PASAA
Volume 59
January - June 2020

Idea Sharing: Are Analytic Assessment Scales More Appropriate than Holistic Assessment Scales for L2 Writing and Speaking?

Nathan Thomas
UCL Institute of Education
20 Bedford Way, London, WC1H 0AL, UK
Email: nathan.thomas.19@ucl.ac.uk

Abstract

Classroom assessment practices can be confusing for many teachers. Terminology is numerous and elusive. Different types of assessment serve different purposes. This short discussion paper's contribution originates from my own attempt to determine whether analytic scales would be more appropriate than holistic scales for assessing the L2 writing and speaking of young adults in classroom settings. A scoping search and subsequent review of the relevant literature seems to demonstrate that analytic scales tend to be more precise than holistic scales. If the purpose of the assessment is to provide feedback for learning, analytic scales are more appropriate. Conversely, holistic scales are acceptable if the assessment is not intended to provide implications for improvement. To expand on these general statements, this paper first discusses terminology necessary to understand academic texts on assessment and assessment scales. It then discusses the assessment of writing as performed by teachers, followed by the assessment of speaking performed by both students and teachers, focusing on a small number of studies selected for their relevance and applicability to practice. I hope this paper serves its 'idea sharing' purpose by providing a gentle introduction to the discussion of assessment for teachers with little to no experience in this area.

Keywords: writing assessment, speaking assessment, analytic scales, holistic scales

Introduction

Classroom assessment plays a major role in both teaching and learning. Assessment is defined by Black and William's (1998, p.2) in their often-cited definition as "all those activities undertaken by teachers, and by their students in assessing themselves, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged". Black and William's dual lens of both teacher and student self-assessment will be used to guide this paper. Although often conflated with the terms *testing* and *evaluation*, the scope of assessment is much broader. Distinguishing between these and other seemingly analogous terms is the first step for teachers to understand better the purpose and successful implementation of classroom assessment. Understanding what authors mean when they discuss common scales used in assessment is also crucial as each scale serves a different purpose. Making sense of this terminology is the first aim of this paper.

A secondary aim is to determine briefly, and at the exploratory level, whether analytic assessment scales are more appropriate than holistic assessment scales when assessing L2 writing and speaking in EFL classrooms. My motivation for this topic stems from my initial confusion regarding assessment scales for use in my own classroom—a feeling I am sure many other teachers experience since often insufficient training is provided. Previously, I tended to favor holistic scales, trusting my own intuition. However, now it is clear to me that analytic scales are necessary to enable the students to learn from the assessment. Though, practicality is a major concern, as analytic scales take additional time to complete and benefit from rater training (discussed later; see also Chinda, 2013). I found it rewarding to review the available research to guide my own assessment and hope other teachers may find some of the implications beneficial to their own practice.

To guide this essay, I will discuss several recent studies from the perspective of a practitioner, seeking to glean potentially useful information for others like me. I am not an expert in assessment, but like many practitioners, I have been tasked with designing assessments for students for over a decade and have had to piece together my own understanding of best practices through a critical reading of published studies. Much of this work is quite dense, so my aim in this idea-sharing paper is to distill, in short, particular pros and cons of different assessment scales, a basic understanding of which has improved my own assessment practices.

To narrow down the scope, studies discussed in-depth have been limited to those at the upper secondary and university level, where most of my recent experience lies (having taught in both China and Thailand). Focus is on studies that may provide implications for in-tact classrooms in upper secondary and university settings rather than large-scale assessments. In university settings, students will have most likely already achieved the necessary scores from large-scale, standardized tests (e.g. TOEFL, IELTS) to gain entry into university but may still need "to develop their language proficiency in ways which will enhance their academic performance *at* the university" (Read, 2016, p. 15, emphasis added; see also Read, 2015).

In the second half of this paper, I have centered my discussion on two studies in particular: one study involving teacher assessment of writing and the other on student assessment of speaking. This is due to the fact that in my previous context, I was expected to assess writing formally whereas speaking assessment was usually self-assessed by students more informally (i.e. as a pedagogical tool). I have chosen these studies because they offer unique perspectives and do not claim that they are representative of all available studies. Those looking for additional empirical support should consult the relevant literature relating to their own context. But first, a discussion of terminology is needed.

Coming to Terms with Terminology

Bachman (2004) describes the relationship between assessment, testing, and evaluation in the following way: a 'test' is a form of 'measurement' to conduct an 'assessment'. How the assessment is used and then interpreted can be delineated into either 'evaluation' (making judgements about a learner compared to a predetermined standard) or as 'description' (to be used as feedback for learning) (Bachman & Palmer, 2010). In this way, assessment can be both *formative* and *summative*, as originally proposed by Scriven (1967) and later adapted by Bloom, Hastings, and Madaus (1971) to mean on-going assessments to improve performance (formative) and assessment at the end of a teaching period for the purpose of judging performance and/or assigning a grade (summative). More recently, the terms assessment *for* learning, assessment *of* learning, and assessment *as* learning have been used to describe the roles of formative, summative, and self-assessment; in this case, assessment as learning (and self-assessment by extension) is defined as "when students reflect on and monitor their progress to inform their future learning goals" (Cheng & Fox, 2017, p. 6).

Regardless of the type of assessment, when assessing performance of productive skills (writing and speaking), rating scales are used. Bachman and Palmer (1996) describe two kinds: global scales (henceforth referred to as *holistic* scales to represent more common usage) and analytic scales. Holistic scales rely on a single, overall performance score, whereas analytic scales dissect speaking and writing ability into subskills that are scored separately and may then be combined to generate a total score if necessary. The use of holistic and analytic scales depends on a variety of factors but often begins with how the assessor conceptualizes language ability. According to Bachman and Palmer (1996), if language ability is viewed as 'a single unitary ability' (p. 208), then a holistic scale may be used; if language ability is viewed as component parts (e.g. pronunciation, fluency, and accuracy—although many analytic scales drill down into much more specific subcategories), then an analytic scale may be used. How an assessor defines the construct and then operationalizes that construct through the task type(s) in the assessment process largely influences the nature of the scale; validity and reliability are also influenced by the assessments' theoretical underpinnings (Bachman & Palmer, 1996; see also McNamara, 1996).

Classic interpretations of validity and reliability view these concepts as separate but related entities, dependent on the assessment itself (Chapelle, 1999). From this perspective, validity is decided by an evaluation of whether the assessment measures what it is intended to measure, as per Lado's (1961) original definition. Reliability is determined by whether a test is consistent and dependable in the outcomes it produces (Brown, 2003). More recent conceptualizations of validity see it as an interpretive process, heavily dependent on how the assessment is used, with reliability viewed as evidence of validity (Chapelle, 1999; Kane, 2013). For the purpose of this paper, validity and reliability are viewed using this more recent conceptualization that homes in on interpretation and usage (see also Bachman, 1990; Harding, 2018; Messick, 1989). It is also important to note that, as one insightful reviewer pointed out, when we discuss the validity and reliability of a particular type of assessment or scale, we should keep in mind that these two aspects of language assessment can be to a large extent affected by the skills and expertise in using the scale. Therefore, rater training is highly beneficial.

In my view, the process of designing or selecting an assessment scale should be guided by a series of questions in five stages: 1) What am I trying to assess? What am I going to do with this assessment? (Purpose); 2) How much time do I have to complete the assessment? What is feasible in my context? (Practicality); 3) Based on my purpose, is the scale valid? (Validity); 4) As evidenced by the scale's validity, is it reliable? (Reliability); and 5) Do I have the skills and expertise necessary to use this scale? Is training required? If so, do I have access to such training? The figure below serves as a guide (see Figure 1).

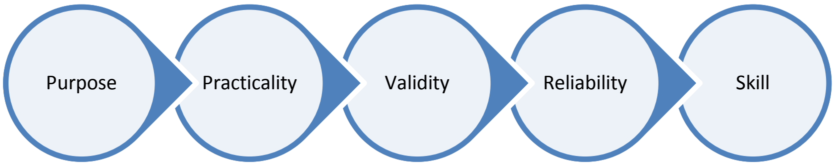

Figure 1: The decision-making process

Scales for Assessing Writing

Writing is an important part of secondary and university education. Students must be prepared to write in a variety of genres—often in an L2 such as English—as it has become the lingua franca of education in many contexts (Galloway & Rose, 2015). Nevertheless, writing is one area where assessment practices can significantly affect learning, since assessment provides direct feedback on the learning process (Hamp-Lyons, 2016). Yet, in order for assessment to be effective, the purpose that underscores the assessment process must be made clear (Bachman, 2004), and adequate rating scales must be developed that are both valid and reliable (Fulcher, 2003; Fulcher, Davidson, & Kemp, 2011). In comparing holistic and analytic scales for writing assessment, Weigle (2002) argued that analytic scales have higher reliability, validity, and are more useful in terms of providing assessment for learning, because isolated traits that are assessed independently provide additional data points for consideration.

Knoch (2009) drew attention to the fact that rating scales are often criticized for being intuitively developed. She built on Weigle's (2002) claim that analytic scales provide better assessment and questioned whether an empirically-developed scale designed to accommodate specific discourse-analytic criteria identified in student essays would have even higher validity and reliability than a more general analytic scale. As her findings confirmed her theory, in essence, Knoch (2009) further solidified the superiority of analytic scales to holistic scales as the former is typically seen as the more specific instrument. At the same time, Knoch (2009) created a third distinction, stating that intuitively developed analytic scales are not as effective at providing assessment for learning as empirically developed scales.

Nearly ten years later, rating scale development is still an area of immense interest among writing assessment researchers (see Becker, 2018; Lallmamode et al., 2016; Rakedzon & Baram-Tsabari, 2017). During the intervening years, methods of investigation have advanced, with more researchers exploring data-driven models and constructing the empirically developed scales that Knoch (2009) called for. Yet still, the dichotomy between holistic and analytic scales remains, as thorough analytic assessment is time-consuming and often impractical (Weigle, 2002). In one previous school that I worked at, I found holistic assessment to be the only viable option. I taught up to eight different classes at any given time, each with nearly 50 students in each class. However, as the following study and resulting discussion shows, there are now technological innovations that can help by targeting specific aspects using analytic scales.

In one very recent study, Vögelin, Jansen, Keller, Machts, and Möller (2019) investigated teachers' judgements of students' argumentative essays using both holistic and analytic scales. The holistic scale enabled the participants to score the writing as a single entity, while the analytic scale consisted of seven dimensions: "[f]rame of essay (introduction and conclusion), body of essay (organization of paragraphs), support of arguments, spelling and punctuation, grammar, vocabulary, and overall task completion," each with four descriptors ranging from "fully" – "mostly" – "partly" – "no(t)" (Vögelin et al., 2019, pp. 54-55). Thirty-seven pre-service teachers each assessed four essays in which the vocabulary profiles were manipulated to vary lexical diversity (see Jarvis, 2013) and lexical sophistication (see Nation & Webb, 2011), thus, establishing different proficiency levels. A high degree of lexical diversity and sophistication is typically considered representative of high-quality writing (Nation, 2013).

In Vögelin et al.'s study, the teachers were given both holistic and analytic scales and then asked to evaluate (i.e. grade) the four passages individually. The findings show that text quality did not have an effect on the holistic scores; however, text quality did have a significant effect on the analytic scores. High-quality texts were distinguished from low-quality texts using both scales, but the greater effect using analytic scales may illustrate the preciseness of such scales, which allow for more nuanced scoring and increased validity. If the scores were shared with students, the analytic scale could offer them specific areas for improvement. An additional finding worth mentioning is that vocabulary levels (as controlled by the researchers) affected scores regarding structural and grammatical judgements, not just the vocabulary criterion (see discussion below).

While it is not possible to make widespread assertions from just one study, given the sound experimental design (incorporating both holistic and analytic scales), Vögelin et al. (2019) provides support for the claim that analytic scales may be better equipped to help teachers make accurate judgements about student writing. The distortion of text structure and perceived grammatical accuracy due to the manipulation of vocabulary, an unrelated criterion, could cause feedback via formative assessment that is inaccurate and a grade via summative assessment that is incorrect. Because the participants in the study were pre-service teachers, it would be interesting to see how experienced teachers/raters would evaluate the texts with two different scales. If there were differences in rating based on experience, additional rater training would need to be conducted.

Furthermore, this study indirectly serves as an example of how software can be used as formative assessment with little effort on the part of the teacher. Since Vögelin et al. found vocabulary levels to be a unique indicator of text quality, students could self-assess their writing using an analytic scale and then compare it with a set of findings based on a specialized corpus (i.e. a collection of texts). Callies and Götz's (2015) volume on incorporating learner corpora in language assessment shows that this trend has already begun in various contexts. The next section will discuss student self-assessment of speaking, as one-to-one assessment has rarely been possible in my previous contexts—an issue teachers often face in settings where students far outnumber the teachers and time is limited.

Scales for Assessing Speaking

Like writing, the assessment of speaking can be influenced by the type of assessment and rating scale. Assessing speaking is difficult for a number of reasons. First, in classroom settings, the language is fleeting, and it can be difficult to determine in real time exactly what features students need to improve. In university settings, large class sizes may render any form of speaking assessment challenging. To mitigate this issue, as with the assessment of writing discussed above, many researchers are exploring the use of technology and self-assessment (e.g. Isaacs, Trofimovich, & Foote, 2017; Litman, Strik, & Lim, 2018; Purpura, 2016). But while new approaches to assessment begin to emerge, it is important that classic issues regarding scales and measures of validity and reliability do not get overlooked.

Babaii, Taghaddomi, and Pashmforoosh (2016) investigated the ability of 29 EFL university students in Iran to assess their own speaking and compared the students' self-assessments with those of six of their teachers. Three topics were chosen that would require the participants to perform different tasks. Upon completion of the tasks, the participants were asked to write down criteria they felt were important for assessing their speech. Then, they were asked to listen to their own recordings and self-assess their speaking, assigning a score based on their own criteria. The teachers, who were trained in testing and evaluation, were also asked to develop criteria from which to assess speaking, prior to hearing the participants' speech samples. The teachers' criteria consisted of ten items, ranging from fluency to time management, representing an analytic scale. After a 40-day interval, the participants were asked to rate their responses again, this time with the teachers' analytic scale. Both teachers and students evaluated the speaking samples independently using the same scale. Students were also asked to write brief reflections about their own speaking assessment.

Seven themes emerged from the students' reflections on their own speaking: topic management, confidence, fluency, time management, grammar, vocabulary, and pronunciation. This shows that the students critically considered elements often found on analytic scales. The students' self-assessment from the second scoring session more closely aligned with those of the teachers after being presented with the analytic scale and basic rater training. The students reported increased self-awareness and confidence in conducting self-assessment after using the teachers' scale, although some learners were still skeptical of their ability to self-assess accurately. Fourteen students expressed concern with maintaining a self-assessment schedule, which they feel could be beneficial if used on a regular basis.

These findings corroborate both previous and more recent studies where students have overestimated their ability to perform on certain speaking tasks (e.g. Dolosic, et al., 2016; Fay et al., 2008; Heilenman, 1990; Sadeghi et al., 2017). However, in one very recent study, Ma and Winke (2019) found that the intermediate-level participants were not as accurate in their self-assessment as those at the novice and advanced levels. Students whose self-assessment scores did not align with their actual score tended to underestimate rather than overestimate their speaking ability. Ma and Winke's (2019) findings should be treated with caution, however, since a binary coding scale was used that classified skills only as mastered or not mastered, representing what I consider to be a holistic self-assessment scale, generally accurate for performing summative assessment but lacking in pedagogical value. Moreover, such a holistic scale would not have been appropriate for the students in Babaii et al.'s (2016) study, as the students were in fact interested in specific aspects of their speech (as reported in their reflections) to improve their performance. The subjectivity of self-assessment has been criticized (see Brantmeier, Vanderplank, & Strube, 2012), but studies such as Babii et al. (2016) and Chen (2008) demonstrate improved alignment of self and teacher assessment after training and monitoring. These studies give way for new pedagogical implications. For instance, formative assessment using matching analytic scales can be administered at regular intervals by both teachers and students. Administering formative assessment in this way can lead to increased reliability between the scores and a subsequent lessening of teacher-as-rater time as the students become more proficient as raters. With the use of collaborative online spaces, it is also possible for students to become proficient at raters each other's speech, turning assessment into a collaborative learning activity. Increased validity and students' self-assessment can be corroborated with the teacher's assessment using the same scale (cf. Ma & Winke's, 2019, all-or-nothing holistic scale).

Conclusion

Over 30 years ago, Alderson (1988, 1990) drew attention to the fact that technology is changing the way we assess students' performance. Even more advancements have been made in modes of assessment and the types of rating scales that are available today. Those seeking assessment of English in secondary and university settings should seek out these new methods and embrace alternative assessment with their students. Computer software that enables the use of learner corpora and automatic or collaborative speaking and writing evaluation can be used as forms of both teacher assessment and student self-assessment. Specific criteria could be applied using analytic scales designed by teachers and students in collaboration with one another to create localized instruments. Ongoing rater training could be provided to students in the form of classroom lessons to enhance their own ability to self-assess.

Regarding the question of whether analytic assessment scales are more appropriate than holistic assessment scales, analytic scales tend to be more precise in the feedback they provide for learning and evaluation. However, the main factor to consider is purpose. For the most part, formative assessment should be continuous and systematic; it benefits from the use of analytic scales. Summative assessment that occurs at the end of a teaching cycle can be more holistic if the assessment is not intended to provide implications for improvement. In simplifying the discussion regarding Figure 1 above, there are three main questions that assessors should ask themselves when deciding what rating scale to use: 1) What exactly am I trying to assess? 2) What am I going to do with this assessment? 3) What is feasible in my context?

The Author

Nathan Thomas is currently a postgraduate researcher at the UCL Institute of Education, London, UK. Previously, he taught the English language in China and Thailand for 10 years. He has published in leading academic journals such as *Applied Linguistics*, *Applied Linguistics Review*, *ELT Journal*, *Language Teaching*, *System*, and *TESOL Quarterly*.

References

Alderson, J. C. (1988). Innovations in language testing: Can the microcomputer help? *Special Report No 1 Language Testing Update*. Lancaster: University of Lancaster.

Alderson, J. C. (1990). Learner-centered testing through computers: Institutional issues in individual assessment. In J. A. L. de Jong (Ed.), *Individualizing the assessment of language abilities* (pp. 20–37). Clevedon: Multilingual Matters.

Babaii, E., Taghaddomi, S., & Pashmforoosh, R. (2016). Speaking self-assessment: Mismatches between learners' and teachers' criteria. *Language Testing, 33* (3), 411-437.

Bachman, L. F. (1990). *Fundamental considerations in language testing* . Oxford: Oxford University Press.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford: Oxford University Press.

Becker, A. (2018). Not to scale? An argument-based inquiry into the validity of an L2 writing rating scale. *Assesing Writing, 37*, 1-12.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80* (2), 139-148.

Bloom, B. S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on the formative and summative evaluation of student learning.* New York, NY: McGraw-Hill.

Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. *System, 40*, 144-160.

Brown, H. D. (2003). *Language assessment: Principles and classroom practices.* London: Longman.

Callies, M., & Götz, S. (Eds.). (2015). *Learner corpora in language testing and assessment* . Amsterdamn: John Benjamins.

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19*, 254-272.

Chen, Y.-M. (2008). Learning to self-assess oral performance in English: A longitudinal case study. *Language Teaching Research, 12*, 235-262.

Cheng, L., & Fox, J. (2017). *Assessment in the language classroom: Teachers supporting student learning* . London: Palgrave Macmillan.

Chinda, B. (2013). Considerations in Performance-Based Language Assessment: Rating Scales and Rater Training. *PASAA, 46*, 141-159.

Dolosic, H. N., Brantmeier, C., Strube, M., & Hogrebe, M. C. (2016). Living language: Self-assessment, oral production, and domestic immersion. *Foreign Language Annals, 49*, 302-316.

Fay, N., Page, A. C., Serfaty, C., Tal, V., & Winkler, C. (2008). Speaker overestimation of communication effectiveness and fear of negative evaluation: Being realistic is unrealistic. *Psychonomic Bulletin & Review, 15* (6), 1160-1165.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28 (1)*, 5-29.

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

Galloway, N., & Rose, H. (2015). *Introducing global Englishes.* Abingdon: Routledge.

Hamp-Lyons, L. (2016). Purposes of assessment. In D. Tsagari, & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 13-27). Boston: De Gruyter.

Harding, L. (2018). Validity in pronunciation assessment. In O. Kang, & A. Ginther (Eds.), *Assessment in second language pronunciation* (pp. 30-48). London: Routledge.

Heilenman, K. L. (1990). Self-assessment of second language ability: The role of response effects. *Language Testing, 7* (2), 174-201.

Isaacs, T., Trofimovich, P., & Foote, J. A. (2018). Developing a user-oriented second language comprehnsibility scale for English-medium universities. *Language Testing, 35* (2), 193-216.

Jarvis, S. (2013). Defining and measuring lexical diversity . In S. Jarvis, & M. Daller (Eds.), *Vocabulary knowledge - Human ratings and automated measures* (pp. 13-43). Amsterdam: John Benjamins.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50* (1), 1-73.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26* (2), 275-304.

Lado, R. (1961). *Language testing: The construction and use of foreign language tests.* New York: McGraw-Hill.

Lallmamode, S. P., Mat Daud, N., & Abu Kassim, N. L. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing, 30* (1), 44-62.

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and oppotunities. *Language Assessment Quarterly, 15* (3), 294-309.

Ma, W., & Winke, P. (2019). Self-assessment: How reliable is it in assessing oral proficiency over time? *Foreign Language Annals, 52*, 66-86.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)* (pp. 13-103). New York : Macmillan.

Nation, I. S. P. (2013). *Learning vocabulary in another lanuguage (2nd ed.).* Cambridge: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary.* Boston: Heinle.

Purpura, J. (2016). Second and foreign language assessment. *The Modern Language Journal, 100* (S), 190-208.

Rakedzon, T., & Baram-Tsabari, A. (2017). To make a long story short: A rubric for assessing graduate students' academic and popular science writing skills. *Assessing Writing, 32* (1), 28-42.

Read, J. (2015). *Assessing English proficiency for university study.* Basingstoke: Palgrave Macmillan.

Read, J. (2016). Some key issues in post-admission language assessment . In J. Read (Ed.), *Post-admission language assessment of university students* (pp. 3-22). Springer.

Sadeghi, K., Mousavi, M. A., & Javidi, S. (2017). Relationship between EFL learners' self-perceived communication competence and their task-based and task-free self-assessment of speaking. *Journal of Research in Applied Linguistics, 8* (2), 31-50.

Scriven, M. (1967). *The methodology of evaluation (Vol. 1).* Washington, DC: American Educational Research Association.

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL . *Assessing Writing, 39*, 50-63.

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