Chapter 19
Writing Assessment: A Pedagogical Perspective on Variations and Dichotomies

Farah Bahrouni
Sultan Qaboos University, Oman

ABSTRACT

This chapter reports on the findings of a contextual mixed-method study investigating the factors that influence how teachers at the Language Center (LC) of Sultan Qaboos University, Oman (SQU), assess students’ academic writing and the features on which they focus. Results from the quantitative data analysis indicate that the influence of raters’ first language (L1) is statistically significant, while the impact of their experience of teaching a particular course is effectively insignificant. Findings from the qualitative data analysis, however, reveal other influential factors, notably teaching experience in general terms, learning experience, educational background, culture, and personality. As for the features teachers focus on while assessing writing, both quantitative and qualitative data analyses show different patterns for the L1 groups involved. Findings from this study could be useful for informing decision makers on the various ways teachers from different L1 and experiential backgrounds assess writing and the constructs they follow.

INTRODUCTION

Over the years, the writing skill, whether in the learner’s first, second, or even foreign language, has been assessed mainly through two approaches, direct and indirect assessment (Breland, 1983).

Indirect assessment is so called because the test takers’ writing ability is inferred through observations of specific discrete bits of knowledge about writing, usually by means of multiple-choice questions. This approach draws heavily on structural linguistics, particularly contrastive analysis and behaviorist psychology (Shimada, 1997). A basic assumption in the structural approach is that knowledge of the elements of a language is equivalent to knowledge of the language (McNamara, 2000; Shimada, 1997) hence ‘atomization’ (McNamara, 2000) of language complexities and its ‘decontextualization’ (Resnick, in Wolcott & Legg, 1998, p. 18) into discrete fragments. This is a fundamental tenet that controls the construction of the test and its method. This results in the criticism that such a
test is devised to measure only learners’ knowledge about language, ‘linguistic competence,’ instead of their language use, ‘linguistic performance,’ with questions revolving mainly around discrete grammar points (McNamara, 2000). Furthermore, indirect tests are criticized for relying heavily on multiple-choice questions, which suffer from a serious lack of emphasis on high order skills and a defective transmission view of learning, in which knowledge is delivered and memorized, not constructed and used, (Myers, 1994 cited in Wolcott & Legg, 1998, p. 18). Put succinctly, what language test takers have traditionally been asked to produce differs radically different from what is normally produced by humans in daily life.

Despite their questionable construct validity, indirect tests were more advantageous in terms of easy and objective scoring, a virtue that direct writing assessment has been yearning to acquire since it took over with the advent of the communicative approach in the early 1970s. While it has gained in construct validity, direct writing assessment has lost in reliability of the scores and validity of the entailed decisions. Actually, it has lost where its indirect counterpart has gained. With direct writing assessment, the challenge has moved to developing rating procedures that do not jeopardize construct validity. In other words, the focus now is on the learners’ linguistic performance, not on their linguistic competence, as was the case with indirect assessment. Ever since this focus shift, direct assessment has become the most widely used method to assess writing around the world. As early as 1974, Diederich captured the rationale behind this change:

As a test of writing ability, no test is as convincing to teachers of English, to teachers in other departments, to prospective employers, and to the public as actual samples of each student’s writing, especially if the writing is done under test conditions in which one can be sure that each sample is the student’s own unaided work. People who uphold the view that essays are the only valid test of writing ability are fond of using the analogy that, whenever we want to find out whether young people can swim, we have them jump into a pool and swim (in Du, 1995, p. 3).

Direct writing testing, however, is not the panacea for the perennial problems writing assessment is suffering from. It has its inherent problems, among which is the great deal of subjectivity involved in its assessment. The very fact that direct writing assessment involves the judgment of a human rater or raters makes it susceptible to subjectivity, error, variability, and unpredictability. Even with the clearest possible detailed scoring instructions and the most efficient rater training, an element of subjectivity always remains in the judgments of the raters involved (Lumley & McNamara, 1995; Weigle, 1994, 1998, 2002).

Aside from the writer and the assignment variables, the rater variable is a major source of measurement error in direct writing assessment. This source becomes more conspicuous, and more influential in terms of its effects on the scores, when the writing performance is produced in the context of a second language (Carrell, 1995) or a foreign language (Haas & Flower, 1988), where raters’ multicultural backgrounds as well as test takers’ L1 transfer and rhetoric, further aggravate the situation. Such is the case with the context of the present study: the composition writers are Arabic-speaking EFL learners, while their raters come from different corners of the world, obviously with different linguistic and experiential backgrounds. In such a context, rater reliability emerges as an issue of serious concern. Raters are human, and humans are different in terms of background, culture, and personality. Discrepant rating behavior in a writing assessment context is, therefore, congruent with human nature. What is actually unnatural, and may raise questions about the reliability of the ratings, is raters’ agreement. Natural and quite uncontrollable disagreement is nonetheless undesirable as it seriously affects