Scoring Approaches

Scales/Rubrics

Kuiken, F.; Vedder, I.

DOI
10.4324/9781351034784-14

Publication date
2021

Document Version
Final published version

Published in
The Routledge Handbook of Second Language Acquisition and Language Testing

12

Scoring Approaches

Scales/Rubrics

Folkert Kuiken & Ineke Vedder

Background

Language learners are often eager to know how proficient they are in speaking or writing a particular language. Also, teachers want to be able to assess accurately and reliably the proficiency level of their learners (Leclerq et al., 2014). Language proficiency can be expressed in simple terms like adequate/inadequate, sufficient/insufficient, pass/fail, etc. Although such terms seem to be straightforward and easy to use, they do not tell us exactly how well learners perform at a particular stage in their acquisition process. For this reason, "simple" ways of arriving at a score for language proficiency have developed into more complex systems for assessing second language (L2) proficiency, such as rubrics or scales, containing various bands or level descriptors.

In our definition of a rating scale, we follow Davies et al. (1999) who characterize a rating scale as "a scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged" (p.153). Typically, scales range from zero mastery via partial mastery to full mastery. A proficiency band or level allows some variation, but still, a given level has some characteristics that distinguish it from the level below and the one above. Band descriptors "describe in words performances that illustrate each level of competence defined on the scale" (McNamara, 2000, p. 40).

In the literature rating scales and scoring rubrics are synonymous, with the term rubric being used more often in North American contexts; in what follows we will use the term (rating) scale. As Davies et al. (1999) advocate, scales and descriptors "are not in themselves test instruments and need to be used in conjunction with tests appropriate to the population and test purpose" (pp. 153–154). It is thus important to keep in mind that any score in language assessment is the outcome of an interaction that involves not only the test-taker and the test, but also the prompt or task, the performance itself, the rater, the rating scale, the construct to be measured, and the theory on which the scale is based (Bygate, 2011; Weigle, 2002).

In what follows we discuss which type of scale is suitable for a particular purpose and what the descriptors of a scale should look like. We stress the importance of good-quality rating scales, taking into consideration reliability, validity, and practicality in use of the instrument. We then zoom in on the assessment of speaking and writing viewed from the perspective of second language acquisition (SLA), with a focus on complexity, accuracy, and fluency (CAF). Special attention is paid to the spoken and written performance elicited by means of real-world tasks, on the basis of two studies conducted within the framework of task-based language assessment (TBLA).

Key Concepts

***Rating scale/rubric*:** "[A] series of ascending bands of proficiency" against which language performances are judged. "It may cover the whole conceptual range of learner proficiency, or it may just cover the range of proficiency relevant to the sector or institution concerned" (Council of Europe, 2001, p. 40).

***Rating criteria*:** Aspects of the construct in terms of which a test-taker's performance will be evaluated, e.g., task achievement, linguistic accuracy, organization, fluency, etc.

***Scoring band*:** A level in a rating scale which describes the specific extent to which a language learner can do something in the language and what the linguistic features are of their language use. A rating scale comprises multiple bands.

***Descriptor*:** A statement that stipulates the characteristics of a language performance at a particular band level. A scoring band comprises one or more descriptors per rating criterion.

Key Issues

Several classifications of rating scales have been proposed. The most commonly cited categorization, with respect to type of scoring, is that of holistic versus analytic scales (Hamp-Lyons, 1991; Weigle, 2002). According to the way scales are constructed, so-called intuitive methods can be distinguished from empirical methods (Fulcher, 2003). Other possible distinctions are those between norm-referenced and criterion-referenced scales (Bachman & Palmer, 1996; Luoma, 2004) or between task-dependent and task-independent scales (Brown et al., 2002). Over the years, the advantages and disadvantages of each type, in terms of reliability, validity, and practicality, have been debated.

Holistic Versus Analytic Scales and Scoring Approaches

In essence, the distinction between holistic and analytic scales depends on the question whether a single—holistic—score is given to a task performance, or whether several features of performance are scored separately. Different types of holistic assessment are often distinguished: holistic scoring, primary-trait scoring and multiple-trait scoring. Because the latter partly overlaps with analytic scoring—as demonstrated by Weigle (2002)—in this chapter the two are combined.

Holistic Scoring. In holistic scoring a single score is assigned to a speech or text sample based on the overall impression of the performance. This score is given either impressionistically, or on the basis of a rating scale. An example of holistic scoring is the rating approach used for the TOEFL iBT® Writing Test, with scales that contain one holistic descriptor of the syntactic and rhetorical characteristics at each of six levels of writing proficiency. Figure 12.1 shows the lowest (0) and highest (5) level descriptors of the test's integrated-writing scale.

In the literature numerous advantages of holistic scoring have been mentioned (e.g., Fulcher, 2003; Knoch, 2011): fast and easy to use, and practical for decision-making because it only gives one score. It is flexible in that it allows many different combinations of strengths and weaknesses within a level. There are, however, also some disadvantages linked to holistic scoring (e.g., Bachman & Palmer, 1996). Probably the most important one is that a single score does not allow raters to distinguish between various aspects of speaking or writing, and therefore does not provide useful diagnostic information about a test-taker's speaking or writing ability. Holistic scores are also not always easy to interpret, as raters do not necessarily use the same criteria to arrive at the same judgments. For that reason it may be more difficult to obtain high rater reliability in holistic scoring (Ohta et al., 2018).

0	A response at this level merely copies sentences from the reading, rejects the topic or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.
5	A response at this level successfully selects the important information from the lecture and coherently and accurately presents this information in relation to the relevant information presented in the reading. The response is well organized, and occasional language errors that are present do not result in inaccurate or imprecise presentation of content or connections.

Figure 12.1 Level 0 and 5 of TOEFL iBT® test integrated writing rubrics (ETS, 2019).

Primary-Trait Scoring. Similar to holistic scoring, in primary-trait scoring a single score is allocated to a performance, but the scoring procedure is specifically designed for each individual task, or class of tasks, and the rating scale is defined with respect to the specific speaking or writing assignment. Primary-trait scoring includes a description of the task to perform, a statement of the primary-trait (construct) to be measured, a rating scale with level descriptors, samples of performance to illustrate each level, and explanations of why each sample was graded in the way it was (Weigle, 2002). An example of primary-trait scoring can be found in Lloyd-Jones (1977), where students are submitted to a narrative writing task: They receive a picture based on which they have to describe to a good friend what is happening from the perspective of one of the people in the picture. The resulting text is scored on four categories: overall judgment of the performed task, use of dialogue, point of view expressed in the performance, and use of tense.

The advantage of primary-trait scoring is that very careful and explicit statements are made with respect to a specific task. On the negative side, the process of primary-trait scoring is often long and complex, as each task or task type requires the development of a specific rating scale. The score also provides information about the ability of the student to perform that particular task or type of tasks only, rather than tasks in general. This means that the gain in explicitness and a stronger validity are offset against a reduction in the generalizability of the meaning the score provides (Fulcher & Davidson, 2007).

Multiple-Trait and Analytic Scoring. Multiple-trait and analytic scoring have in common that language performance is rated separately on several aspects or criteria. For that reason, as already mentioned, Weigle (2002) does not distinguish between multiple-trait and analytic scoring. Although we agree with Hamp-Lyons (2016) that multiple-trait scoring is not identical to analytic scoring, we follow Weigle's line of reasoning, as differences between the two scoring methods have more to do with procedures for developing and using the scales, rather than with characteristics of the scales themselves. Multiple-trait scoring is, so to say, on the edge of holistic scoring as discussed above and analytic assessment.

In multiple-trait or analytic scoring each score is representative of some performance feature or construct underlying performance. For example, in rating speaking fluency, raters may be asked to provide separate assessments of pausing behavior, repetitions, repairs, and the like. Analytic scales contain a number of criteria, usually three to five (Luoma, 2004, p. 68), each of which have descriptors at the different levels of the scale. For instance, in the ESL Composition Profile (Jacobs et al., 1981) texts are rated on five aspects of writing: content, organization, vocabulary, language use and mechanics, with multiple scoring criteria, each with their own descriptors (see Figure 12.2 for the scoring of content).

Probably the most cited advantage of multiple-trait or analytic scales is that they provide more specific information about learners' language proficiency and are therefore more reliable (see e.g., Hamp-Lyons, 2016). The explicit descriptors provide detailed guidance to raters to evaluate specific strengths and weaknesses of a learner's performance. This leads to more diagnostic

	Score	Descriptors
Content	30-27	EXCELLENT TO VERY GOOD: knowledgeable – substantive – thorough development of thesis – relevant to assigned topic
	26-22	GOOD TO AVERAGE: some knowledge of subject – adequate range – limited development of thesis – mostly relevant to topic, but lacks detail
	21-17	FAIR TO POOR: limited knowledge of subject – little substance – inadequate development of topic
	16-13	VERY POOR: does not show knowledge of subject – non-substantive – not pertinent – OR not enough to evaluate

Figure 12.2 Scoring of content in the ESL composition profile (Jacobs et al., 1981).

information about their abilities. As the scales comprise more detailed descriptors, inexperienced raters can more easily understand and apply the criteria. A drawback of this type of scaling is that it takes longer to evaluate the testees' performances (Weigle, 2002).

Other Possible Distinctions

Rating scales can also be categorized according to the way they are constructed. From that perspective, intuitively developed rating scales are distinguished from empirically based rating scales. Other distinctions concern norm-referenced versus criterion-referenced scales and task-dependent versus task-independent scales.

Intuitively Developed Scales Versus Empirically Based Scales. An intuitively developed scale is designed on the basis of what scale developers consider to be common features at various levels of proficiency. The scale may be developed by either an experienced teacher or language tester, starting with expert judgments, followed by elaboration and refinement by those who use it.

In recent years, several researchers have proposed that scales should be developed based on empirical methods, which can be done in different ways (Fulcher, 2003). For example, the scale may be data-based or data-driven, with bands and descriptors developed on the basis of the features observed by experts in a set of learner data. Another option is an Empirically Based Boundary definition (EBB) scale, where experts divide language samples into better or poorer performances and then develop a sequence of yes/no questions that characterize differences between the performance sets. These questions are then used as scoring guidelines by the raters (e.g., Does the writer display comfort with the use of English? No—Level 1 or below; Yes—Level 2 or 3 ⟶ Can the writer sustain an argument? No—Level 2; Yes—Level 3; see e.g., Ewert & Shin, 2015).

Nowadays, mixed-methods approaches (intuitive—qualitative—quantitative) are increasingly used for the development of rating scales (Green & Hawkey, 2012). For instance, in a study by Chan et al. (2015) on developing a rating tool to assess reading-into-writing skills, methodologies used included a questionnaire, expert panel judgment, group interview, automated textual analysis and analysis of rater reliability.

Norm-Referenced Versus Criterion-Referenced Scales. An issue relevant to scale design is whether the scale should be norm-referenced or criterion-referenced (Bachman & Palmer, 1996; Luoma, 2004). Norm-referenced scales allow comparison of learner performances against each other or against standards set by a norming group. Criterion-referenced scales imply comparison against some external criterion (e.g., the ability to perform a certain job). The latter allow test users to make inferences about how much language ability a test-taker has, rather than how well (s)he performs relative to other individuals (which constitutes norm-referencing).

Task-Dependent Versus Task-Independent Scales. Another distinction between scales is made by Brown et al. (2002) with respect to task-based performance. They distinguish between task-dependent scales—which they consider to be inefficient, as an entire domain of tasks has to be tested one-by-one, and task-independent scales—which estimate the learners' abilities to accomplish similar other tasks. Examples of both types are provided in the section "Recommendations for practice."

No One Best Scale Type. Presented as such, the above-described rating scales all seem to constitute a dichotomy and suggest sharp contrasts. In reality, however, most scales are often positioned on a continuum somewhere in between the two scale extremities. Another important point to note is that the choice for a particular rating scale—whether holistic or analytic, intuitively developed or empirically based, task-dependent or task-independent, norm-referenced or criterion-referenced—depends on the construct to be measured, the purpose for which the scale will be used, and how the score should be reported. In that sense, there is no one best scale type which fits all goals (Turner, 2013).

Scale Levels And Descriptors

Weigle (2002), stressing the importance of the use of explicit levels and scale descriptors, both for teachers, raters and students, lists the following criteria: 1) standardization, i.e., use of descriptors provides the instructor with a standard by which to assess performance efficiently; 2) consistency, i.e., descriptors are valuable for maintaining consistency of standards across different instructors and also within instructors over time; 3) self-assessment, i.e., when students are given scale descriptors in advance, they are aware of the criteria on which they will be judged, resulting in "more productive use of self-assessment" (Covill, 2012).

A particular challenge for scale design, however, is the scarcity of solid theoretical and empirical evidence about language learning/acquisition/development, both in general and with respect to the specific construct that is assessed. This makes it difficult to characterize levels of a proficiency scale in precise, distinctive terms.

Scale Levels. When developing a rating scale, the number of scale levels is an important issue to consider. Rating scales typically contain between three and nine levels. Luoma (2004) argues that the optimum is probably somewhere in the middle: The more levels there are, the more specific the feedback will be, and the easier it will be to show progress; the lower the number of levels, the more consistent the decisions. She therefore envisages four to six levels in a scale.

Ideally, the definition of each level should be independent of the ones above and below it on the scale. However, wordings frequently involve comparative statements, with one level descriptor relative to one or more others. Often, descriptors include quantifiers like *many*, *a few* and *little*, as in Ekiert et al.'s (2018) discourse appropriateness scale, with descriptors ranging from "completely discourse appropriate" to "completely discourse inappropriate." Raters, however, might interpret these quantifiers in different ways, which may result in lower interrater reliability (see Pill & Smart, Chapter 13, this volume).

Scale Descriptors. Referring to Weigle's (2002) earlier mentioned criteria of standardization, consistency and self-assessment, Luoma (2004) listed factors that should to be taken into account when designing descriptors. These are: be brief, be clear, be concrete, be explicit, be usable (practical), and be interpretable and comprehensible independently without reference to other descriptors. Depending on whether the scale is intuitively developed or empirically based, descriptors may be defined on *a priori* grounds, deriving from theoretical notions of what the test is meant to be measuring, or empirically, on the basis of an analysis of data from performances on test tasks. Another important consideration is the way in which performance at the top end of the scale is defined. Particularly in the past, many rating scales made specific reference to the assumed performance of native speakers. This is problematic, however, since performance of

native speakers is highly variable, related to educational level, and actually covers a range of positions on the scale (Hulstijn, 2015).

Related Issues

Lately, an increase in the use of integrated scales can be noticed. This goes hand-in-hand with increases in integrated assessments of language skills, particularly reading-into-writing. Research into the development and use of this type of scales has been rare until now, but there are some exceptions (see e.g., Chan et al., 2015; Ewert & Shin, 2015; Fulcher et al., 2011; Ohta et al., 2018).

Another issue which has been given more attention recently is the reliability and validity of scales (see e.g., Turner, 2013). In a study on the validity of L2 writing scales, Becker (2018) concluded that L2 practitioners must make principled and justified decisions about the scoring criteria that they include in scales when assessing students' writing performance. Rater training is crucial in that respect. As has been indicated by, for example, Rezaei and Lovorn (2010), rater training regarding effective scale creation and use will lead to higher reliability and validity (for more on rater training, see Pill & Smart, Chapter 13, this volume).

In order to ensure reliable and valid score inferences, rating scales also need to be regularly monitored and, if necessary, modified. Rating scales often are adopted in different local contexts and therefore need to be systematically reviewed and revised in relation to that context. A good example is the writing section of the Examination for the Certificate of Proficiency in English (ECPE), which has been extensively reviewed by Banerjee et al. (2015), resulting in a completely revised rating scale. Another example is the revision of the above mentioned ESL Composition Profile (Jacobs et al., 1981) by Janssen et al. (2015) for use in placement testing on an English for PhD students program in Colombia. In the latter study the authors made use of multi-faceted Rasch measurement (MFRM), by means of which it is possible to disentangle interactions between the traits of a scale, the consistency of scoring of each rater, and the relative difficulty of each trait, to evaluate the scale's functioning in the new context and to inform scale revisions.

Recommendations for Practice

As demonstrated by the language testing literature discussed above, many studies underline the importance of selecting and developing rating scales in accordance with the construct to be assessed, the intended use, the type of learners, and the L2 proficiency level for which the scale will be employed (e.g., Alderson, 2005; Knoch, 2011). A second recommendation is to empirically try out a rating scale before using it for language testing purposes. Thirdly, (re)validation of an existing rating scale—if used in a different learning context—is crucial, in order to test out its suitability in the new context.

To illustrate these recommendations, we now describe two empirical studies conducted within the framework of TBLA. TBLA aims to investigate how performance may be assessed through linguistic data collected by means of real-world instructional tasks in which the primary focus is on meaning (Ellis & Shintani, 2014). In this type of research various rating instruments have been employed, as shown by the two studies presented below, one by Kuiken and Vedder (2014, 2017, 2018) on the assessment of functional adequacy, the other by Gilabert and Barón (2018) on measurement of cognitive complexity and pragmatic competence.

Assessment of Functional Adequacy

Kuiken and Vedder (2014, 2017, 2018) developed a six-point Likert rating scale for the assessment of functional adequacy. Functional adequacy, as a construct, refers to the appropriateness of the utterances of the speaker/writer within a particular setting and on the basis of a specific target task (e.g., making a phone call to the dentist, taking part in a discussion).

1	2	3	4	5	6
None of the questions and the requirements of the task have been answered.	*Some (less than half)* of the questions and the requirements of the task have been answered.	*Approximately half* of the questions and requirements of the task have been answered.	*Most (more than half)* of the questions and the requirements of the task have been answered.	*Almost all* the questions and the requirements of the task have been answered.	*All* the questions and the requirements of the task have been answered.

Figure 12.3 Descriptors of task requirements in the scale for functional adequacy.

Positioned on the continuum holistic-analytic, the rating scale comes closer to the analytic pole. The rating scale is task-independent, and both theoretically and empirically based. Inspired by the conversational maxims of Grice (1975), the following dimensions of the construct are distinguished: task requirements, content, comprehensibility, coherence and cohesion. Task requirements focus on the extent to which the target task is adequately completed, in accordance with genre, register, speech acts, and specific task instructions. Content refers to the adequacy of the number and type of information units in the text. Comprehensibility takes into account the comprehensibility of the text and the amount of effort required from the listener/reader. Coherence and cohesion refer to the adequacy of the utterances of the speaker/writer, in terms of the occurrence of cohesive ties and coherence breaks. An example of the descriptors of the six bands of one of the four scale dimensions (task requirements) is presented in Figure 12.3.

To validate the scale a number of experiments were set up, for oral and written L2 and L1 production and for two different target languages: Dutch L2/L1 and Italian L2/L1 (for a full account, see Kuiken & Vedder, 2014, 2017, 2018). Two training sessions were organized, to familiarize the raters with working with the scale. In order to shed light on raters' employment of the scale and the ways in which they interpreted the different scale dimensions, bands and descriptors, a retrospective panel discussion was organized.

The rating scale has subsequently been tested out for different types of learners and various source and target languages. For instance, it has been tested for the assessment of writing and speaking development of native and non-native speakers of Italian in a research project involving twelve primary and three secondary schools in northern Italy (Pallotti, 2017), and for undergraduate university students of English in Spain (Herraíz Martínez, 2018). The scale has also been tried out with various types of tasks (e.g., decision-making, narrative, instructive, expository and information gap tasks).

What does the development of a scale for the assessment of functional adequacy tell us about the construction of a rating scale in general? First of all, the studies mentioned above—which led to some minor adaptations of the scale descriptors—confirm the importance of piloting and evaluating the reliability, validity, and applicability of the rating tool. Secondly, the results demonstrate that this functional adequacy scale has proven to be a reliable, valid, efficient, and task-independent instrument, which can be employed in different settings, for various task types, and for different source and target languages. Finally, the findings show the necessity of rater training and monitoring of raters over time (see Pill & Smart, Chapter 13, this volume).

Measurement of Cognitive Complexity and Pragmatic Competence

Gilabert and Barón (2018) investigated the relationship between task-complexity and pragmatic competence in the L2 writing of intermediate learners of English as a Foreign Language (EFL). The participants had to write a response to four e-mail messages, at different levels of task complexity. Fifteen expert raters, teachers of EFL, participated in the study.

1 The opening and closing markers are not used.
The adressee is not acknowledged.
The degree of formality is very low and inadequate.
Pragmatic expressions (such as requests and apologies) are awkward and inappropriate.
Mitigation is not used in the pragmatic expressions produced (such as requests or apologies).

6 The use of opening and closing markers is correct for the context.
The addressee is fully acknowledged and is addressed accordingly to his/her position.
The structure of the e-mail follows the expected level of formality.
Pragmatic expressons are fully pragmatically and linguistically appropriate.
Mitigation is always used in pragmatic expressions.

Figure 12.4 Level 1 and 6 of the pragmatics grid in Gilabert and Barón (2018).

In this study, different rating tools were used. All raters first answered a subjective perception questionnaire (nine-point Likert scale) on the required mental effort and perceived difficulty of the different versions of the e-mail task, to obtain a measure of task complexity. The raters also provided qualitative comments to motivate their decisions. For the assessment of pragmatic competence a holistic seven-point Likert scale was employed to rate the pragmatic appropriateness of the learners' e-mail messages.

Different from the analytic and task-independent scale used by Kuiken and Vedder, the two rating scales in Gilabert and Barón included only one scale dimension and are task-dependent. Whereas their task complexity scale was intuitive, their pragmatic rating scale was empirically-based. The scale descriptors, ranging from zero ("pragmatically inappropriate") to six ("fully pragmatically appropriate"), were based on the CEFR (Council of Europe, 2001, 2018) and focused on appropriateness of register, speech act, and politeness of learner output (see Figure 12.4). For a detailed account, the reader is referred to Gilabert and Barón (2018).

A first lesson learnt from this study is that, depending on rating objectives and constructs to be assessed, different rating instruments should be used. A second implication from the study is that expert judgments by experienced teachers can serve as a reliable way to discriminate among different complexity levels: Asking raters to evaluate task complexity is a readily available technique that does not require expensive equipment (eye-trackers, reaction-time software) and may provide valuable insights for performance assessment, task design and language pedagogy.

Testing Tips

- The choice of the type of rating scale should depend on the assessment purpose for which the scale is used, in relation to the construct and rating objectives.
- Scale dimensions of the rating scale should be derived from the construct to be assessed.
- Scale descriptors should be explicit and clear, and relevant to the target language task which the learner has to complete.
- Prior to implementation in research or testing, a scale should be piloted by raters on sample performances representing a range of performance levels, to evaluate the scale's overall quality and practicality.
- Rater training is necessary: Training sessions, panel discussions and practice materials are crucial for raters to become accustomed with working with scoring bands and descriptors.

Recommended Readings

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

This book can be recommended, particularly with regard to Chapter 4: Speaking scales (pp. 59–95), which contains many examples of various types of scales.

Weigle, S. (2002). *Assessing writing*. Cambridge University Press.

This book is a good resource for those who need practical advice for designing tasks and scoring procedures for writing tests, especially Chapter 6: Scoring procedures for writing assessment (pp. 108–139).

Crusan, D. (Ed.). (2015). The use of rubrics to assess writing: Issues and challenges [Special issue]. *Assessing Writing*, *26*.

This volume bundles a collection of articles looking into the development process, role and uses of rating scales in writing assessment in various contexts.

References

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Bachman, L. F., & Palmer, A. S. (1996). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing*, *26*, 5–19. https://doi.org/10.1016/j.asw.2015.07.001

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing scale. *Assessing Writing*, *37*, 1–12. https://doi.org/10.1016/j.asw.2018.01.001

Brown, J. D., Hudson, T., Norris, J. M., & Bonk, W. (2002). *An investigation of second language task-based performance assessments*. University of Hawai'i Press.

Bygate, M. (2011). Teaching and testing speaking. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 412–441). Wiley-Blackwell.

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20–37. https://doi.org/10.1016/j.asw.2015.07.004

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Council of Europe.

Covill, A. (2012). College students' use of a writing rubric: Effect on quality of writing, self-efficacy and writing practices. *The Journal of Writing Assessment*, *5*(1), 1–9.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge University Press.

Ekiert, M., Lampropoulou, S., Révész, A., & Torgersen, E. (2018). The effects of task type and L2 proficiency on discourse appropriacy in oral task performance. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 247–263). John Benjamins.

Ellis, R., & Shintani, N. (2014). *Exploring language pedagogy through second language acquisition research*. Routledge.

ETS. (2019). *TOEFL iBT® test independent and integrated writing rubrics*. https://www.ets.org/s/toefl/pdf/toefl_writing_rubrics.pdf

Ewert, D., & Shin, S.-Y. (2015). Examining instructors' conceptualizations and challenges in designing a data-driven rating scale for a reading-to-write task. *Assessing Writing*, *26*, 38–50. https://doi.org/10.1016/j.asw.2015.06.001

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

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, *28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Gilabert, R., & Barón, J. (2018). Independently measuring cognitive complexity in task design for interlanguage pragmatic development. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 159–190). John Benjamins.

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

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Speech acts* (pp. 41–58). Academic Press.

Hamp-Lyons, L. (1991). *Assessing second language writing in academic contexts*. Ablex.

Hamp-Lyons, L. (2016). Farewell to holistic scoring. Part two: Why build a house with only one brick? *Assessing Writing*, *29*, A1–A5. https://doi.org/10.1016/j.asw.2016.06.006

Herraíz Martinez, A. (2018). *Functional adequacy: The influence of English-medium instruction, English proficiency and previous language learning experience* (Unpublished doctoral dissertation). University Jaume I, Castellón, Spain.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.

Jacobs, H., Zinkgraf, S., Wormuth, D., Hartfiel, V., & Hugley, J. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Janssen, G., Meier, V., & Trace, J. (2015). Building a better rubric: Mixed methods rubric revision. *Assessing Writing*, *26*, 51–66. https://doi.org/10.1016/j.asw.2015.07.002

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*, 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing*, *31*(3), 329–348. https://doi.org/10.1177/0265532214526174

Kuiken, F., & Vedder, I. (2017). Functional adequacy in L2 writing: Towards a new rating scale. *Language Testing*, *34*(3), 321–336. https://doi.org/10.1177/0265532216663991

Kuiken, F., & Vedder, I. (2018). Assessing functional adequacy of L2 performance in a task-based approach. In N. Taguchi & Y. Kim (Eds.), *Task-based approaches to teaching and assessing pragmatics* (pp. 265–285). John Benjamins.

Leclercq, P., Edmonds, A., & Hilton, H. (2014). *Measuring L2 proficiency: Perspectives from SLA*. Multilingual Matters.

Lloyd-Jones, R. (1977). Primary trait scoring. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing* (pp. 33–69). National Council of Teachers of English.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

McNamara, T. (2000). *Language testing*. Oxford University Press.

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing*, *38*, 21–36. https://doi.org/10.1016/j.asw.2018.08.001

Pallotti, G. (2017). Applying the interlanguage approach to language teaching. *International Review of Applied Linguistics*, *55*(4), 393–412. https://doi.org/10.1515/iral-2017-0145

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing*, *15*, 18–39. https://doi.org/10.1016/j.asw.2010.01.003

Turner, C. E. (2013). Rating scales for language tests. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Blackwell/Wiley. https://doi.org/10.1002/9781405198431.wbeal1045

Weigle, S. (2002). *Assessing writing*. Cambridge University Press.