ARTICLES

Problems in Assessing EFL Writing on High-stakes Tests: A Guide to the Research

Bernard SUSSER
Department of International Studies,
Faculty of Liberal Arts

Abstract

The research on writing assessment shows that testing writing is a difficult task. This paper reviews this literature, focusing on validity, reliability, and scoring methods. Problems concerning validity include theoretical issues such as the definition of the construct "writing" as well as practical issues such as task authenticity. The main problem concerning reliability is whether or not holistic scoring, the most common scoring method for high-stakes assessments, gives the score user a reliable estimate of the test taker's writing ability; this problem becomes more serious with ESL/EFL writers whose writing and language abilities may differ markedly. A related issue is the training of the raters who produce the scores. The conclusion, while emphasizing that the present method of assessing writing in high-stakes examinations is producing dubious and unreliable results, also presents some lessons from the research to help EFL writing instructors prepare their students for these important tests.

1. Introduction

Essay questions are included in most high-stakes tests for native speakers (e.g., SAT) and ESL/EFL students (e.g., TOEFL iBT and IELTS) but a review of the research on writing assessment shows that there are a number of problems. Referring specifically to "large-scale, formal L2 writing assessments," Leki, Cumming, and Silva (2008, pp. 91-92) found four "issues and types of analyses": (1) reliability of scoring; (2) rater training; (3) task types; and (4) washback. The present paper examines the research literature on these issues, focusing specifically on validity, reliability, and scoring methods. Two similar reviews of research have appeared previously: Grabe and Kaplan's (1996, pp. 399-414) chapter is a good review but is now more than 10 years old; Leki, Cumming, and Silva's coverage is recent but brief (2008, pp. 87-92). This paper has two main purposes: first, it will serve as a guide to the literature on this topic; for that reason multiple references are given when several researchers have discussed a specific point. This guide is intended to encourage EFL writing instructors to contribute their own research on writing assessment. Second, in the conclusion, I draw some lessons from the research to help EFL writing instructors prepare their students for these important tests.

2. Validity and reliability in general

Many writers have discussed the problems of validity and reliability in assessing both first- and second-language writing (e.g., Davies & Elder, 2005; Hamp-Lyons, 1990; Henning, 1991; Williamson, 1993). Here I will take up just two important points: (1) critiques of the concepts themselves; and (2) the gap in the understanding of validity between statisticians and writing teachers.

(1) In an extreme case, Lynne (2004) rejects the entire practice of basing writing assessment on the objectivist principles of validity and reliability, claiming that they conflict with the social constructivist principles that govern contemporary writing theory: "continued reliance on these terms [validity and reliability] perpetuates an intolerable incongruity between the ideals of literacy education and the practice of writing assessment" (p. 13; see also Broad, 2003, pp. 5 ff.; Wilson, 2006, pp. 49 ff.). Scharton (1996) offers a critique of validity from a political aspect. While few other writers go so far, many have expressed concerns over "the limitations of validity theory" (Stoynoff and Chapelle, 2005, pp. 148 ff.; see also Hamp-Lyons, 2007, p. 501).

(2) A related problem, described by Huot (2002, pp. 45 ff., 93-94, 156 ff.) is a significant gap in the understanding and use of validity between "the college writing assessment community" and the "educational measurement community"; his point is that college English teachers continue to hold a simpleminded understanding of validity as being sure that the test measures what it is supposed to measure, while educational measurement specialists have developed quite a complex definition of validity, emphasizing among other things "the decisions and the consequences of those decisions made on behalf of an assessment" (p. 57). Elliot (2005, pp. 266 ff.) offers support for this view by showing how the concept of validity has changed, with the present emphasis being on the uses of tests. Further, Shale (1996) makes a similar point in his criticism of the use of "the classical measurement approach to 'reliability'" (p. 93) with respect to inter-rater reliability, arguing that "efforts at standardizing marker behavior are conceptually ill founded" (p. 94). The dominance of "classical measurement practice and theory" in writing assessment can be explained by the fact that these procedures are easy to use but also because writing teachers "have assumed an unequal partnership with an epistemically privileged class of researchers" who have become the "determiners of truth and knowledge" vis-à-vis inferior composition teachers "who are viewed as technicians" (p. 95). These critiques suggest that writing teachers and program administrators might need to rethink how they make use of the scores from high-stakes tests to evaluate and place students.

3. Problems with validity

Typical high-stakes tests require an essay to be written within a strict time limit (30 minutes to an hour is typical) in response to a given prompt. There are several problems with the validity of such a task, both theoretical and practical.

3.1. Theoretical problems: The construct of "writing"

Four theoretical issues are of particular importance. (1) First is the difficulty of defining the construct of writing itself, which in turn affects the construct validity of the test

(see e.g., Cumming, 2002, p. 79; Hamp-Lyons, 1990, p. 80; Weigle, 2002, pp. 41 ff., 78 ff.). Breland, Bridgeman, & Fowles (1999), following the Flower and Hayes model of writing, state that the "cognitive writing processes consist of planning, text generation, and revision" (p. 3) but given the time constraints of the "usual writing assessment situation," two of these three (planning and revision) are not usually assessed. However, in their discussion of construct validity, they cite Messick's point that one of the "two greatest threats to construct validity" is "*construct underrepresentation*," and give as an example "the use of extremely constrained time limits, allowing no time for planning or revision" (p. 4; see also Camp, 1993, p. 61). They seem to be casting doubt on the 30-minute essay's construct validity, a serious charge. In fact, Brindley and Ross (2001) use as their example of the problem of content representation (validity) "the traditional timed essay," which is quite different from what students "would normally produce in the classroom" (p. 152). From a different point of view, Norton (2000) analyzed marking memoranda from three different high-stakes writing assessments and found that the three "have different assumptions about competent writers and readers" (p. 25). These statements cast doubt on the claim that such essay tests assess "writing" in a meaningful way.

(2) A related problem is that "the common practice of a single sample of a student's writing is insufficient for a valid assessment" (Grabe & Kaplan, 1996, p. 414). The issues here are the decontextualized setting of the test, the fact that individual testees will have more or less interest in any given topic, that different topics will draw on some writing skills and not others, etc. In short, as Hamp-Lyons & Kroll (1997, p. 32) pointed out, "the use of single-sample context-stripped prompts with 30 minutes to write responses no longer has validity in the eyes of the ESL composition community" (see also Hamp-Lyons & Kroll, 1996/2001, p. 226) a sentiment shared by their colleagues in mainstream L1 composition (Camp, 1993, p. 52). It is for this reason that there is a trend towards assessment of portfolios, although this obviously is not a practical alternative for high-stakes admissions tests.

(3) A third theoretical issue is the difficulty of separating language ability from writing ability in the case of ESL/EFL testees. Cumming (1989) argued that "writing expertise and second-language proficiency each make quite different contributions to the processes and products of writing in a second language" (p. 118). (See Barkaoui, 2007a, for a thorough review of the research on ESL/EFL essay tests, and Kroll, 1998, and Hedgcock, 2005, pp. 606-609, for reviews of research on ESL writing assessment.)

(4) Finally, a fourth issue is that the test rubrics come to standardize the writing, leading to a construct validity problem because the assessment cannot produce scores that support valid inferences about writing achievement (Nichols and Berliner, 2005, pp. 95-97); instead, "our tests can end up measuring not the construct of *writing achievement*, but the construct of *compliance to the rubric*" (p. 97).

3.2. Practical problem 1:
Task authenticity (construct validity)

Numerous writing experts have argued that a short essay on a general topic simply is not authentic. In fact, Educational Testing Service (ETS) publications make the spe-

cific claim that "the writing tasks presented in TWE [Test of Written English] topics have been identified by research as typical of those required for college and university course work" (ETS, 2004, p. 6). However, as Weigle (2006, pp. 224-225) pointed out, "…research on writing in undergraduate courses… suggests that graded writing is virtually always done in response to other texts that have been read and/or discussed orally. Thus the task of writing an essay on a previously unseen topic, with little or no opportunity to explore the topic through interaction with other texts on the topic, is a highly inauthentic task as it does not represent the contextual factors of authentic academic writing"; this view is supported by many other writing experts (e.g., Bailey, 1998, p. 186; Braine, 1989; Green, 2007, pp. 52, 216-217; Horowitz, 1986; 1989, p. 33; Leki & Carson, 1997, p. 49; Rosenfeld, Leung, & Oltman, 2001, p. 49; Scharton, 1996, p. 71; Shih, 1986 p. 621; Weigle, 2002, p. 52; Zhu, 2004). (The research on this topic is surveyed in: Cooper & Bikowski, 2007, pp. 207-210; Paltridge, 2004, pp. 87-99; Reid, 2001, pp. 147-151; and Waters, 1996, pp. 9-22.) Roemer (2002, p. 16) is almost alone in evaluating the TWE writing task as authentic.

The claim that this task is not authentic is based in part on research investigating actual writing assignments and tests given in college and university courses. Horowitz (1986) found that college writing assignments other than essay test questions largely asked the writer to "find, organize, and present data according to fairly explicit instructions" (p. 455). He also found that essay tests too were "based on a body of knowledge to which all the examinees had equal access in the recent past" (1991, p. 81). The Hale, et al. (1996) study prepared for the ETS studied a much larger sample of authentic university writing tasks (162 tasks from eight North American universities). They found that in-class short writing tasks "typically consisted of questions on tests" (p. 32; see also p. 46), and that the most common out-of-class assignments were short tasks and essays (p. 31; see also p. 41). They argued that essays and library research papers were similar except that the latter "called for seeking out sources of material to be incorporated and/or cited in the paper" (p. 16), implying that "essays" did not require "sources" (the same implication appears again on p. 41; only once did the authors admit that an essay may require reading something [p. 47]). Although the authors claimed (pp. 42-43) that their results were basically similar to those of Horowitz, in fact they studiously avoided throughout their report Horowitz's main finding, namely that most university essay assignments are based on a specific "body of knowledge" quite different from the "content-free writing assessments such as the TWE" (1991, p. 75). Kroll (1991, p. 22) uses the term "expository writing that is non-content based" and Hamp-Lyons calls the prompts "anodyne" (1996, p. 231). Further, of the 34 example writing assignments given in Hale, et al. (1996, pp. 52-61), only two do not require specific reading or background knowledge. More recently, Moore and Morton (2005; see also 2007) studied 155 writing tasks from two Australian universities and found that "almost all tasks involved a research component of some kind, requiring the use of either *primary* or *secondary* sources or a combination of the two" (p. 52). They concluded that the kind of writing demanded in actual university courses was quite different from the essay question on the IELTS (pp. 63-64). In a related study, Coffin (2004) was surprised

to find that the argument structures used by successful candidates in the IELTS essays used an "approach to argumentation…more reminiscent of letters to the press than of academic prose" (p. 243).

Another aspect of the lack of authenticity is that no information is given "about the audience, purpose, etc., to help test-takers contextualize their essay" (Chalhoub-Deville & Turner, 2000, p. 534). Shaw and Falvey (2008) note that one of the "minimum requirements for task instructions in a direct test of writing" should be "a specification of the target audience and purpose of writing" (p. 178; see also Shaw and Weir, 2007, pp. 71 ff. and the example in Bachman & Palmer, 1996, pp. 274-275).

Of course, the designers of such writing tests are not unaware of these issues (e.g., Cumming, et al., 2004, p. 136; Cumming, et al., 2006, p. 8). For example, Cumming, et al. (2000) recognize that undergraduate writing often consists of "telling people about the knowledge one has" (p. 5) so the focus is on transmitting rather than creating knowledge (p. 5). The problem is that the test must use "tasks that represent key genres for writing which [sic] are integral to as wide a range of university or college contexts as possible, without biasing this selection in favor of (or against) particular groups, areas of interest or knowledge, or specific situations" (p. 5). In other words, the prompt cannot require specific knowledge but must be "academic" so they redefine the common academic task of displaying one's knowledge of a topic in an essay to a display of "writing and language abilities" (p. 5), in other words, writing devoid of content.

3.3. Practical problem 2: Question types

The next point is the questions themselves. It goes without saying that to be fair to all test takers, the questions cannot require knowledge of any specific content, nor should the topic be culturally biased (Carlson & Bridgeman, 1986, p. 139; see also Hamp-Lyons, 1996, p. 231; Kroll, 1991, pp. 22 ff.; Kroll & Reid, 1994, pp. 235 ff., gave several examples of culturally problematic essay prompts); this unfortunately brings us back to the problem of "content-free writing assessments," as discussed above. As a result, the questions ask testees to draw on their personal experience. However, White (1986, p. 67) points out, unfortunately without any references, that "there is a surprisingly low correlation between scores on personal experience and on analytic topics"; if true, this may defeat the purpose of the test. In fact, Tedick (1990) found that a field-specific topic produced "a marked increase in holistic scores" compared to a general topic (p. 132); she argued that field-specific topics discriminate levels of writing proficiency better than general topics (pp. 132 ff.) Further, as He and Shi (2008) argued, general topics "encourage memorization of sample essays and result in little or surface learning" (p. 143). They reported further that most of the Chinese students they interviewed felt that they had passed the TWE because of test preparation training that required them to memorize "generic sentences" or whole model essays (p. 157); the interviewees claimed that this training "did not help them develop writing skills. However, it did help them pass TWE the first time they took it" (p. 137). Lee, Breland, and Muraki (2004) noted "that some examinees can somehow compensate for their low ELA [English language ability] by using a strategy

of memorizing a template of an exemplary essay and replacing some key words in the essay for a new writing prompt" (p. 22); they singled out "examinees of the East Asian language group" as likely suspects.

3.4. Practical problem 3: Task message (face validity)

Many experts on writing instruction have pointed out that questions asking about personal opinions give the wrong message about what good academic writing is: "it is permissible, and even encouraged, to express a strong opinion on a topic that one has not read or thought much about and for which one has no ready access to data that will support one's point of view. This message can easily undermine a teacher's insistence on critical reading and the use of appropriate sources as essential components of academic writing skills" (Weigle, 2006, p. 226; see also Hillocks, 2002, pp. 77, 201 ff.; Murphy, 2007, pp. 54-55; Walker & Piu, 2008, pp. 18-19; Weigle, 2002, pp. 95, 146-147). Further, most academic writing teachers employ the process approach: "writing as a process of discovering meaning, writing from sources, or writing as revision" (Weigle, 2006, p. 222; see also Susser, 1994; Wolcott & Legg, 1998, pp. 12-18). Consequently, "they feel undermined when an externally mandated timed impromptu essay examination gives a different message: Good writing is a good first draft" (Weigle, 2006, p. 222).

Condon (2006, pp. 212 ff.) gives a particularly damming critique of the short, timed essay used for placement purposes: he found that when his colleagues scored "a set of timed writings for placement and then again for critical thinking, the resulting scores actually show a negative correlation. In effect, if students choose to think, their placement scores suffer..." (p. 214). This happens because the type of reading required for assessing this type of placement essay imposes, as Charney (1984) claimed, "a very unnatural reading environment, one which intentionally disallows thoughtful responses to the essays" (p. 74; Huot, 1990, p. 211 and 2002, pp. 145 ff. makes a similar point [see below]).

3.5. Summary: The validity of essay tests

This section has surveyed the research on writing assessment using timed essay questions. There is no consensus among the experts but a few points are clear. First, despite its basis in the science of statistics, the concept of validity is a contentious one when applied to writing tests and claims about validity in this case should be taken as provisional (if not doubtful). Specifically, there are serious problems with the authenticity and message of the assessment task, so that consumers of test scores should exercise caution in their use. In the same way, writing instructors should make clear to their students the difference between test preparation and typical academic writing.

4. Scoring and reliability

A variety of scoring methods are used in high-stakes writing assessments; here I will focus on holistic scoring, which is used for the TOEFL iBT independent writing task. This essay "is scored on the overall quality of the writing: development, organization, and appropriate and precise use of grammar and vocabulary" (ETS, 2008, p. 26); the essays are scored holistically by "certified raters" on a scale of 0 to 5 using the "Independent Writing Rubrics" (ibid., p. 46). Several interrelated problems relating to the way the essays are

scored and the reliability of such scores have been addressed in the literature; here I will take up four: (1) issues of holistic scoring, including its relationship to analytic or multitrait scoring, and to essay length and content; (2) rater effects and rater training; (3) holistic scoring and EFL writers; and (4) issues of reliability.

4.1. Holistic scoring

4.1.1 Holistic assessment and holistic scoring

It is important first to clarify the distinction between holistic writing assessment and holistic scoring. According to Hamp-Lyons (1992, ¶ Holistic Writing Assessment), holistic writing assessment refers to tests that "test writing wholly through the production of writing" and, citing Cooper (1977, p. 4), do not require counting of "linguistic, rhetorical, or informational features" of the writing. She contrasts holistic writing assessments to objective and analytic tests; the former use recognition rather than production skills and the latter involve counting features such as the number of words but do not consider discourse-level aspects of writing quality. Holistic writing assessment covers several scoring methods, including holistic, primary trait, and multiple trait scoring (Hamp-Lyons, 1992, ¶ Scoring methods for holistic writing assessment; see also Hamp-Lyons, 1991, pp. 243 ff.). Cooper (1977, pp. 4 ff.) lists seven types of holistic evaluation, including "general impression marking" (pp. 11-12), which corresponds most closely to the "holistic scoring" discussed below.

4.1.2. Pros and cons of holistic scoring

Charney (1984, p. 74) defines holistic rating as "a quick, impressionistic qualitative procedure for sorting or ranking samples of writing …according to previously established criteria." Many writing assessment experts defend holistic scoring; Cooper (1977, p. 3), for example, wrote that "holistic evaluation of writing remains the most valid and direct means of rank-ordering students by writing ability" (see Wolcott & Legg, 1998, pp. 71-87 for a review of the research on holistic scoring). Evans, Pearson, and Bundrick (1999) claimed that holistic scoring correlates with other test results. The "principle virtue" of holistic scoring is its "reliance on the complex, richly informed judgments of skilled human raters to interpret the quality of students' writing performance" (Cumming, Kantor, & Powers, 2002, p. 68). Connor and Carrell (1993) showed that both writers and raters using a TWE prompt and scoring guide shared the same assumptions about both the purpose and the evaluation of the task, which they saw as confirming the value of holistic assessment (p. 156); on the other hand, their raters "did not think it important for the writers to address the specific requirements of the prompt" (p. 153). White (1986, pp. 68 ff.) argued that holistic scoring is good only if a number of pitfalls are avoided, such as weaknesses in the community of readers and problems with the scoring guide (rubric); he objected strongly to the "use of an all-purpose scoring guide, designed to meet the requirements of all questions that are designed for a particular testing program" (p. 71), claiming that it is impossible to use the same scoring guide for different questions. In the revised edition of his book on writing assessment (1994, pp. 231 ff., 281 ff.), White continued to favor holistic scoring as "the triumph of the human" (p. 281) but devoted a whole section to "problems with holistic scoring" (pp. 283-289); Williamson (1993) also defended holistic scoring with a very judicious examination of

its problems. Huot (1993) found that "holistic scoring procedures actually promote the kind of rating process that insures a valid reading and rating of student writing" (p. 227).

On the other hand, quite a few writing experts are unhappy with holistic scoring (e.g., Elbow, 1993/1996a, pp. 200 ff.; Elbow, 1996b, with an appendix listing 19 "works that question holistic scoring," pp. 132-133; Haswell, 1998, pp. 237 ff.; Lynne, 2004, pp. 32-37; Murphy, 1999, pp. 116 ff.; Shaw & Falvey, 2008, pp. 28-29, 37; Vaughan, 1991; Wilson, 2006, p. 23). Weigle (2006, pp. 224-225) complained about the "reductive nature of holistic scoring"; Haswell (2006, pp. 72 ff. & n. 6 p. 248) cited research showing that holistic scores explain very little and add almost zero information for placement decisions. Cooper (1977), cited above as a strong advocate of holistic assessment, argued that holistic evaluation can be reliable when the raters have similar backgrounds and are properly trained but also only when "we have at least two pieces of a student's writing" (p. 19) and two independent ratings; further, "there are theoretical reasons to believe that the writing task we set for the students should specify a speaker role, audience, and purpose" (p. 20). Charney (1984) pointed out that in holistic scoring, judgment is usually influenced by salient but superficial characteristics such as length, handwriting, spelling, and mature vocabulary (pp. 76, 78; see also Weigle, 2002, pp. 69-70). Wilson (2006) argued that the rubrics used in holistic scoring are reductive, "don't honor the complexity of what we [teachers] see in writing" (p. 41), and ignore the rhetorical purpose of writing (p. 76). Huot (1990) argued that "perhaps the most important criticism of holistic scoring is the possibility that a personal stake in reading might be reduced to a set of negotiated principles, and then a true rating of writing quality could be sacrificed for a reliable one" (p. 211). Finally, Camp's (1993) comment effectively summarizes the position that the "single impromptu writing sample…no longer seems a strong basis for validity" (p. 52). She continues: "Performance on the writing sample no longer appears to be an adequate representation of the accepted theoretical construct for writing, nor does it seem an adequate representation of students' likely experiences with writing, past or future…" (p. 52).

4.1.3. Holistic scoring and essay length

One way to clarify this issue of holistic scoring is to look at the issues of length and content. Criticisms of the SAT writing test focused on claims that essay length was the most important factor in the evaluation and that content, even erroneous or nonsensical content, was ignored. Perelman (2005) argued that "longer essays consistently score higher" and that the test disregards factual accuracy and basically encourages the wrong kind of writing. In response, Kobrin, Deng, and Shaw (2007), researchers for the College Board, responded that length explains only 39% of the variance of essay scores (p. 10). A similar problem has appeared with respect to the TOEFL essay. Even ETS researchers have noted the strong relationship between essay length and holistic scores (Frase, et al., 1999, p. 24; Lee, Gentile, & Kantor, 2008, pp. 35-36). Jarvis, Grant, Dikowski, & Ferris (2003) found that "text length … appears to be a rather consistent predictor of perceived writing quality" (p. 400), adding that writers can compensate for deficiencies in some areas by just writing more (p. 399). Schaefer (2008, p. 472) cited several studies showing that essay length is one predictor of scores in a variety

of situations; see also Carrell (1995, pp. 182, 185-186). Reid (1990, pp. 195-196) cited several studies showing that essay length correlates highly with writing quality for both native and nonnative speakers.

4.1.4. Holistic scoring and essay content

Concerning content, ETS researchers Cumming, Kantor, and Powers (2002, pp. 72 ff.) found that experienced essay raters looked for qualities such as rhetorical organization, coherence, accuracy and fluency of language, quantity, etc.; content was conspicuously absent; Erdosy (2004), another ETS researcher, found the emphasis on content "unbelievable, not to mention depressing" (p. 10). However, concerning writing assessment in general, research on what affects rater decisions about writing quality usually has found content to be at or near the top (Huot, 1990, p. 207; Sakyi, 2000, p. 140). Weigle (2002, p. 132) claimed that TOEFL users "are interested primarily in a general sense of a person's ability to create a coherent written text, not the quality of the ideas or the persuasiveness of the essay." ETS researchers Lee and Kantor (2005, p. 3) seem to confirm this claim when they argued that raters of the TOEFL independent writing task "mostly focus on language and ideas developed by the writer," specifically contrasting this to the need to "attend to content accuracy" when rating the integrated writing question. Concerning persuasiveness, "the development of a reasonable argument," Connor (1991, p. 222) pointed out that the TWE scoring guidelines do not mention this explicitly. The results of her small-scale study suggest that the TWE guidelines "may not reflect the kind of 'communicative competence' that previous research and the raters in this study consider important in argumentative/persuasive writing" (p. 222). These contradictory claims suggest that TOEFL writing scores are not evaluating what the profession considers to be important in writing.

4.2. Rater effects and rater training

There has been extensive research on how raters assess essays written for examinations (reviewed in Leki, Cumming, & Silva, 2008, pp. 91-92; Lumley, 2005, pp. 23 ff.). The main "rater effects" have been drawn from this research by Knoch, Read, and von Randow (2007, p. 27): (1) severity effect (raters are too harsh or too lenient); (2) halo effect (rating on the basis of an overall impression without discriminating among distinct categories); (3) central tendency effect (avoiding extreme ratings); (4) inconsistency; and (5) bias effect (see also Schaefer, 2008, on rater bias patterns, the review of research in Eckes, 2008, pp. 155 ff. and his theory of "rater types," and the review of research on rater characteristics, preferences, etc. in Shaw & Weir, 2007, pp. 168 ff.). Carrell (1995) found that grades assigned by raters using a modified version of the TWE rubrics varied significantly by the rater's personality type (p. 175).

It is well known that untrained raters give unreliable results and that training can be effective in improving reliability (e.g., Weigle, 1994; see surveys of the literature in Elder, Barkhuizen, Knoch, & von Randow, 2007, pp. 37 ff. and in Shaw & Weir, 2007, pp. 181 ff.). Powers and Kubota (1998, p. 6) provided a description of what ETS rater training consists of; Hamp-Lyons noted that it is "very draconian" (2003, p. 182). However, many studies of rater training and the actual scoring process itself have shown numerous problems, even with trained raters (e.g., Belanoff, 1991, p. 59; Charney, 1984; Connor-Linton,

1995; Hamp-Lyons, 2003, p. 178-179; Huot, 1990; Knoch, Read, & von Randow, 2007, p. 27; Lumley, 2002; 2005, pp. 239 ff.; Sakyi, 2000; Shaw & Falvey, 2008, p. 15; Vaughn, 1991; Weigle, 2002, pp. 70 ff.; Wolcott & Legg, 1998, 60-70); Carrell (1995), however, found "no statistically significant effects for raters' training" (p. 175). Specifically, Cumming, Kantor, and Powers (2001) found some differences between native and non-native graders (e.g., p. 56) and a tendency for raters to look at rhetoric/organization in higher-scoring essays and grammar in lower-scoring ones (e.g., pp. 59-60). Erdosy (2004) pointed out that raters pay most attention to grammatical competence at the lowest level of proficiency, sociolinguistic competence in the middle levels, and discourse competence at the top level (p. 8) and that "any composition will receive a higher score if preceded by weak compositions than if preceded by strong ones" (p. 7); Sakyi (2000, pp. 144-145) also found that some raters' judgments were influenced by a contrast with the essay they had read previously. He and Shi (2008, pp. 141-142) showed the "devastating" effect of interrater unreliability on test takers in a high-stakes university writing test. Hamp-Lyons and Zhang (2001), using a TOEFL-like prompt, found that even trained native-speaker raters were unable to discount their disapproval of the ideology expressed by Chinese examinees. Finally, Huot (2002, pp. 145 ff.) argued that rater training limits the ways readers read student writing, producing "an environment for reading that is unlike any in which most of us ever read" (p. 146), resulting in the production of "reliable" numerical scores "regardless of the decisions" (p. 147) that need to be made. He noted that rater training is often called a calibration process (p. 145); Herrington & Moran (2006, p. 126) go further to claim that raters have been "normed" and "made, arguably, into something like reading machines" (see also Haswell, 2006, pp. 72ff & n. 6 p. 248; Huot, 1996, pp. 236-237; Wilson, 2006, p. 77). However, Lumley's (2005) study suggested quite the opposite, that rating is "conflict" (pp. 240 ff.), "a social procedure organized around the need to bring *intuitive reactions* into conformity with the requirements of the testing institution" (p. 240); his point was that raters do not so much mechanically calibrate essays against the given scale as go through a process of "squeezing, shaping, defining, arbitrating, comparing, and rejecting" to express their "instinctive feeling" about an essay's quality in terms of the official rating instrument (ibid.; see also pp. 289 ff.). He gave numerous examples of how raters struggled with the limitations of the official rubric.

4.3. Holistic scoring for ESL writers

Holistic scoring is especially problematic for ESL writers because of "the mix of strengths and weaknesses often found in ESL writings" (Hamp-Lyons & Kroll, 1997, p. 29; see also Cumming, et al., 2005, pp. 6-7; Hamp-Lyons, 1991, pp, 253 ff.; Hamp-Lyons, 1992, ¶ Multiple trait scoring and LLEP writers; Hamp-Lyons, 1995; Hamp-Lyons, 1996, pp. 232 ff.; Hamp-Lyons, 2003, p. 176; Weigle, 2002, p. 114). Cumming (1990) found that raters tend to distinguish students' language proficiency from their writing expertise so that "students who are poor writers may be disadvantaged even if their language skills are good" (p. 42); likewise, Carlson and Bridgeman (1986, p. 144) point out that unlike native speakers, it is often the case that ESL writers show a "greater disparity between organizational skills and mechanical

competence," so raters using a holistic score "must agree on how to score essays that present a large discrepancy between organization and mechanical skill" (p. 144). They noted further (pp. 143-144) that Freedman (1979) found that "content and organization had the greatest influence on holistic scores" for essays by native speakers but cited Breland and Jones's (1982) finding that this may not hold true for ESL students; in their case, grammar and vocabulary "were particularly strong correlates of holistic scores" (p. 143; see also Hamp-Lyons & Kroll, 1996/2001, pp. 233-234). Tedick and Mathison (1995) found problems with holistic scoring of ESL essays, including cases in which writers who did not address the task nevertheless received high holistic scores (pp. 222 ff.); they urge those "involved in ESL writing assessment to move beyond the limits that holistic scoring places on us" (p. 225). Kroll (1990) demonstrated this in a large-scale study, finding that "the writers were able to show control over the level of either syntax or rhetoric while simultaneously showing poor control at the other level" (p. 150); this disparity was masked by the holistic score. Hughes (1989, p. 91) noted that holistic scoring rubrics similar to the TOEFL's "assume that a particular level of grammatical ability will always be associated with a particular level of lexical ability," an assumption that he finds "highly questionable."

Hamp-Lyons (e.g., 1995), among others, has urged the use of analytic or multiple-trait scoring to solve this problem. In a small-scale study, Barkaoui (2007b) found several interesting differences between ratings based on holistic and on multiple-trait scales. Lee, Gentile, and Kantor (2008, pp. 1 ff.) discussed the merits and demerits of analytic vs. holistic scoring; their own research showed that analytic scores correlate reasonably well with holistic measures (p. 34). One explanation for this may be Haswell's claim that analytic evaluation based on several aspects of writing, such as the well-known "Profile" of Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey (1981, p. 30), "is identical to holistic rating. The 'Profile' just asks the rater to perform the holistic five times" (2005/2007, p. 5). White (1994, p. 233) opposed analytic scoring because (he claimed with no reference) that there is no agreement "about what, if any, separable subskills exist in writing."

4.4. Score reliability

The final issue is that of the reliability of the scores produced by holistic grading. Cherry and Meyer (1993) give a thorough critique of reliability issues in holistic assessment; they are particularly critical of the common practice of resolving cases of discrepant scores by having the essay read by a third reader (pp. 121 ff.). ETS researchers are well aware that "writing assessments based on single essays, even those read and scored twice, have extremely low reliability—usually less than .60" (Breland, Bridgeman, & Fowles, 1999, p. 14: see also Elliot, 2005, pp. 344-345, for a survey of the dismal history of inter-rater correlations at the College Board).

5. Conclusion

The above analysis of the research on assessing writing by EFL learners on high-stakes tests leads to two important conclusions. First, it is clear that there is much disagreement among the experts and conflicting research results. This can be explained, at least in part, by recalling certain conditions inherent in this topic: the amorphous nature of writing as a construct; the compromises

necessitated by actual testing conditions, including financial and physical limitations, the great variety of students being tested, the balance between fairness and the need for depth, etc.; and the competing interests of the various stakeholders. The second conclusion is that the present method of assessing writing in high-stakes examinations is producing dubious and unreliable results that do not reflect well the testees' ability in academic writing and therefore are a poor standard upon which to evaluate learners for admission or placement. Unfortunately, no practical solution to this problem has yet been proposed. In this situation, EFL writing instructors must not only give their students sufficient practice in writing essays under test conditions but also help students to do the kind of writing required in coursework.

Despite these grim conclusions, the literature reviewed in this paper does provide some help to ESL/EFL teachers who are preparing students to take high-stakes writing tests. The first lesson concerns task authenticity: we saw above that most researchers distinguish the typical test writing task from the tasks students usually are assigned in content courses. This suggests that test preparation courses should teach the test prompt essay as a separate writing genre, with its own set of conditions, structure, voice, etc. This may help to avoid confusion on the part of students and teachers. The second lesson is similar: the research on question types and task message summarized above suggests implicitly writing exercises and classroom activities that will help students do well on these tests. This is an area where many test preparation books can be helpful (see, e.g., Lucas, et al., 2009, pp. 30-31). The final lesson is in the literature on holistic scoring. Students do not need to be made aware of the many problems with this type of scoring but familiarity with the rubrics used and study of model essays will help them write essays that will get high scores. Here again, some textbooks have good exercises to develop these skills or teachers can develop their own (e.g., Susser, 2008, p. 3). It is my hope that the above review of research will encourage EFL/ESL writing teachers to conduct their own research on these issues, and consider new ways to help their students prepare for high stakes writing assessments.

References

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

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions*. Boston, MA: Heinle & Heinle.

Barkaoui, K. (2007a). Participants, texts, and processes in ESL/EFL essay tests: A narrative review of the literature. *The Canadian Modern Language Review, 64*(1), 99-134.

Barkaoui, K. (2007b). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107.

Belanoff, P. (1991). The myths of assessment. *Journal of Basic Writing, 10*(1), 54-66.

Braine, G. (1989). Writing in science and technology: An analysis of assignments from ten undergraduate courses. *English for Specific Purposes, 8*(1), 3-15.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). *Writing assessment in admission to higher education: Review and framework*. College Board Report #99-3. New York: College Entrance Examination Board. Retrieved January 27, 2010 from http://www.ets.org/Media/Research/pdf/RR-99-03-Breland.pdf

Breland, H. M., & Jones, R. J. (1982). *Perceptions of writing skills*. ETS Research Report #82-47.

Princeton, NJ: Educational Testing Service.

Briggs, D. C. (2001). The effect of admissions test preparation: Evidence from NELS:88. *Chance, 14*(1), 10-18. Retrieved August 25, 2008 from http://www.amstat.org/PUBLICATIONS/chance/pdfs/141.briggs.pdf

Brindley, G., & Ross, S. (2001). EAP assessment: Issues, models, and outcomes. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 148-166). Cambridge, UK: Cambridge University Press.

Broad, B. (2003). *What we really value: Beyond rubrics in teaching and assessing writing*. Logan, UT: Utah State University Press.

Camp, R. (1993). Changing the model for the direct assessment of writing. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 45-78). Cresskill, NJ: Hampton Press.

Carlson, S., & Bridgeman, B. (1986). Testing ESL student writers. In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 126-152). White Plains, NY: Longman.

Carrell, P. L. (1995). The effect of writers' personalities and raters' personalities on the holistic evaluation of writing. *Assessing Writing, 2*(2), 153-190.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System, 28*(4), 523-539.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65-81.

Cherry, R. D., & Meyer, P. R. (1993). Reliability issues in holistic assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 109-141). Cresskill, NJ: Hampton Press.

Coffin, C. (2004). Arguing about how the world is or how the world should be: The role of argument in IELTS tests. *Journal of English for Academic Purposes, 3*(3), 229-246.

Condon, W. (2006). Why less is not more: What we lose by letting a computer score writing samples. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 211-220). Logan, UT: Utah State University Press.

Connor, U. (1991). Linguistic/rhetorical measures for evaluating ESL writing. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 215-225). Norwood, NJ: Ablex.

Connor, U., & Carrell, P. (1993). The interpretation of tasks by writers and readers in holistically rated direct assessment of writing. In J. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 141-160). Boston, MA: Heinle & Heinle.

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762-765.

Cooper, A., & Bikowski, D. (2007). Writing at the graduate level: What tasks do professors actually require? *Journal of English for Academic Purposes, 6*(3), 206-221.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 3-31). Urbana, IL: NCTE.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning, 39*(1), 81-141.

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

Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing, 8*(2), 73-83.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*(1), 107-145.

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL prototype writing tasks: An investigation into raters' decision making and development of a preliminary*

analytic framework. TOEFL Monograph Series, MS-22. Princeton, NJ: Educational Testing Service. Retrieved January 27, 2010 from http://www.ets.org/Media/Research/pdf/RM-01-04.pdf

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper*. TOEFL Monograph Series, MS-18. Princeton, NJ: Educational Testing Service. Retrieved January 27, 2010 from http://www.ets.org/Media/Research/pdf/RM-00-05.pdf

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5-43.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL*. TOEFL Monograph Series, MS-30. Princeton, NJ: Educational Testing Service. Retrieved January 27, 2010 from http://www.ets.org/Media/Research/pdf/RR-05-13.pdf

Davies, A., & Elder, C. (2005). Validity and validation in language testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 795-813). Mahwah, NJ: Lawrence Erlbaum Associates.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155- 185.

Educational Testing Service. (2004). *Test of written English: Guide*. Fifth edition. Princeton, NJ: Educational Testing Service. Retrieved January 27, 2010 from http://www.ets.org/Media/Tests/TOEFL/pdf/tweguid.pdf

Educational Testing Service. (2007). *TOEFL iBT tips: How to prepare for the TOEFL iBT*. Princeton, NJ: Educational Testing Service. Retrieved January 27, 2010 from http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_Tips.pdf

Elbow, P. (1996a). Ranking, evaluating, and liking: Sorting out three forms of judgment. In B. Leeds (Ed.), *Writing in a second language: Insights from first and second language teaching and research* (pp. 200-215). New York: Longman. (Reprinted from *College English, 55*(3), 1993, 187-206).

Elbow, P. (1996b). Writing assessment: Do it better, do it less. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 120-134). New York: Modern Language Association of America.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37-64.

Elliot, N. (2005). *On a scale: A social history of writing assessment in America*. New York: Peter Lang.

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions*. TOEFL Research Report #70. Princeton, NJ: Educational Testing Service. Retrieved January 27, 2010 from http://www.ets.org/Media/Research/pdf/RR-03-17.pdf

Evans, R., Pearson, L. C., & Bundrick, M. (1999). Reaffirming construct, convergent and predictive validity between objective tests and holistically scored essays. *College Student Journal, 33*(1), 2-6.

Frase, L. T., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL Test of Written English*. TOEFL Research Report #64. Princeton, NJ: Educational Testing Service. Retrieved January 27, 2010 from http://www.ets.org/Media/Research/pdf/RR-98-42.pdf

Freedman, S. W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*(3), 328-338.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistic perspec-*

tive. New York: Addison-Wesley Longman.

Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge, UK: Cambridge University Press.

Hale, G.A, Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs*. TOEFL Research Report #54. Princeton, NJ: Educational Testing Service.

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

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp- Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241- 276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1992). Holistic writing assessment for LEP students. *Proceedings of the Second National Research Symposium on Limited English Proficient Student Issues: Focus on Evaluation and Measurement*. United States Department of Education, Office of Bilingual Education and Minority Language Affairs. Washington, DC. Volume 2. Retrieved March 24, 2009 from http://www.ncela.gwu.edu/pubs/symposia/second/vol2/holistic.htm

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

Hamp-Lyons, L. (1996). The challenges of second-language writing assessment. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 226-240). New York: Modern Language Association of America.

Hamp-Lyons, L. (2003). Writing teachers as assessors of writing. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162-189). Cambridge, UK: Cambridge University Press.

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

Hamp-Lyons, L., & Kroll, B. (2001). Issues in ESL writing assessment: An overview. In T. Silva & P. K. Matsuda (Eds.), *Landmark essays on ESL writing* (pp. 225-240). Mahwah, NJ: Lawrence Erlbaum Associates. (Reprinted from *College ESL, 6*(1), 1996, 52-72.)

Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000--Writing: Composition, community and assessment*. TOEFL Monograph Series, MS-5. Princeton, NJ: Educational Testing Service. Retrieved January 28, 2010 from http://www.ets.org/Media/Research/pdf/RM-96-05.pdf

Hamp-Lyons, L., & Zhang, B. W. (2001). World Englishes: Issues in and from academic writing assessment. In J. Flowerdew & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 101-116). Cambridge, UK: Cambridge University Press.

Haswell, R. H. (1998). Rubrics, prototypes, and exemplars: Categorization theory and systems of writing placement. *Assessing Writing, 5*(2), 231-268.

Haswell, R. H. (2006). Automatons and automated scoring: Drudges, black boxes, and dei ex machina. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 57-78). Logan, UT: Utah State University Press.

Haswell, R. H. (2007). Researching teacher evaluation of second language writing via prototype theory. In P. K. Matsuda & T. Silva (Eds.), *Second language writing research: Perspectives on the process of knowledge construction* (pp. 105-120). Mahwah, NJ: Lawrence Erlbaum Associates, 2005. Revised 2007 version retrieved January 7, 2010 from http://www.writing.ucsb.edu/wrconf08/Pdf_Articles/Haswell-Article.pdf

Haswell, R., & Wyche-Smith, S. (2009). Adventuring into writing assessment . In B. Huot & P. O'Neill (Eds.), *Assessing writing: A critical sourcebook* (pp. 203-217). Boston, MA: Bedford/St. Martin's. (Reprinted from *College Composition and Communication, 45*(2), 1994, 220-236.)

Hedgcock, J. S. (2005). Taking stock of research and pedagogy in L2 writing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 597-613). Mahwah, NJ: Lawrence Erlbaum Associates.

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. 279-291). Norwood, NJ: Ablex.

Herrington, A., & Moran, C. (2006). WritePlacer *Plus* in place: An exploratory case study. In P. F. Ericsson & R. H. Haswell (Eds.), *Machine scoring of student essays: Truth and consequences* (pp. 114-129). Logan, UT: Utah State University Press.

Hillocks, G. Jr. (2002). *The testing trap: How state writing assessments control learning*. New York: Teachers College Press.

Horowitz, D. M. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20*(3), 445-462.

Horowitz, D. (1989). Function and form in essay examination prompts. *RELC Journal, 20*(2), 23-35.

Horowitz, D. (1991). ESL writing assessments: Contradictions and resolutions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 71-85). Norwood, NJ: Ablex.

Hughes, A. (1989). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41*(2), 201-213.

Huot, B. A. (1993). The influence of holistic scoring procedures on reading and rating student essays. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 206-236). Cresskill, NJ: Hampton Press.

Huot, B. (1996). Computers and assessment: Understanding two technologies. *Computers and Composition, 13*(2), 231-243.

Huot, B. (2002). *(Re)articulating writing assessment for teaching and learning*. Logan, UT: Utah State University Press.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*(4), 377-403.

Johns, A. M. (2001). Interpreting an English competency examination: The frustrations of an ESL science student. In T. Silva & P. K. Matsuda (Eds.), *Landmark essays on ESL writing* (pp. 117-135). Mahwah, NJ: Lawrence Erlbaum Associates. (Reprinted from *Written Communication, 8*(3), 1991, 379-401.)

Knoch, U., Read, J., & von Randow, J. (2007). Retraining writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26-43.

Kobrin, J. L., Deng, H., & Shaw, E. J. (2007). Does quantity equal quality? The relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology, 8*(1), 1-15.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 140-154). New York: Cambridge University Press.

Kroll, B. (1991). Understanding TOEFL's Test of Written English. *RELC Journal, 22*(1), 20-33.

Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219-239.

Kroll, B., & Reid, J. (1994). Guidelines for designing writing prompts: Clarifications, caveats, and cautions. *Journal of Second Language Writing, 3*(3), 231-255.

Lee, Y.-W., Breland, H., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts for different native language groups*. TOEFL Research Report #77. Princeton, NJ: Educational Testing Service. Retrieved January 28, 2010 from http://www.ets.org/Media/Research/pdf/

RR-04-24.pdf

Lee, Y.-W., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and E-rater*. TOEFL Research Report #81. Princeton, NJ: Educational Testing Service. Retrieved January 28, 2010 from http://www.ets.org/Media/Research/pdf/RR-08-01.pdf

Lee, Y.-W., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Evaluating prototype tasks and alternative rating schemes*. TOEFL Monograph Series, MS-31. Princeton, NJ: Educational Testing Service. Retrieved January 28, 2010 from http://www.ets.org/Media/Research/pdf/RR-05-14.pdf

Leki, I., & Carson, J. (1997). "Completely Different Worlds": EAP and the writing experiences of ESL students in university courses. *TESOL Quarterly, 31*(1), 39-69.

Leki, I., Cumming, A., & Silva, T. (2008). *A synthesis of research on second language writing in English*. New York: Routledge.

Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing, 11*(1), 5-21.

He, L., & Shi, L. (2008). ESL students' perceptions and experiences of standardized English writing tests. *Assessing Writing, 13*(2), 130-149.

Lucas, M. D., Carty, P., Christmas-Nishibata, J., & Susser, B. (2009). Issues in teaching to the writing test: Preparing students for the TOEFL iBT independent writing task. *Bulletin of the Institute for Interdisciplinary Studies of Culture, 26*, 23-32.

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.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Frankfurt am Main: Peter Lang.

Lynne, P. (2004). *Coming to terms: A theory of writing assessment*. Logan, UT: Utah State University Press.

Moore, T., & Morton, J. (2005). Dimensions of difference: A comparison of university writing and IELTS writing. *Journal of English for Academic Purposes, 4*(1), 43-66.

Moore, T., & Morton, J. (2007). Authenticity in the IELTS academic module writing test: A comparative study of Task 2 items and university assignments. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (pp. 197-248). Cambridge, UK: Cambridge University Press.

Murphy, S. (1999). Assessing portfolios. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: The role of teachers' knowledge about text, learning, and culture* (pp. 114-135). Urbana, IL: NCTE.

Murphy, S. (2007). Some consequences of writing assessment. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 33-50). New York: Routledge.

Nichols, S. L., & Berliner, D. C. (2005). *The inevitable corruption of indicators and educators through high-stakes testing*. Education Policy Research Unit, Arizona State University. Retrieved January 11, 2010 from http://epicpolicy.org/files/EPSL-0503-101-EPRU.pdf

Norton, B. (2000). Writing assessment: Language, meaning, and marking memoranda. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 20-29). New York: Cambridge University Press.

Paltridge, B. (2004). Academic writing. *Language Teaching, 37*(2), 87-105.

Perelman, L. (2005). New SAT: Write long, badly and prosper. *Los Angeles Times*, May 29,2005. Retrieved January 28, 2010 from http://articles.latimes.com/2005/may/29/opinion/oe-perelman29

Powers, D. E., & Kubota, M. Y. (1998). *Qualifying readers for the online scoring network: Scoring argument essays*. Research Report, RR-98-28. Princeton, NJ: Educational Testing Service. Retrieved January 28, 2010 from http://www.ets.org/Media/Research/pdf/RR-98-28.pdf

Raimes, A. (1992). The author responds to Traugott, Dunkel, and Carrell. *TESOL Quarterly, 24*(1), 186-190.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 191-210). New York: Cambridge University Press.

Reid, J. (2001). Advanced EAP writing and curriculum design: What do we need to know? In T. Silva & P. K. Matsuda (Eds.), *On second language writing* (pp. 143-160). Mahwah, NJ: Lawrence Erlbaum Associates.

Roemer, A. (2002). A more valid alternative to TOEFL? *College and University (C & U Journal), 77*(4), 13-17.

Rosenfeld, M., Leung, S., & Oltman, P. K. (2001). *The reading, writing, speaking, and listening tasks important for academic success at the undergraduate and graduate levels*. TOEFL Monograph Series, MS-21. Princeton, NJ: Educational Testing Service. Retrieved January 28, 2010 from http://www.ets.org/Media/Research/pdf/RM-01-03.pdf

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 129-152). NewYork: Cambridge University Press.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Scharton, M. (1996). The politics of validity. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 53-75). New York: Modern Language Association of America.

Shale, D. (1996). Essay reliability: Form and meaning. In E. M. White, W. D. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Politics, policies, practices* (pp. 76-96). New York: Modern Language Association of America.

Shaw, S., & Falvey, P. (2008). *The IELTS writing assessment revision project: Toward a revised rating scale*. University of Cambridge ESOL Examinatiion Research reports #01. Cambridge, UK. Retrieved December 23, 2009 from http://www.cambridgeesol.org/assets/pdf/research_reports_01.pdf

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

Shih, M. (1986). Content-based approaches to teaching academic writing. *TESOL Quarterly, 20*(4), 617-648.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Longman.

Stoynoff, S., & Chapelle, C. A. (2005). *ESOL tests and testing*. Alexandria, VA: TESOL.

Susser, B. (1994). Process approaches in ESL/EFL writing instruction. *Journal of Second Language Writing, 3*(1), 31-47.

Susser, B. (2008). Teaching to the test with technology. Paper presented at WorldCALL 2008, Fukuoka, Japan. Retrieved March 27, 2010 from http://www.j-let.org/~wcf/proceedings/d-021.pdf

Tedick, D. J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes, 9*(2), 123-143.

Tedick, D. J., & Mathison, M. A. (1995). Holistic scoring in ESL writing assessment: What does an analysis of rhetorical features reveal? In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 205-230). Norwood, NJ: Ablex.

Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111-125). Norwood, NJ: Ablex.

Walker, R., & Riu, C. P. (2008). Coherence in the assessment of writing skills. *ELT Journal, 62*(1), 18-28.

Waters, A. (1996*). A review of research into needs in English for Academic Purposes of relevance to the North American higher education context*. TOEFL Monograph Series, MS-6. Princeton, NJ: Educational Testing Service. Retrieved January 28, 2010 from http://www.ets.org/Media/Research/pdf/RM-96-07.pdf

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*(2), 197-223.

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

Weigle, S. C. (2006). Investing in assessment: Designing tests to promote positive washback. In P. K. Matsuda, C. Ortmeier-Hooper, & X. You (Eds.), *The politics of second language writing: In search of the promised land* (pp. 222-244). West Lafayette, IN: Parlor Press.

White, E. M. (1986). Pitfalls in the testing of writing. In K. L. Greenberg, H. S. Wiener, & R. A. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 53-78). White Plains, NY: Longman.

White, E. M. (1994). *Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance* (2nd ed.). San Francisco: Jossey-Bass.

Williamson, M. M. (1993). An introduction to holistic scoring: The social, historical, and theoretical context for writing assessment. In M. M. Williamson & B. A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theorical and empirical foundations* (pp. 1-43). Cresskill, NJ: Hampton Press.

Wilson, M. (2006). *Rethinking rubrics in writing assessment*. Portsmouth, NH: Heinemann.

Wolcott, W., & Legg, S. M. (1998). *An overview of writing assessment: Theory, research, and practice*. Urbana, IL: NCTE.

Zhu, W. (2004). Writing in business courses: An analysis of assignment types, their characteristics, and required skills. *English for Specific Purposes, 23*(2), 111-135.