOL Press.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3-12.

Van Gorp, K., & Deygers, B. (2014). Task-based language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 2, pp. 578-593). John Wiley & Sons, Inc.

Van Moere, A. (2014). Raters and ratings. In A. J. Kunnan (Ed.), *The companion to language assessment* (Vol. 3, pp. 1358-1374). John Wiley & Sons, Inc.

Vasylets, O., Gilabert, R., & Manchón, R. M. (2017). The effects of mode and task complexity on second language production. *Language Learning, 67*(2), 394-430. https://doi.org/10.1111/lang.12228

Vercellotti, M. L. (2018). Finding variation: Assessing the development of syntactic complexity in ESL speech. *International Journal of Applied Linguistics, 29*(2), 1-15. https://doi.org/10.1111/ijal.12225

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199-225. https://doi.org/10.1177/0265532214557113

Appendix A

Constructing Analytic Rubrics for Assessing Open-Ended Tasks in the Language Classroom: Checklist

Choosing Analytic Rubrics

- Rubrics are tools to assess open-ended performance-based tasks.
- Rubrics are used when assessing longer language samples, especially monologic tasks.
- Analytic rubrics results provide information to instructors and students.

Categories

- Map each category to a skill listed in the learning objectives of the course or assignment.
- Check that categories reflect observable and measurable skills.
- Check that each category is independent.
- Limit the rubric to no more than five categories.

Levels of Performance

- Describe the expected level of performance (e.g., meets standard) for each category first.
- Describe the performance one level higher (e.g., exceeds standard) and/or one level lower (e.g., approaching standard) after the expected performance is set.
- Use descriptive language (rather than evaluative terms) to help learners understand what they should do to reach the next higher level.
- Describe quality rather than quantity.
- Consider the number of levels of performance in relation to the number of categories for ease of use.
- Check that point values make sense and that category weights are representative of course and activity foci.

Pre-use Review

- Review the descriptions for context appropriate expectations.
- Check for consistency among descriptions across levels.
- Evaluate if descriptions can be shortened while maintaining meaning.
- Evaluate formatting or text enhancement for ease of use.
- Consider how well the rubric's format facilitates students' understanding of the feedback.

Post-use Evaluation

- Consider whether the assigned scores adequately represent the quality of the performances.
- Confirm the independence of each category in the rubric.
- Evaluate whether each category was assessed consistently, following the descriptions at each level.
- Evaluate how useful the levels of performance are for this context and these learners.
- Evaluate how easy the rubric was to use.

Appendix B

Table 5. Analytic Rubric: Travel Advertisement – Low-Advanced Learners

	20 **Strong student example**	**19***	**18** **Great work**	**17***	**16** **Good work**	**15***	**14** **Not yet meeting expectations**
Content	factual information claims well-supported relevant and specific details		factual information statements generally supported relevant but perhaps vague details		factual but incomplete information claims minimally supported some details less relevant		some incorrect or irrelevant information confusing or misleading to audience in parts
Word Choices	variety of words appropriate for audience word choice supports meaning		variety of words appropriate for audience word choice may be too specific or general but meaning is clear		words may be repetitive word choice may reduce understanding of meaning in isolated parts		words often unsuitable for task and/or audience errors interfere with meaning
Grammar	variety of structures grammar choices support meaning		variety of structures grammar choices may be generic but appropriate for meaning		some variety of structures but some overused some choices reduce effective expression of meaning		limited variety of structures errors interfere with meaning
Accuracy	isolated errors do NOT interfere with meaning		repeated, predictable errors do NOT interfere with meaning		multiple, repeated errors distract from understanding meaning		multiple, repeated errors slows down understanding of meaning
*Fulfills the expectations of the lower level but not quite fulfills the expectations of the higher level							

[i] Students' proficiency levels are approximately B1 and B2 on the CERF scale