Direct assessment of second language writing: Holistic and analytic scoring

Ling He, PhD

linghe@westcliff.edu

Abstract

Direct assessment of second language (L2) writing skill, in which a student is usually assessed by means of an essay on a topic during a limited time period, has been used as a valid measurement of ability to write in comparison to indirect assessment of writing through multiple-choices. Although both direct and indirect assessments have a risk of reliability, it is effectively argued that direct assessment is more representative of the integrative aspects of writing (Coffman,1966; Morris-Friehe & Leuenberger, 1992). In this era of globalization, a large number of culturally and linguistically diverse students seek higher education in North America. Direct assessment of L2 writing skill is widely used to assess these newly-arrived students' English proficiency for placement decisions at the beginning of their university programs in the target language. As such, understanding direct assessment of L2 writing is paramount to improve assessment procedures for test validity and fairness. This article reviews direct assessment of L2 writing in the existent research literature with an attention to its two major scoring rubrics: holistic and analytic. The literature review concludes that the purpose of the writing task is significant in deciding which scoring method to use in specific contexts.

Introduction

Direct assessment of second language (L2) writing skill, in which a student is usually assessed by means of an essay on a topic during a limited time period, has been using as a valid measurement of ability to write in comparison to indirect assessment of writing ability through multiple-choices. Although both direct and indirect assessments have a risk of reliability, it is effectively argued that direct assessment is more representative of the integrative aspects of writing (e.g., Coffman, 1966; Morris-Friehe & Leuenberger, 1992) because "it has a face validity since it requires the candidate to perform the actual behavior which is being measured" (Eley, 1955, p. 11). In contrast, indirect measure using multiple-choice assessment of writing ability is less laudable because it does not "require the examinee to perform the actual behavior being measured —he does not actually write… [and he] makes little or no attempt to measure the 'larger organization, and content'" (Beaddock, Lloyd-Jonese, & Shoer, 1963, p. 42). In this era of globalization, a large number of culturally and linguistically diverse students seek higher education in North America. Direct assessment of L2 writing skill is widely used to assess these newly arrived students' English proficiency for placement decisions at the beginning of their university programs. As such, understanding direct assessment of L2 writing is paramount to improve assessment procedures for test validity and test fairness. This article reviews direct assessment of L2 writing skill in the existent research literature with an attention to its two major scoring rubrics: holistic and analytic. The literature review concludes that the purpose of the writing task is significant in deciding which scoring method to use in specific contexts and that a valid, reliable rubric can enhance direct assessment of L2 writing for either placement decisions or diagnostic purposes, thereby effective teaching and learning in L2 writing.

Discussions

Direct assessment of L2 writing skill for English proficiency

Direct assessment of L2 writing skill has been acting as a gate-keeper in most North American universities to decide nonnative English speaking (NNES) students' English language proficiency levels, whose score is often used for placement at the beginning of program studies. That is, in addition to the required large-scale international standardized test of English language proficiency such as TOEFL (Test English as a Foreign Language) or IELTS (International

English Language Testing System) for admission to a university in an English-speaking country, NNES students are usually required to attend a writing test when they start their program in the target language country. Those students having a low test score of writing must take English courses in a bridge English program before having regular English courses. Thus, direct assessment of L2 writing is a high-stakes test as its placement decision instantly influences NNES students' time, tuition, living expense, academic plan, and motivation for their program study. To avoid raters' subjective bias and a risk of reliability, selecting a valid, reliable method of scoring written texts is vital in assessing L2 writing ability.

Numerous research investigations have demonstrated that direct assessment of writing performance tends to yield low reliabilities or a poor consistency of raters' grading on a writing score. Among many factors influencing a rating score of direct assessment of writing skill, reliability and validity are two major concerns (Hamp-Lyons & Kroll, 1996; Henning, 1991); namely, a test cannot be valid without being reliable. Scoring methods are emphasized as one of the significant factors that can affect direct assessment scores of writing ability McNamara, 1996; Brown, 1996). An effective rating rubric is the heart of the validity of direst assessment of writing because the rubric "represents, implicitly or explicitly, the theoretical basis upon which [a] test is founded" (Wiggle, 2002, p. 109), and it operationally defines the construct of being measured (McNamara, 1996). Choosing a right scoring method becomes the first decision for direct assessment of L2 writing to reduce unsystematic grading that potentially threatens scoring validity. Thus, it is paramount to minimize measurement errors for the attainment of reliability and validity in direct assessment of L2 writing through a well-developed, effective rating scale with explicitly defined criteria and standards.

Types of scoring methods in direct assessment of L2 writing skill

While many other rating scales exist such as primary trait scales (Lloyd-Jones, 1977; Weigle, 2002) and multiple-trait scales (Hamp-Lyons, 1990; Hamp-Lyons & Henning, 1991), holistic and analytic rating rubrics have been mainly used for direct assessment of L2 writing skill in test situations (Canale, 1981; Carroll, 1980; Perkins, 1983). Accordingly, these two scoring methods are discussed in the following, respectively.

Holistic scoring. As impressionistic marking, holistic scoring aims to rate overall

proficiency level by assigning a single score to each written text based on raters' immediate and general impression of the examinees' final written products using a rating scale, often a five- or six-point continuum, which uses a set of scoring criteria where each point corresponds to a descriptor that defines good performance at each score point. See the holistic scoring for iBT TOEFL Test Independent Writing Rubrics by Educational Testing Services (Appendix 1). Concerns from the literature are mainly about the validity of the procedure of using holistic scoring due to low reliability among the raters' scores. For example, Diederich (1964) conducted the earliest study about holistic scoring in the large-scale tests wherein 53 raters measured 300 essays and yielded low reliability or a big difference among raters' rating scores of writing. Similarly, Breland and Jones (1984) showed the same low raters' reliability of 800 essays. Other concerns arising from the literature criticize that using holistic scoring for direct assessment of L2 writing provides little useful diagnostic information about a test-taker's writing ability (Elbow, 1996), language accuracy, control of syntax, lexical range, and organization (Davies, Brown, Elder, Hill, Lumley, & McNamara, 1999), the inadequate inter-rater reliability check for measures of linguistic accuracy (Hamp-Lyons, 1990; Henning, 1991; Polio, 1997; Raimes, 1990), and "problematic for second-language writers, since different aspects of writing ability develop at different rates for different writers" (Weigle, 2002, p. 114); that is, the same holistic score assigned to two different texts may represent two entirely different sets of characteristics. The central concern is that these drawbacks of holistic scoring may cause raters to confound L2 writing skill with language proficiency (Cohen, 1994).

Research shows various sources causing low reliability of using holistic scoring in direct assessment of L2 writing. These sources include scoring methods (Hamp-Lyons, 1991; Reid, 1993; Shohamy, 1995), rating behavior (Kim, 2010), native and nonnative raters (Shi, 2001), sampling bias (Newell, 1984), writing tasks (Shaw & Weir, 2007), and scoring methods and raters' own intuitive impression (Lumely, 2002). Among these identified sources, two primary sources are rater inconsistency and sampling bias. Rater training is considered the important way for the attainment of rating consistency or reliability (e.g., Bachman & Palmer,1996; Brown 1995); however, sample bias like using more than one writing sample in a large-scale writing assessment seems to have not feasibly improved given practicality.

Although the low reliability of rating is shown in the literature, holistic scoring has been used as the primary measure of writing skill (e.g., Cohen, 1994; Shaw & Weir, 2007) for its distinct advantages of practicality and diagnostic information. Holistic scoring is practical as the essays can be measured rapidly; thus, the process is more economical than analytic scoring. The practicality of holistic scoring echoes Bachman and Palmer (1996)'s criteria of assessment scales that "the most important consideration in designing and developing a language test...is its usefulness" (p. 17). Weigle (2002) further analyzes usefulness in this regard and uses practicality as the most important criterion while comparing the scoring methods. That is, a valid, reliable scoring method must be first of use. In terms of diagnostic purposes, research in both L1 (English as a first language) and L2 writing studies have a consensus that holistic scoring is reliable in giving useful ranking information in an efficient way with a low cost when rater training and rating session administration are faithfully adhered to (Perkins, 1983; White, 1994). Because of these two distinct strengths, holistic scoring is commonly used in large-scale assessments of writing performance, especially in high-stakes tests for making decisions about placement for L2 writers (e.g., Cumming, 1990; Hamp-Lyons, 1990; Reid, 1993).

Analytic scoring. Different from holistic scales, analytic scoring typically provides separate or component scores of writing on specific features, such as relevance and adequacy of content, organization, and lexical breadth and depth, thereby having higher discriminating power (e.g., Mendelsohn & Cumming, 1987). See Analytic Scoring Rubric for Writing (Appendix 2), which was originally developed by scholars in Virginia in the 1990s and was adapted by Wright (2015). The multiple ratings of different components of L2 writing in an analytic rubric are awarded to the same essay in an attempt to enhance the reliability of assessment (e.g., Hout, 1996; Shaw & Weir, 2007; Weir, 1990). As such, analytic scoring is preferred over holistic scoring by many writing specialists for explicit diagnostic information about NNES students' writing, which helps determine proficiency levels for placement and assist NNED students as well. As Shaw and Weir (2007) state,

> Analytic scales are more suitable for second-language writing as different features of writing develop at different rates. This method, therefore, lends itself more readily to full

profile reporting and could well perform a certain diagnostic role in delineating students' respective strengths and weaknesses in overall written production. (pp. 151-152). That is, analytic scoring is popular for measures of specific textual features which NNES writers may have developed unevenly. For example, some NNES writers may have an excellent control of sentence structure and grammar but lack knowledge in organizing their ideas in the manner expected in the target language. Given separate or component scores of writing features, an analytic rubric is easier in training raters for its practical and efficient rating procedure (Cohen 1994; McNamara, 1996). Compared with holistic scoring, analytical scoring has a high inter-rater and intra-rater reliability.

Analytic scoring is mainly disadvantageous for its time-consuming and costly in large-scale writing assessment. It is sometimes challenging to assign numerical scores based on certain descriptors even for experienced essay raters (Hamp-Lyons, 1989). Also, measuring the quality of individual aspects may maximize the role of autonomous text features and diminish the inter-language correlation of written discourse (Hillock, 1995; Hughes, 2003; White, 1994). Thus, analysis scoring alone cannot always easily accommodate qualitative judgments concerning content, coherence, style, and language.

Conclusion

Research has demonstrated that direct assessment of L2 writing skill using holistic and analytic scoring methods are reliable and valid to inform test users (e.g., an educational institution) of NNES students' proficiency levels. While both scoring methods have their pros and cons, holistic scoring, which uses a single score representing a reader's general overall assessment of a written text, has been used as the primary measure of L2 writing due to its usefulness or practicality to differentiate NNES students by their relative ranking on a continuum across a range of scores. In contrast, analytic scoring, which specifies separate scores for specific features of writing, is welcomed for its diagnostic information (Brown & Hudson, 2002) for classroom evaluations of learning and a call for student attention to areas of needed improvement or their achievement (Brown,1996).

The purpose of this article aims to discuss two major scoring rubrics for direct assessment of L2 writing: holistic and analytic. The review of these two rubrics helps improve

the validity of assessment procedures in assessing L2 writing and raise teachers' awareness of different features of scoring rubrics. It is arbitrary to assume that analytical rubrics are better for assessing individual components of the various features of a written text than holistic rubrics assigning an overall score to a piece of writing (Weigle, 2002; Hyland, 2002). The purpose of the writing task is significant in deciding which scoring method to use. A well-developed, effective scoring rubric with explicitly defined criteria, standards, and scales should be encouraged for scoring L2 writing to avoid subjectivity and a risk of reliability, thereby ultimately enhancing effective teaching and learning in L2 writing. What writing components should be assessed in a scoring rubric would depend upon L2 writing construct according to a sound writing theory.

References

Braddock, R., Lloyd-Jones, R., & Schoer, L. (1963). *Research in written composition*. Champaign, Illinois: National Council of Teachers of English

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

Breland, H. M., & Jones, R. J. (1984). Perceptions of writing skills. *Written Communication*, *1*, 101-109

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Carroll, B. J. (1980). *Testing communicative performance.* London: Pergamon Institute of English.

Canale, M. (1981). Communication: How to evaluate it? *Bulletin of the Canadian Association of Applied Linguistics, 3*, 77-94.

Coffman, W. E. (1966). On the Validity of Essay Tests of Achievement. *Journal of Educational Measurement 3*, 151-156.

Cohen, A. (1994). *Assessing language ability in the classroom*. Boston, MA: Heinle and Heinle.

Conner, U. (1996). *Contrastive rhetoric: Cross-cultural aspects of second-language writing*. New York: Cambridge University Press.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, *7* (1), 31-51.

Eley, E. G. (1955). Should the general composition test be continued? The test satisfied an educational need. *College Board Review, 25*, 10-13

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

Diederich, P. B. (1964). Problems and possibilities of research in the teaching of written composition. In D. H. Russell, M. J. Early, & E. J. Farrell (Eds.), *Research Design and the Teaching of English: Proceedings of the San Francisco Conference 1963* (pp. 52-73). National Council of Teachers of English, Champaign, IL

Elbow, P. (1996). Writing assessment: Do it better, do it less. In W. Lutz, E. White, & S. Kamusikiri (Eds.), *The Politics and Practices of Assessment in Writing* (pp. 120-34). Modern Language Association of America.

Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H.W. Dechert & Raupauch(Eds.), *Interlingual processes* (pp. 229-244). Tubingen: Gunter Narr.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge: Cambridge University Press.

Hamp-Lyons, L. & Henning, G. (1991). Communicative writing profiles: Investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, *41* (3), 337-373.

Hamp-Lyons, L. & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL*, *6* (1), 52-72.

Henning, G. (1991). Issues in evaluating and maintaining an ESL writing assessment program. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 279-292). Norwood, NJ: Ablex Publishing Corp.

Hillocks, G. (1995). T*eaching writing as reflective practice*. New York: Teachers College Press.

Hughes, A. (2003). *Testing for language teachers.* Cambridge: Cambridge University Press.

Hout, B. (1996). Toward a new theory of writing assessment. *College Composition and Communication, 47*, 549-566.

Hyland, K. (2002a). *Teaching and researching writing*. London: Longman.

Kim, H. J. (2010). Investigating raters' development of rating ability on a second language speaking assessment. *Unpublished doctoral dissertation.* Teachers College, Columbia University.

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

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

Mendelsohn, D. and Cumming, A. (1987). Professors' ratings of language use and rhetorical organizations in ESL compositions. *TESL Canada Journal*, *5*(1), 9-26.

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

Morris, M. and Leuenberger, J. (1992). Direct and indirect measures of writing for nonlearning disabled and learning disabled college students. *Reading and Writing, 4* (3), 281-296
Newell, G.E. (1984). Learning from writing in two content areas: A case study/protocol analysis. *Research in the Teaching of English*, *18*(3), 265-287

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17*, 651-671.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47* (1), 101-143.

Reid, J. (1993). *Teaching ESL Writing*. Englewood Cliffs, NJ: Regents Prentice Hall.
Raimes, A. (1990). The TOEFL test of written English: Causes for concern. *TESOL Quarterly, 25,* 407-430.

Shohamy, E. (1995). Performance assessment in language testing. *Annual Review of Applied Linguistics*, *15*, 188-211.

Shaw, S. D., & Weir, C. J. (2007). Studies in Language Testing: Examining writing. *Research and practice in assessing second language writing* (No. 26). Cambridge, UK: Cambridge University Press.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303-325.

Weir, C. J. (1994). *Understanding and developing language tests*. London: Prentice Hall.
Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance*. San Francisco: Jossey-Bass Publishers.

Wright, W. E. (2015). *Foundations for Teaching English Language Learners: Research, Theory, Policy, and Practice* (2nd ed.). Philadelphia, PA: Caslon Publishing.

Appendix 1 Holistic Scoring

iBT TOEFL Test: Independent Writing Rubric (Scoring Standards)

Score	Task Description
5	**An essay at this level largely accomplishes all of the following:** • effectively addresses the topic and task • is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details • displays unity, progression, and coherence • displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
4	**An essay at this level largely accomplishes all of the following:** • addresses the topic and task well, though some points may not be fully elaborated • is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details • displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections • displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning
3	**An essay at this level is marked by one or more of the following:** • addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • displays unity, progression, and coherence, though connection of ideas may be occasionally obscured • may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • may display accurate but limited range of syntactic structures and vocabulary
2	**An essay at this level may reveal one or more of the following weaknesses:** • limited development in response to the topic and task • inadequate organization or connection of ideas • inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • a noticeably inappropriate choice of words or word forms • an accumulation of errors in sentence structure and/or usage
1	**An essay at this level is seriously flawed by one or more of the following weaknesses:** • serious disorganization or underdevelopment • little or no detail, or irrelevant specifics, or questionable responsiveness to the task • serious and frequent errors in sentence structure or usage
0	**An essay at this level** merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Appendix 2 Analytic Scoring

Analytic Scoring Rubric for Writing

Domain Score	**1**	**2**	**3**	**4**
Composing	Lack of a central idea; no details, random digressions	Lack of a focused central idea, or more than one idea; limited details and many digressions	Central idea but with fewer details and some digressions	Central idea with relevant details in a well-organized text
Style	**1**	**2**	**3**	**4**
	Limited vocabulary; choppy sentences; flat tone	Basic vocabulary; limited to no sentence variety; inconsistent tone	Acceptable vocabulary choices; some sentence variety; consistent but less appealing tone	Well-chosen vocabulary; excellent sentence variety; tone that appeals to readers
Sentence Formation	**1**	**2**	**3**	**4**
	Frequent non- standard word order; mostly run- on sentences or sentence fragments; omissions of many words; errors frequently detract from meaning.	Some non-standard word order; several run-on sentences; several sentence fragments; omissions of several words; errors somewhat detract from meaning	Mostly standard word order, some run-on sentences; some sentence fragments; occasional omission of words; errors do not detract from meaning	Standard word order; no run-on sentences; no sentence fragments; effective transitions
Usage	**1**	**2**	**3**	**4**
	Little to no correct use of inflections; frequent tense shifts; little to no subject-verb agreement; many errors in word meaning; errors fully detract from meaning	Some correct use of inflections; some consistency in tense and subject-verb agreement; several errors in word meaning; errors somewhat detract from meaning	Mostly correct use of inflections; Mostly consistent tense and subject-verb agreement; mostly standard word meaning; errors do not detract from meaning	Correct use of inflection (e.g., verb conjugations, plurals, prefixes, suffixes, adverbs); consistent tense; consistent subject-verb agreement; standard word meaning
Mechanics	**1**	**2**	**3**	**4**
	Little to no correct use of mechanics or formatting; errors fully detract from meaning	Some correct use of mechanics and formatting; errors somewhat detract from meaning	Mostly correct use of mechanics and formatting; errors do not detract from meaning	Correct use of mechanics (capitalization, punctuation, spelling), and formatting

COMMENTS	TOTAL SCORE
*4 = consistent control; 3 = nearly consistent control; 2 = inconsistent control; 1 = little or no control *Source:* Adapted from O'Malley & Pierce in 1996, originally from Virginia Department of Education in 1990s	

Source: Wright, W. E. (2015). *Foundations for Teaching English Language Learners: Research, Theory, Policy, and Practice* (2nd ed.). Philadelphia, PA: Caslon Publishing. *Note:* The rubric is formatted with some features by the author of this article.