PASAA
Volume 46
July-December 2013

CONSIDERATIONS IN PERFORMANCE-BASED LANGUAGE ASSESSMENT: RATING SCALES AND RATER TRAINING

Bordin Chinda
Chiang Mai University

Abstract

Performance-based assessment has gained more attention from ELT practitioners since the actual performances produced by students are evaluated in this type of assessment. However, the assessment of students' performances involves more complicated procedures when compared with more traditional testing methods. This paper, therefore, points out crucial considerations in adopting this type of assessment in a language class. First, the article introduces the concepts of performance-based language assessment and its major characteristics. Then, the two main characteristics: rating scales and rater training are discussed. In the rating scales section, different types of rating scales, as well as approaches in developing rating scales, are explored with the emphasis on the scales used for assessing writing. Finally, the paper presents the roles of rater training in performance-based assessment and how to conduct such training.

Keywords: Performance-based assessment; Rating scales; Rater training

Introduction

Though discrete point items of traditional testing have been the dominate mode of assessment in Thailand (e.g. Ordinary National Educational Test), with the arrival of communicative language teaching, language testing and assessment in many institutions has shifted to focus more on the actual performance of the students (Chinda, 2009, 2012). Traditional testing emphasises "the rank ordering of students, privileges quantifiable data for isolation, individual test performances, and in general promotes the idea of neutral, scientific measurement as the goal of educational evaluation"; whereas, the "alternative assessment" or performance-based assessment is based on "an investigation of developmental sequences in student learning, a sampling of genuine performances that reveal the underlying thinking processes, and the provision of an opportunity for further learning" (Lynch 2001, pp. 228 - 229).

McNamara (1996) states that a defining characteristic of performance testing is that "the assessment of the actual performances of relevant tasks are required of candidates, rather than the more abstract demonstration of knowledge, often by means of paper-and-pencil tests" (p. 6). Moreover, Davies, Brown, Elder, Hill, Lumley, and McNamara (1999) define performance-based assessment as "a test in which the ability of candidates to perform particular tasks... is assessed" (p. 144). Tasks, in the assessment of second language performance, are designed to measure learners' productive language skills through performances, which allow learners to exhibit the kinds of language skills that may be required in a real world context (Wigglesworth, 2008, p. 111).

Furthermore, Wigglesworth (2008), drawing from McNamara (1996) and Norris, Brown, Hudson and Yoshioka (1998) reports that there are three factors which distinguish performance tests from traditional tests of second language: (1) there is a performance by the candidate; (2) the performance is judged using an agreed set of criteria; and (3) there is a degree of authenticity of the assessment tasks (p. 113). Wigglesworth, based on the same sources, points out

that, based on the criteria used for judging the performance, there are two types of performance-based assessment. In the first type of performance-based assessment, tasks are used to elicit language to reflect the kind of real world activities learners will be expected to perform, and in which the focus is on interpreting the learners' ability to perform such tasks in the real world, with language being the means of fulfilling the task requirement rather than an end in itself. McNamara (1996) calls it a "strong" form of second language performance-based assessment or "task-based performance assessments" as termed by Norris, et al. (1998). In the second type of performance-based assessment, the tasks are used to elicit language samples for the purpose of rating, that is, the focus of the assessment is less on the task and more on the language produced. McNamara (1996) considers it a "weak" form of second language performance-based assessment, whereas Norris, et al. (1998) use the term "performance based testing".

Another important characteristic of performance-based assessment discussed by McNamara (1996) is "a new type of interaction, that between the rater and the scale; this interaction mediates the scoring of the performance" (p. 121). The figure below presents this characteristic of performance-based assessment.

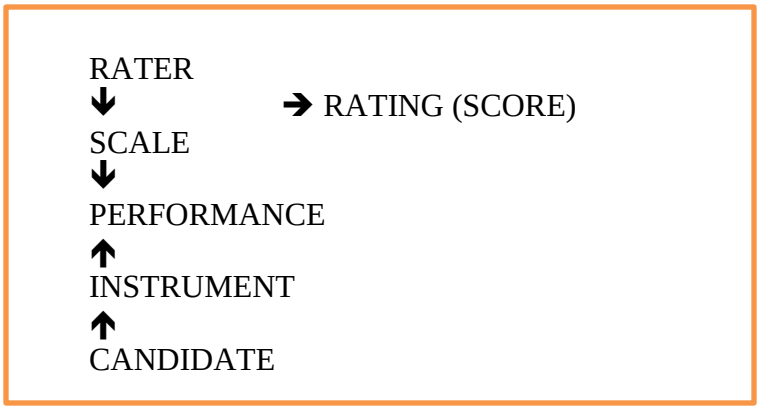

Figure 1: Characteristics of performance assessment (McNamara, 1996, p. 120)

According to the figure, the rater needs to use a rating scale in rating a performance to arrive at a score for that performance. In marking any performance-based assessment tasks, whether in the classroom context or in large-scale proficiency tests, the markers/ raters, or teachers in classrooms, are required to make more complicated judgements than the right-wrong decisions in multiple-choice, true/false, error-recognition, and other item types of testing situations where the candidate's responses can be marked as either "correct" or "incorrect". In this type of marking, or as it is sometimes referred to as subjective marking, Alderson, Clapham and Wall (1995) stress that the examiners' job is to assess "how well a candidate completes a given task", for which they need a "rating scale" (pp. 106 - 107). Therefore, this article explores two major aspects of performance-based language assessment: rating scales and rater training.

Rating scales

It should be noted that in the literature, different terms have been used to refer to a rating scale. For instance, Hudson (2005) reports that sometimes there is a clear distinction between the terms "rubric" and "scale" and sometimes they are conflated (p. 207). In this paper, the term rating scale is used. A rating scale (or proficiency scale) is a "scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged... The levels or bands are commonly characterised in terms of what the subjects can do with the language... and their mastery of linguistic features" (Davies, et al., 1999, p. 153). Rating scales also represent the most "concrete statement of the construct being measured" (Weigle, 2002). The statements in rating scales are commonly referred to as "descriptors" which describe "the level of performance required of the candidates at each point on a proficiency scale" (Davies, et al., 1999. p. 43).

According to Alderson (1991), rating scales can be categorised into three types depending on their function and intended audience:

- **User-oriented scales**, with a reporting function, aimed at enabling test users – for example, employers and admissions officers – to interpret test results by providing information about the typical behaviour of the students at any given level;
- **Assessor-oriented scales**, with a guiding of the rating process function, is aimed at describing guidance for the assessors who rate performances by providing typical performances by students at each level;
- **Constructor-oriented scales**, with the function of guiding the construction of the tests, aims to provide guidelines for test constructors by providing a set of specifications that students should be able to do at a given level.

In recent language testing and assessment literature, rating scales or scoring methods have been categorised differently by different researchers (e.g. Alderson et al.1995; Arter & McTighe, 2001; Davies, et al., 1999; Hamp-Lyons 1991; Shaw & Weir, 2007; Weigle, 2002). For instance, Hamp-Lyons (1991) identifies three types of scoring methods: holistic scoring, primary trait scoring, and multiple trait scoring. Weigle (2002), on the other hand, identifies three main types of rating scales: primary trait scales, holistic scales, and analytic scales. Weigle does not distinguish multiple-trait scales from analytic scales because she considers that the characteristics of multiple trait scales "have to do more with procedures for developing and using the scales, rather than with the description of the scales themselves" (p. 109). This article uses the terms *multiple trait scale* and *analytic scale* interchangeably. In addition, this paper only explores two types of scales: holistic scales and analytic scales because the primary trait scoring method has not been widely used in second-language assessment (Weigle, 2002. p. 110) but is generally used in research situations particularly in very large-scale data collection (Hamp-Lyons,1991).

Analytic versus holistic rating scales

With an analytic scale, raters are asked to judge several components of a performance separately, on the basis of traits, criteria, or dimensions of performance. These components are divided so that they can be judged separately rather than expecting the assessor to give a single score for the entire performance (Alderson et al. 1995; Arter & McTighe, 2001; Weigle, 2002). Arter and McTighe (2001) state that analytic scales are used when planning instruction to show relative strengths and weaknesses of a performance, when teaching students the nature of a quality performance, when giving detailed feedback, and when knowing how to precisely describe quality is more important than speed (p. 25). One main advantage of the analytic scoring method over the holistic counterpart is that it provides a higher reliability (Goulden, 1994). Weigle (2002) also agrees that compared to holistic scoring, analytic scoring is more useful in rater training, and is particularly useful for second-language learners, as it is more reliable. Moreover, Hamp-Lyons and Kroll (1997) have commented that "a detailed scoring procedure [i.e. multiple trait scoring] requiring the readers to attend to the multidimensionality of ESL writing, may ensure more valid judgement of the mix of strengths and weaknesses often found in ESL writing" (p. 29).

Furthermore, Hamp-Lyons and Kroll (1997) reported that a multiple trait scoring system "helps raters balance their judgments of characteristic ESL features of writing, principally a high frequency of low-order sentence grammar problems, against higher order elements of the writing…" (p. 29). However, Weigle (2002) recognises that the rating time that is necessary for analytic scoring takes longer than that of holistic scoring because raters need to make more than one decision for every script. She also adds that a good deal of the information provided by the analytic scale is lost when scores on different scales are combined to make a composite score (p. 120).

In contrast, with a holistic scale, raters are asked to give a judgement on a candidate's performance as a whole, or in other

words, a single score for an entire performance based on an overall impression of a candidate's work (Alderson et al. 1995; Arter & McTighe, 2001; Weigle, 2002). Thus, the scale used in this method is sometimes called an impression scale. Arter and McTighe (2001) state that holistic scales are used when the speed of scoring is more important than knowing precisely how to describe quality, when the performances are simple, and when a quick snapshot of overall achievement is the objective (p. 25). This type of scoring method, nevertheless, has been heavily criticised, especially in an EFL/ESL writing assessment context. Hamp-Lyons (1995, pp. 760-761) points out that:

> a holistic scoring system is a closed system, offering no windows through which teachers can look in and no access points through which researchers can enter. Scores generated holistically cannot be explained to other readers in the same assessment community; diagnostic feedback is, therefore, out of the question.

Furthermore, in the report for the Educational Testing Service (ETS), Hamp-Lyons and Kroll (1997, p. 28) point out the inherent nature of holistic scoring as being impression marking in a speed dependent manner. They state that "many raters make judgments by responding to the surface of the text and may not reward the strength of ideas and experiences the writer discuss".

Drawing from Bachman and Palmer (1996), Weigle (2002) provides a useful approach to making a decision in choosing between holistic scales and analytic scales in writing assessment. Table 1 below presents a comparison of the two types of rating scales based on the six qualities of test usefulness (for more detailed information on test usefulness, see Bachman & Palmer, 1996).

Table 1: A comparison of holistic and analytic scales based on six qualities of test usefulness (Weigle, 2002, p. 121)

Quality	**Holistic Scale**	**Analytic Scale**
Reliability	Lower than analytic, but still acceptable	Higher than holistic
Construct Validity	Holistic scales assume that all relevant aspects of writing ability develop at the same rate and can thus be captured in a single score; holistic scores correlate with superficial aspects, such as length and handwriting	Analytic scales are more appropriate for L2 writers as different aspects of writing ability develop at different rates
Practicality	Relatively fast and easy	Time-consuming; expensive
Impact	Single scoring may mask an uneven writing profile and may be misleading for placement	More scales provide useful diagnostic information for placement and/or instruction; are more useful for rater training
Authenticity	White (1995) argues that reading holistically is a more natural process than reading analytically	Raters may read holistically and adjust analytic scores to match holistic impressions
Interactiveness	n/a	n/a

Nevertheless, it should be noted that raters could evaluate the work with a "halo effect" when they employ an analytic rating scale. A halo effect is a rater's failure to discriminate among conceptually distinct and potentially independent aspects of a candidate's performance, or a rater's tendency to allow the overall impression of a candidate's performance to influence his or her judgement (Saal et al., 1980; King et al., 1980; cited in Myford & Wolfe, 2003). From the above discussion, it can be concluded that in an EFL context, like Thailand, analytic rating scales might be more applicable, as they provide more useful information for learners and teachers to make improvements in a rater training system (discussed below). Furthermore, according to Chinda (2009)'s study, teachers have found that analytic scales could help them rate students more reliably, even though they found the scales rather difficult to use when they first encountered them.

Approaches in designing rating scales

After the decision on the type of scale to be adopted, the subsequent step of equal importance is to design the scales. However, before designing the rating scales, there is another crucial decision to be made, which is to choose a designing approach. From the perspective of designing rating scales in a large-scale testing context, Hudson (2005) identifies two types of rating scales in relation to criterion-referenced task-based assessment: decontextualised and contextualised rating scales. Drawing from Brindley (1998), Hudson describes that the former scale is "defined independently of content and context… and is derived from a theoretical model of language, and attempts to define a decontextualized ability or proficiency" (p. 209); whereas the latter scale "is behaviourally based and attempts to describe proficiency according to 'real-world' performance in specific contexts" (p. 210). Within the behavioural scales, Hudson also identifies two main developmental approaches: intuitive approach, and empirical approach. However, this article only explores the contextualised

approach in developing a rating scale because a decontextualised rating scale is not relevant to teaching English in an EFL context.

In developing scales in terms of assessing speaking, Luoma (2004, pp. 83 - 86) identifies three methods (within the contextualised approach). The first is "the intuitive method" in which the development of a scale is based on a principled interpretation of experience. The developers, who are usually experienced in teaching and/or material development, may consult existing scales or a course syllabus, and then design the scales afterward. The second method is "the qualitative method". In this method, the developers ask groups of experts to analyse data related to the scale, which may be the descriptors or samples of the performances at different levels. Finally, the third method, "the quantitative method" requires a certain expertise in statistics, such as multidimensional scaling, linear regression, and with a knowledge of item response theory. This method which mainly addresses scale validation, is usually carried out by large-scale testing systems or research institutions, as it may require the collection of large data sets.

From another perspective concerning writing assessment, Weigle (2002, pp. 122-124) proposes that once a decision has been made about the kind of rating scale is to be adopted, holistic or analytic, the following factors should be considered:

- Who is going to use the scale?
- What aspects of writing are most important, and how will they be divided up?
- How many points, or scoring levels, will be used?
- How will scores be reported?

After these questions are addressed, the descriptors for levels/bands of the scale can be written. According to Weigle, there are two approaches: *a priori* and empirical. In the *a priori* approach, the "inherent" ability (for example, a student has ability x) being measured is defined in advance; whereas in the empirical approach,

descriptors are derived through an examination of actual performances. Shaw and Weir (2007), in addition, state that the design and development of the rating scales for the tests of writing has traditionally relied on an *a priori* approach, which is based on the experience of an expert and intuitive judgement (p. 162). Nevertheless, they point out that researchers have advocated for more application of the empirically-based approach in developing rating scales. In this approach, samples of actual performances are analysed to construct or re-construct assessment criteria and scale descriptors.

Furthermore, Turner (2000, 2001), Turner and Upshur (2002), and Upshur and Turner (1995) stress the advantageous aspects of empirically derived criteria. Upshur and Turner (1995) strongly believe that scales that are locally developed by teachers could create positive wash-back effects on teaching. They point out that because there are no restrictions upon the development of the scale descriptors, the descriptors derived from the interaction among the scale development team reflect the instructional objectives. In addition, the development process of the scales and descriptors "can lead to greater agreement on the aims of teaching" (p. 11), which can increase the validity of the assessment.

Moreover, Knoch (2009) expands Weigle's (2002) classification of rating scales (cf. Table 1) by illustrating the differences between the intuitive and empirically developed analytic scales. Table 2 summarises these features.

Table 2: A comparison of intuitively developed and empirically developed analytic scales (adapted from Knoch, 2009, p. 299)

Quality	Intuitively developed	Empirically developed
Reliability	Higher than holistic.	Higher than intuitively developed analytic scales.
Construct Validity	Analytic scales are more appropriate for L2 writers [than holistic] as different aspects of writing ability develop at different rates. But raters may choose to rate with a halo effect.	Higher construct validity as based on real student performance; assumes that different aspects of writing ability develop at different speeds.
Practicality	Time-consuming; expensive.	Time-consuming; most expensive.
Impact	More scales can provide useful diagnostic information for placement, instruction and diagnosis, but might be used holistically by raters; useful for rater training.	Provides even more diagnostic information than intuitively developed analytic scales; especially useful for rater training.
Authenticity	Raters may read holistically and adjust analytic scores to match holistic impressions.	Raters assess each aspect individually.

As illustrated in Table 2, and in the discussions by Turner (2000, 2001), Turner and Upshur (2002), and Upshur and Turner (1995), a more suitable approach in designing rating scales for a second language classroom could be an empirical approach, as the

information gathered during the designing process can be used for improving teaching and rater training (discussed below).

Rater training

Alderson, Clapham and Wall (1995) point out that one of the most important issues to consider in teacher assessment is rater monitoring. Alderson, et al (1995) state that training the examiners or raters could provide them with "competence and confidence" (p. 128). In other words, rater training must be provided for the examiners. A rater training prepares raters for the task of judging candidate performance. It mainly involves the process of the familiarising raters with the test format, test tasks, rating criteria, and exemplar performances at each criterion level (Davies, et al., 1999, p. 161). In order to improve the quality of rater-mediated assessment, McNamara (2000) emphasises the moderating meeting scheme, providing initial and on-going training to raters. Alderson, et al. (1995) also add that on a regular basis, tests should be routinely monitored, after each administration item and subtest analyses and a descriptive statistic analyses should be conducted, additionally, raters should be monitored, and post-test reports should contain information for any future modification. In the same vein, Davies, et al. (1999, p. 161) state that the reliability of the raters depends, partially, on the quality of their training, which aims to ensure a high degree of both inter- and intra-rater training. In addition, Lumley (2002) stresses that rater training and reorientation allow raters to "learn or (re)develop a sense of what the institutionally sanctioned interpretations are of the task requirements and the scale features, and how other related personal impressions of text quality are relevant to the rating scale provided", which increases the reliability of rating, overall (p. 267). It is, however, important to be aware that training on its own cannot guarantee that raters will mark as they are supposed to (Alderson et al., 1995, p. 128). In addition, Hamp-Lyons (2007) states that rater training can influence how teachers judge their students' language performances, but making judgements

still remains subjective, because it is based on the individual teacher's experiences. Davies et al. (1999, p. 161) state that:

> rater training shows that training reduces extreme differences in severity between raters and makes raters more internally self-consistent, but that significant differences in severity between raters remain; further, rater characteristics (relative severity, self-consistency) vary over time.

Following are 2 directions for conducting a rater training. Building on White (1984), Weigle (2002, pp. 130 - 131) sets up a set of guidelines for training raters of writing assessment. In the first step, the leader (or preferably a team) should read through the scripts to find anchor/benchmark scripts that exemplify the different bands/levels on the rating scale. The scripts that exemplify certain problematic situations should be included. After that, the first set of scripts is generally given to the raters in order (e.g. from highest to lowest) with the appropriate scores indicated. Nonetheless, the purpose of this activity is to familiarise the raters with the scale and to illustrate certain features of the rating criteria. When the raters are comfortable with the scale, a set of scripts, including one script at each level in random order, should then be given. Finally, raters should work with more problematic sets of scripts, which may have more than one script at a given level, or, may be less clearly representative of certain points of the scale. Furthermore, Weigle recognises that it is important to note that getting a large group of raters to agree on exact scores is virtually impossible, and some disagreement among raters is expected. Thus, it is crucial to inform the raters that they are not required to be perfectly accurate all the time. However, the raters who consistently rate lower or higher than the rest of the group should be given feedback and perhaps be retrained.

However, Alderson, Clapham and Wall (1995) have a rather different view of how to conduct rater training or "standardisation meetings". While Weigle (2002) suggests that the consensus scripts should be given with the scores indicated, Alderson, et al. (1995) state that the raters should not be shown the decisions made by the committee "to prevent examiners from being influenced by the original committee's reasoning before they have had a chance to try out the scale and think for themselves" (p. 112). The consensus scripts are those scripts that represent "adequate" and "inadequate" performances, as well as scripts that present some common problems, which raters often face, but are rarely described in rating scales. The raters should try out the rating scale on the consensus scripts, which are given before the meeting. The first stage of the meeting should be devoted to discussing the consensus scripts to find out if all raters agree on the marks that have been given, and to work out why they have had problems, if they do not agree. The aim of this activity is to help all raters match the marks of the original committee. Thus, the committee's consensus scores should not be indicated on the scripts. After that, the problematic scripts should be presented, together with guidelines on what raters should do in these cases. Then, further practice in marking should be provided with another set of scripts. Similar to Alderson, et al. (1995), McNamara (2000) states that the rater rating system is a process, which involves individual raters independently marking a series of different levels of performance. Then, in groups, they have to share their marks with other raters. The differences are noted and discussed in detail by referring to the interpretation of the different levels of descriptors of the individual raters. The purpose of the meeting is to try to bring about a general agreement on the relevant descriptors and the rating categories.

Conclusion

In a real world context where language learners need to use language skills they have acquired, it is crucial to require the learners to produce the language for the examiners to assess. Since these assessors are required to make more complicated judgements in assessing the learners' performances, they need to use a rating scale for the rating of the students' performances, in order to arrive at a unified system of assessment. To make reliable judgements, the scale developers need to choose the type of rating scales (analytic or holistic) and designing approach (*a priori* or intuitive) to be adopted in order to appropriately fit the context. It should be noted that in an EFL context, particularly in Thailand, in order to promote students' learning, analytic rating scales should be encouraged, although they are time consuming and more expensive to use. Similarly, the rating scales should be empirically developed for the EFL classes, as the information from the designing process provides more useful information about the learners and this information can also be used in rater training sessions. Also, the raters need to be properly and routinely trained to use the scales reliably, though a regular rater training may be time consuming, which could be expensive to run and does require collaboration among the staff members.

The Author

Bordin Chinda is a lecturer at the English Division, Department of Western Languages, Faculty of Humanities, Chiang Mai University. His research interests include impact/wash-back studies, performance-based assessment, innovation in language education, professional development, and English for Academic Purposes. Bordin holds a PhD in Language Testing and Assessment from the University of Nottingham, UK.

References

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71-86). London: Macmillan.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks, CA: Corwin Press.

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

Chinda, B. (2009). *Professional development in language testing and assessment: A case study of supporting change in assessment practice in in-service EFL teachers in Thailand*. Unpublished PhD Thesis, The University of Nottingham, UK.

Chinda, B. (2012). *A Study of Change in Attitude, Belief and Knowledge in Language Testing and Assessment: A Case Study of Teachers at the English Division, Chiang Mai University.* Unpublished manuscript, Chiang Mai University, Chiang Mai, Thailand.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing* (Vol. 7). Cambridge: Cambridge University Press.

Douglas, D. (2001). Language for specific purposes assessment criteria: Where do they come from? *Language Testing, 18*(2), 171-185.

Goulden, N. R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *The Journal of Research and Development in Education, 27*(1), 73-82.

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic context* (pp. 241-278). Norwood, NJ: Ablex Publishing Corporation.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching: Ideologies and alternative. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (Vol. Part I, pp. 487-504). New York: Springer.

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000 - writing: composition, community, and assessment*. Princeton, NJ: Educational Testing Service.

Hudson, T. (2005). Trends in assessment scales and criterion-referenced language assessment. *Annual Review of Applied Linguistics, 55*(1), 205-227.

Jacoby, S., & McNamara, T. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213-241.

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

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.

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

Lynch, B. K. (2001). The ethical potential of alternative language assessment. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language Testing in honour of Alan Davies* (Vol. 11, pp. 228-239). Cambridge: Cambridge University Press.

McNamara, T. (1996). *Measuring second language performance*. London: Addison Wesley Longman Ltd.

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

Myford, C. M. & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part 2. *Journal of applied measurement, 5*(2), 189-227.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawai'i.

Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: Cambridge University Press.

Turner, C. (2001). The need for impact studies of L2 performance testing and rating: identifying areas of potential consequences at all levels of the testing cycle. In C. Elder, A. Brown, E. Grove, K. Hill, N. Iwashita, T. Lumley, T. McNamara & K. O'Loughlin (Eds.), *Experimenting with uncertainty: Language Testing in honour of Alan Davies* (Vol. 11, pp. 138-149). Cambridge: Cambridge University Press.

Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *The Canadian Modern Language Review/La Revue cannadienne des langues vivantes, 56*(4), 555-584.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70.

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

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

Wigglesworth, G. (2008). Task and performance based assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7, pp. 111-122). New York: Springer Science+Business Media.