

Abstract: Essay tests are widely used to assess ESL/EFL learners' writing abilities for instructional, administrative, and research purposes. Relevant literature was searched to identify 70 empirical studies on ESL/EFL essay tests. The majority of these studies examined task, essay, and rater effects on essay rating and scores. Less attention has been given to the effects of examinee factors, scoring methods, and assessment contexts. This absence seems mainly to be the result of a traditional concern with controlling for task and rater variability as 'sources of measurement error.' This article argues for viewing these factors as 'sources of variability' that contribute to the richness and uniqueness of the contexts within which writing performance and assessment occur and for taking them into account when interpreting and using essay test scores. The paper concludes with several implications for research and practice.

Keywords: essay tests; variability; rating scales; essay raters; examinees; writing tasks

Résumé : On utilise largement la production écrite pour évaluer les habiletés écrites d'apprenants en ALS/ALE à des fins administratives, d'enseignement ou de recherche. J'ai recensé la littérature pertinente pour identifier 70 études empiriques sur les épreuves de production écrite en ALS/ALE. La majorité de ces études portent sur l'effet des tâches, des textes et des correcteurs sur le processus de correction et le résultat. On a porté moins d'attention aux effets de facteurs liés aux apprenants, aux méthodes de correction ou au contexte de l'évaluation. Il semble que ce soit principalement attribuable au souci habituel de contrôler la variation au niveau des tâches et des correcteurs comme «sources d'erreur de mesure». Cet article défend l'idée de considérer ces autres facteurs comme des «sources de variation» qui contribuent à la richesse et à la particularité des contextes dans lesquels se réalisent la performance écrite et son évaluation et pour les prendre en compte dans l'interprétation et l'utilisation des résultats aux épreuves de production écrite. L'article conclut avec plusieurs implications pour la recherche et la pratique.

Mots clés : épreuves de production écrite; échelles d'appréciation, correcteurs de productions écrites, candidats à une épreuve; tâches d'écriture

Essay tests have become the most widely used method for assessing writing for administrative and instructional decisions about ESL/EFL (English as second/foreign language) learners, teachers, and programs. Essay tests are also frequently used in research as elicitation techniques to investigate the nature and structure of second language (L2) writing proficiency and development, variability in L2 learners' writing, and the effects of different instructional settings and techniques on learning to write in L2 (Cumming, 1997, 2001; Hamp-Lyons, 1990, 2003; Weigle, 2002). Following Hamp-Lyons (1991a), Weigle (2002) listed seven main characteristics of essay tests that distinguish them from indirect tests of writing, such as multiple-choice tests, and portfolio assessment: examinees must write at least one piece of continuous text; examinees are provided with a set of instructions (a prompt) but have considerable freedom in how to respond; texts are written within a limited time (e.g., 30 minutes); the writing topic is unknown to the examinees in advance; each text is read by at least one, and normally two or more, trained raters; judgements of text quality are based on a common set of criteria in the form of a rating scale and/or sample responses; and judgements are expressed as numbers rather than, or in addition to, verbal descriptions.

These characteristics often lead to variability in writing test performance and scores. For example, examinees may obtain different scores, depending on the task they are assigned and the rater who marks their essays. Traditionally, this variability has been seen as 'measurement error' that lowers the reliability and, hence, the validity of essay tests. As Groot (1990, p. 11) explained, reliability asks, 'To what extent do differences in scores between learners reflect differences in [writing] ability, rather than other factors?' Validity asks, 'Do the test scores indeed reflect the ability the test is intended to measure?' Reliability is a necessary, but not sufficient, condition for validity, in the sense that test scores that are not reliable cannot provide a basis for adequate and appropriate interpretation and use (Bachman, 1990). For example, if raters differ greatly in the way they apply the assessment criteria from one essay to another or they differ from each other in their judgements of the same essay, the scores they assign to the essays will not be valid indicators of students' writing ability and, hence, 'cannot be a valid basis on which to make decisions of great import to the students and the institution' (Jacobs, Zinkgraf, Wormuth, Hartfiel, & Hughey, 1981, p. 23).

To enhance score consistency and validity, measurement practitioners have often recommended such practices as the standardization of tasks, administration conditions, and scoring procedures

(e.g., Coffman, 1971; Diederich, French, & Carlton, 1961; Henning, 1987; Pilliner, 1968; Underhill, 1982; Wood, 1991). Recently, however, these practices have come under attack on the grounds that, although they may improve score consistency, they often reduce test validity. For example, rater training and scoring schemes that are meant to enhance inter- and intra-rater reliability, it is argued, often force raters to ignore their own experiences and expertise when interacting with and judging student writing, thus sacrificing a 'true [or valid] reading' of a text for a 'reliable' one (Huot, 1993, p. 211). Moreover, by attempting to minimize or control for those factors that are most likely to create variation in writing test performance, these practices often simplify the writing and reading experiences and remove from them the particulars of the context in which they typically occur (Broad, 2003; Camp, 1993; Hamp-Lyons & Condon, 2000; Huot, 2002; Moss, 1992, 1994, 1996; White, 1993; Williamson, 1993).

Recent scholarship emphasizes that variability is the norm rather than the exception in language performance and that it is a natural part of people's writing and reading (Broad, 2003; Deville & Chalhoub-Deville, 2006; Huot, 2002; Swain, 1993). Deville and Chalhoub-Deville, for instance, urged practitioners and researchers to 'sidestep the traditional thinking' about variability in L2 performance and scores as 'error' or 'noise that needs to be suppressed' and to see it as 'richness that needs to be understood and mapped' (p. 16; cf. Broad; Camp, 1993; Huot, 2002; Moss, 1992, 1994, 1996). This call shifts the focus from ways to control and standardize writing assessment practices to a focus on identifying and explaining the sources of variability in authentic L2 writing assessment contexts and how they influence the accuracy of essay test scores and, hence, the validity of the inferences and the fairness of the decisions that educators make based on such scores (Broad; Cumming, 1997; Deville & Chalhoub-Deville; Huot, 2002). To this end, the current review of empirical studies on ESL/EFL essay tests is undertaken with the goal of better understanding the sources of variability in essay test performance, drawing lessons for practice, and highlighting areas for further research.

Search procedures

A preliminary survey of previous reviews of writing assessment research (Brossell, 1986; Cumming, 1997; Cumming, Kantor, Powers, Santos, & Taylor, 2000; Hamp-Lyons, 1990, 1991a, 2003; Hamp-Lyons & Kroll, 1997; Kroll, 1998; McNamara, 1996; Ruth & Murphy, 1988;

Weigle, 2002) indicated that any essay test involves at least two participants, a writer-examinee and a reader-rater, and three texts, a writing task, an essay (a written product), and a rating scale, within a specific sociocultural context (e.g., institution) that specifies the purposes, and possibly processes, of writing as well as reading and judging examinees' essays. These texts and participants, which seem to explain a major proportion of the variability in essay test performance and scores (McNamara, 1996; Milanovic, Saville, & Shuhong, 1996; Weigle, 2002), served as a set of themes or categories for locating, reviewing, organizing, and integrating published empirical studies on ESL/EFL essay tests, and for identifying under-explored areas in need of further investigation. For example, under writing task, I was able to identify several studies that have examined writing task effects on examinees' essay characteristics (e.g., Hinkel, 2002), writing processes (e.g., Connor & Carrell, 1993), rating processes (e.g., Cumming, Kantor, & Powers, 2002), and essay scores (e.g., Weigle, 1999). As Weigle (2002) has pointed out, however, few studies have investigated the effects of contextual factors on participants' writing and rating performance in essay tests. These studies examined the effects of essay order, length and time of rating session, and type of training received on L1 (first language) essay test scores (e.g., Daly & Dickson-Markman, 1982; Freedman & Calfee, 1983). Fewer studies have examined how the sociocultural context affects L2 essay writing and rating processes and outcomes.¹ As a result, this paper considers writing assessment contexts only briefly and in relation to the participants and texts identified above. As Kroll (1998) argued, contexts vary in multiple ways, but the two types of participants and three types of texts (described above) are inevitably present in any essay test. Nevertheless, while it is almost impossible to identify all contextual factors and their effects, it is important to keep in mind that 'any assessment takes place in a given social and cultural context and may not be generalizable outside of that context' (Weigle, 2002, p. 60).

The following parameters were set to define which research studies to include in the current review: empirical studies, both quantitative and qualitative, on essay tests (but not other forms of writing assessment) of English as a second or foreign language (ESL/EFL) published in English between 1980 and 2006 in peer-reviewed journals or as technical reports, books, or book chapters (but not dissertations). I first searched a number of educational journals, technical reports, and books manually (see Appendix 1 for a list of journals searched) for relevant studies. I then reviewed the bibliographies of these studies to

TABLE 1
Number of studies included in the review

Writer-examinee	Reader-rater	Writing task	Essay	Rating scale
10	22	21	41	5

identify additional references. When I found relevant references in those bibliographies, I added them to my database. The final corpus included in the review contained 70 studies. Table 1 indicates the number of studies identified for each text and participant in essay tests. It should be noted here that several studies examined two or more of these factors and interactions between them (e.g., Rinnert & Kobayashi, 2001; Weigle, 1999), hence the discrepancy between the total number of studies reviewed and the sum of the numbers of studies in Table 1 (see Appendix 2 for a complete list of the studies reviewed). For ease of presentation, each participant and text is discussed separately in the following sections. However, this should not draw attention away from the interactive nature of essay tests. Assessment processes and scores are the result of the very complex interactions of all the factors discussed below (Hamp-Lyons & Kroll, 1997; Kroll, 1998; McNamara, 1996; Weigle, 2002).

Participants and texts in essay tests

Participant 1: The examinee-writer

Examinees bring a variety of cognitive, affective, linguistic, and sociocultural factors to the ESL/EFL essay test. These factors affect how examinees choose, read, interpret, and respond to writing tasks, as well as their writing processes, texts, and scores. I was able to identify 10 studies on examinee factors in ESL/EFL essay tests. Only two of these studies investigated how examinees choose, read, and interpret writing tasks. Connor and Kramer (1995) found that the task representation of ESL students in a reading-to-write task differed from that of NSE (native English speaker) students. On the basis of this finding, Connor and Kramer speculated that other examinee factors – such as prior knowledge, personal experiences, cultural and educational background, and language proficiency – might affect the way examinees interpret tasks and what constitutes a successful completion of a task. Similarly, Polio and Glew (1996) found that students' backgrounds play an important role in students' choices of prompts in an ESL essay test. Both studies concluded that there is

a need for more research on how examinees from different backgrounds read, interpret, choose, and respond to writing tasks in ESL/EFL essay tests.

I identified five studies that examined the effects of writer factors on ESL/EFL essay features. All of them focused on the effects of examinees' L1, often by comparing texts written by learners from different L1 backgrounds and by NSEs. These studies suggest that essays written by students from different L1 backgrounds differ in their linguistic, stylistic, and rhetorical characteristics (Fraser, Faletti, Ginther, & Grant, 1999; Hinkel, 2002; Park, 1988; Reid, 1990; Scarcella, 1984). A recent example of this line of research is a large-scale study conducted by Hinkel (2002) to compare 68 syntactic, lexical, and rhetorical features of essays written by about 1,500 NSE and ESL university-level students (Chinese, Arab, Japanese, Korean, Vietnamese, and Indonesian) on six topics. Hinkel found significant differences between the essays of NSE and ESL writers as well as between the essays of the different ESL groups in these features. I am not aware of any published studies on the effects of other examinee characteristics on ESL/EFL essay features or examinees' writing processes.

Similarly, there has been little research on the effects of writer factors on ESL/EFL essay test scores. Such studies are usually conducted within the framework of bias analysis or differential item functioning (DIF), which examines whether and how items or tasks function differently for examinees with similar ESL/EFL abilities but from different backgrounds. In one of the earliest studies, Kunnan (1990) found that some items in an ESL test that includes a sub-test of writing error-detection functioned differently across gender and L1 groups. More recently, Breland, Lee, Najaran, and Muraki (2004) and Lee, Breland, and Muraki (2004) examined the comparability of TOEFL writing prompts for different test-taker subgroups in terms of gender and L1 background. Both studies found a small to medium impact of these examinee factors on essay scores and provided several suggestions for task development and review to minimize these potential biasing effects. As the small number of studies in Table 1 indicates, there is a clear need for more research on the effects of examinee factors on ESL essay test performance and scores.

Participant 2: The reader-rater

Twenty-two studies examined rater factors in ESL/EFL essay tests. These studies indicate that rater factors – such as personality,

cultural, linguistic, and educational background, teaching and rating experience – influence rater decision-making behaviour, interpretations and expectations concerning task requirements and scoring criteria, reaction to ESL/EFL essays, severity (inter-rater reliability), and self-consistency (intra-rater reliability). Moreover, raters may display different patterns of behaviour and interactions with aspects of the rating context (e.g., essay order, trainer), which may result in different scores assigned to the same essays across raters and rating occasions (McNamara, 1996; Weigle, 2002). This variability has been the main reason for the use of indirect tests of writing and the development of computer programs that can simulate human readings to generate reliable scores (Broad, 2003; Huot, 2002). Three rater characteristics have received most of the attention in the literature on ESL/EFL essay tests: raters' L1 background, academic background, and teaching and rating experience.

The findings concerning the effects of rater L1 background on essay scores are mixed. Connor-Linton (1995), for example, did not find significant differences between the scores assigned by NSE and Japanese EFL teachers to EFL essays. Shi (2001) also did not find significant differences between the ratings assigned by NSE and Chinese EFL teachers to EFL essays. Both studies, however, found that the raters provided different qualitative reasons for assigning the same scores, which suggests that the raters might have emphasized different characteristics of writing in their evaluation. Hill (1997), in contrast, comparing the holistic ratings of NSE and Indonesian EFL teachers of 100 EFL essays, found that the NSE group were significantly harsher than the EFL group. Kobayashi (1992) also found significant differences between the scores assigned by NSE and Japanese raters to ESL essays. Kobayashi compared scores assigned by 269 NSE and Japanese professors and graduate and undergraduate students to two ESL essays in terms of grammaticality, clarity of meaning, naturalness, and organization. Kobayashi found that the NSE raters were stricter about grammaticality than were the Japanese raters. The NSE professors and graduate students, however, evaluated the clarity of meaning and organization of the essays more positively than did the comparable Japanese-speaking groups. Finally, the Japanese undergraduates evaluated both essays much more positively than did their NSE counterparts.

As Kobayashi's (1992) findings indicate, one issue in studies on the effects of rater language background on essay scores is rater-essay interactions. For example, Kobayashi and Rinnert (1996) – comparing the reaction of NSE and Japanese EFL teachers to essays rewritten to

reflect weaknesses and strengths in rhetorical structure, sentence-level errors, and coherence – found that the Japanese readers reacted more favourably than the NSE readers did to EFL essays that contain Japanese rhetorical patterns (cf. Hamp-Lyons & Zhang, 2001). Rinnert and Kobayashi (2001) reported similar findings for a study that had the same design but compared the ratings of NSE EFL teachers and those of Japanese inexperienced EFL students, experienced EFL students, and experienced EFL teachers. In addition, Rinnert and Kobayashi reported a significant interaction effect between rater L1 background and experience. The inexperienced EFL students attended mainly to content when judging and commenting on the essays; the experienced students and Japanese EFL teachers, like the NSE teachers, attended more to clarity, logical connections, and organization. Rinnert and Kobayashi interpreted this finding as indicating a gradual change in Japanese readers' perceptions of EFL essays from preferring the writing features of their L1 to preferring many of the L2 writing features. Hamp-Lyons (1989) reported a similar finding with NSE raters; she found that experience with other languages altered NSE readers' responses to the English writing of members of those language communities.

Raters' academic and educational backgrounds may refer to whether the rater is a language or content teacher and whether she or he is an ESL or English-as-L1 teacher. Raters from different disciplines have been reported to have different assumptions and expectations about tasks, essays, and rating criteria. Several studies found that faculty from different departments rated and reacted differently to different aspects of ESL essays and disagreed on when various criteria were being met. Santos (1988) and Vann, Lorenz, and Meyer (1991), for instance, found that the reader's academic discipline was a significant predictor of rater tolerance of ESL essay language errors, with humanities and social sciences faculty being more tolerant than faculty in the physical sciences (cf. Leki, 1995; Lukmani, 1996; Van, Meyer, & Lorenz, 1984). Mendelsohn and Cumming (1987) compared the perceptions and scores of 26 engineering, English literature, and ESL instructors to eight ESL essays manipulated to reflect effective and ineffective language use and rhetorical organization. They found complex interactions between essay characteristics and raters' backgrounds, particularly for middle range essays. The instructors from the three disciplines tended to agree on the rating of high and low proficiency essays, but disagreed on middle range essays. When judging these essays, the engineering professors attributed more importance to language use; the ESL

instructors gave more weight to rhetorical organization; while the English teachers did not seem to be biased in either direction.

Four other studies compared the ratings of ESL and English composition teachers. Brown (1991) found no statistically significant mean differences in the holistic ratings assigned by ESL and English teachers to ESL essays, although the two rater groups gave different qualitative reasons for assigning these scores. Similarly, in a study comparing the holistic and multiple-trait scores assigned by ESL and English teachers to ESL essays, O'Laughlin (1994) found no significant differences between the holistic scores of the two groups. However, the English teachers rated the essays significantly more harshly than the ESL teachers did on the multiple-trait scale. In addition, the two groups seem to have weighted the essay features differently in arriving at their holistic judgements of the essays. Song and Caruso (1996), on the other hand, found that while English faculty assigned significantly higher holistic scores to ESL essays than did the ESL teachers, the average multiple-trait scores were similar across the two groups of raters. Finally, in a study that used think-aloud protocols to examine the rating processes of ESL and English composition raters, Cumming et al. (2002) found that while both groups displayed fundamentally the same decision-making behaviours when rating ESL essays holistically, raters with L2 work experience tended to focus more on language issues than did raters with L1 work experience, who focused more on ideas or content.

Regarding the effects of rater experience on essay scores, Song and Caruso (1996) found that experienced teachers assigned higher holistic scores to ESL essays. Shohamy, Gordon, and Kraemer (1992), in contrast, in a study comparing experienced and novice raters of EFL essays, found no significant differences across groups in inter-rater reliability. Other studies found that the effect of rater expertise on essay scores depended on other factors. For example, Weigle (1999) found an interaction effect between rating expertise and writing task. While the experienced raters in her study found it easier to rate ESL essays that adopt a variety of approaches to the same task than essays that are personal and have a limited number of response possibilities, because of over-familiarity with the latter type of essays, the inexperienced raters showed the opposite pattern.² As described above, Rinnert and Kobayashi (2001) found that Japanese raters with different levels of experience attended to different aspects of writing and that, as they gain more experience, their perceptions of EFL essays seem to gradually change from preferring the writing features of Japanese to preferring many of the L2 writing features.

There is a relatively more extensive literature on the effects of rater expertise on the rating processes of ESL essay raters (Cumming, 1990; Delaruelle, 1997; DeRemer, 1998; Erdosy, 2004; Weigle, 1999). This research indicates that experienced and novice raters employ qualitatively different rating processes. In one of the earliest studies in this area, Cumming (1990) found that experienced teachers had a much fuller mental representation of the essay assessment task and used a large and varied number of criteria, self-control strategies, and knowledge sources to read and judge ESL essays. Novice teachers, in contrast, tended to evaluate essays with only a few of these component skills and criteria, using skills that may derive from their general reading abilities or other knowledge they have acquired previously (e.g., by editing essays) (p. 43). Likewise, Delaruelle (1997) and Weigle (1999) found that experienced raters had a broader range of responses and reading repertoires upon which to draw when scoring ESL essays than did novice raters.³ Finally, Erdosy (2004) found that differences in ESL raters' teaching experiences, and to a lesser extent native language, led them to use different assessment criteria.

Text 1: The writing task

From a measurement viewpoint, the writing task should elicit a written response from the examinee, affecting neither the examinee's nor the rater's performance. Recent scholarship, however, indicates that language performance is context dependent and, consequently, that varying task requirements is likely to lead to variation in learners' performance (e.g., Deville & Chalhoub-Deville, 2006; Swain, 1993). Several studies have shown empirically that factors such as task choice, wording, content, rhetorical context (i.e., audience and purpose), cognitive demand, discourse mode (e.g., argumentative, narrative), time allotment, instructions, genre (e.g., letter, essay), transcription mode (e.g., handwritten, word-processed), and input materials (e.g., reading) affect examinees' writing processes and essay features, rater decision-making behaviour and reliability, as well as essay test scores.

Several studies of L2 writing have shown that task factors affect the processes that learners employ and the textual aspects they attend to (e.g., language, discourse) when writing in L2 (e.g., Clachar, 1999; Cumming, 1989; Raimes, 1987). Other studies examined the effects of task factors on examinees' ESL/EFL essays. These studies found that variation in task requirements and instructions

significantly influence the linguistic and rhetorical characteristics of ESL examinees' essays (Campbell, 1990; Cumming et al., 2005; Hinkel, 2002; Park, 1988; Porter & O'Sullivan, 1999; Reid, 1990; Tedick, 1990; Zhang, 1987). For example, Hinkel, comparing the essays of NSE and ESL students on six prompts that differed in wording and content, found that the quality of the essays was determined by the grammar and vocabulary of the prompt, as students tended to insert the lexis and grammatical constructions of the prompt into their own texts. Furthermore, more personal topics resulted in more personal style, while topics that are more distant from the students' personal experiences led to ESL essays that are closer to native-like uses of language features. Contrary to common beliefs, Hinkel concluded that 'the greater writer's familiarity and experience with a topic is and the easier it is to write about, the simpler the text can be' (p. 241).

Task characteristics can also influence examinees' task selection in essay tests. Chiste and O'Shea (1988), for example, found that ESL writers favoured the first and second questions and/or the shortest or second shortest questions in each set of four questions. Choosing the shorter and/or earlier positioned questions in a set did not correlate with better performance, however. Polio and Glew (1996) also found that question type and the specificity of the topic, as well as student background and time constraints, significantly influenced students' choices of prompts in an ESL essay test. To my knowledge, only one study has examined the effects of allowing examinees task choice on ESL essay test scores. Jennings, Fox, Graves, and Shohamy (1999) compared the scores and the textual features of essays written by 254 ESL students randomly assigned to one of two conditions: no choice of topic or choice among five topics. Jennings et al. found that although the scores of the choice group were overall higher than those of the no-choice group and that most participants preferred to have a choice, the differences in scores and essay features across the two conditions were not statistically significant.

Task characteristics can also influence rater performance and reliability. Coffman (1971), for example, argued that inter- and intra-rater reliability are likely to decrease if the essay question allows freedom of response to what and how to write (cf. DeGrujter, 1980; Schoonen et al., 1997). In a study that examined the effects of task type on rater behaviour and scores, Weigle (1999) found that inexperienced raters found it more difficult to apply the scoring rubric to a graph task, which led them to assign lower scores to essays on this task. Experienced raters, in contrast, found it easier to adjust to the variety of approaches taken to the graph task, but the personal nature and

limited number of response possibilities for a choice task caused problems for them because of over-familiarity with such essays.

Several studies have found significant task effects on ESL essay test scores (Breland, Lee, Najarian, and Muraki, 2004; Hamp-Lyons & Mathias, 1994; Porter & O'Sullivan, 1999; Tedick, 1990). Spaan (1993), however, did not find any significant differences between the scores assigned to essays on different tasks. These studies examined the effects of different dimensions of the writing task. Spaan considered task discourse mode (argumentative and narrative) and content (personal and impersonal). Tedick investigated the effects of task content (general and specific). Porter and O'Sullivan focused on the effects of audience specifications. Finally, Breland et al. compared tasks that differ at several levels (content, number of words, topic familiarity, and discourse mode).

Other studies suggest that other factors in the writing assessment context (e.g., examinee, rater) can mediate task effects on essay scores (Ruth & Murphy, 1988; Weigle, 2002). For instance, in a study to investigate the relationship between task difficulty as judged by experienced essay raters and essay scores, Hamp-Lyons and Mathias (1994) found that, surprisingly, the tasks judged to be difficult (argumentative impersonal topics) resulted in higher mean essay scores than those tasks judged to be easy (expository personal topics). Hamp-Lyons and Mathias hypothesized that this might be due to task-examinee and/or rater-task interactions. Some ESL writers may be less comfortable with or not used to writing on personal tasks for cultural reasons, while raters might, consciously or unconsciously, compensate in their scoring for relative task difficulty. Likewise, Weigle (1999) found that task effects depended on rater experience and training. Inexperienced raters in her study assigned significantly lower scores to ESL essays commenting on a graph than to essays on a choice topic before training. After training, however, this difference disappeared. The experienced raters did not show such differences in severity.

These findings indicate that raters do not necessarily share the same assumptions and expectations about a writing task. Few studies have examined this issue, however. Connor and Carrell (1993) compared examinees' and raters' interpretations of the writing task in an ESL essay test. They found that both examinees and raters exhibited little concern with addressing all the parts of the task as specified in the prompt. In particular, the raters' reading and evaluation of the essays were not affected by the way the examinees addressed the rhetorical requirements specified by the task. Rather, the

raters put more emphasis on language use in terms of fluency, infrequency of errors, and general development of ideas – aspects that are independent of task requirements. This finding seems to suggest that the task has little effect on the rating criteria that ESL essay raters employ. Weigle (1999) and Cumming et al. (2002), however, found that task type does affect ESL essay rating processes. As described above, Weigle (1999) found significant interaction effects between rater experience and task requirements. Likewise, Cumming et al. (2002) reported significant task effects on rater behaviour and the essay features that raters attend to. Raters in this study paid more attention to rhetoric, ideas, and task completion than language when rating reading- and listening-based essays, but made more language comments when reading essays on bare topics. There is a need for more research on whether and how elements of the writing task influence the expectations that raters bring to the rating task, their decision-making, and the writing aspects they attend to.

Text 2: The essay

Another way to address the question of task effects on rater performance and essay test scores is to investigate how different tasks affect examinees' essay features and how these, in turn, affect rater performance and scores (Connor & Mbaye, 2002; Cumming, 1997; Upshur & Turner, 1999). Almost all the studies reviewed above considered, in one way or another, aspects of ESL/EFL essays. This finding is not surprising, since the examinee's writing performance or product is the primary focus of any writing assessment hence the large number of studies ($n=41$) under this category in Table 1. These studies investigated the effects of various essay features on scores by examining the relationship between essay holistic scores, on the one hand, and multiple-trait scores (e.g., Bacha, 2001; O'Laughlin, 1994; Song & Caruso, 1996; Tedick & Mathison, 1995) or objective measures of essay features (e.g., Connor, 1991; Frase et al., 1999; Homburg, 1984; Perkins, 1980; Sweedler-Brown, 1993a, 1993b), on the other, using correlational methods (e.g. multiple regression). Other studies isolated specific aspects of ESL/EFL essays such as language errors (e.g., Janopoulos, 1992; Khalil, 1985; Vann et al., 1984, 1991) or rewrote essays to reflect strengths and weaknesses in specific areas (e.g., Kobayashi & Rinnert, 1996; Lukmani, 1996; Mendelsohn & Cumming, 1987; Rinnert & Kobayashi, 2001; Yeh, 1998) to investigate their effects on rater perceptions and essay scores. This line of research has important implications for automated scoring

systems, which use textual features that best predict essay holistic scores to generate scores for other essays (for a review, see Burstein & Chodorow, 2002, and Chodorow & Burstein, 2004).

However, the findings of studies on the relationships between essay features and scores are mixed, perhaps because of the variability in task requirements across studies, the different ways textual features are defined and measured, the complex rater-essay interactions (Mendelsohn & Cumming, 1987; Rinnert & Kobayashi, 2001), the non-linearity of the relationship between essay features and holistic scores (Jarvis, Grant, Bikowskia, & Ferris, 2003), and the possibility that other variables (e.g., task type, rater background) can mediate the relationship between essay features and scores. Almost all essay characteristics (e.g., writing mode, content, organization, length, vocabulary, errors, spelling, coherence, style, support) have been found to significantly correlate with holistic essay scores (Breland, Lee, & Muraki, 2004; Chodorow & Burstein, 2004; Connor, 1991; Engber, 1995; Frase et al., 1999; Gamaroff, 2000; Grant & Ginther, 2000; Homburg, 1984; H.K. Lee, 2004; Lukmani, 1996; Rinnert & Kobayashi, 2001; Santos, 1988; Schneider & Connor, 1990; Song & Caruso, 1996; Sweedler-Brown, 1993a, 1993b; Tedick & Mathison, 1995; Yeh, 1998). To address the diversity of features assumed to affect ESL essay scores, Homburg (1984) proposed a 'funnel model' according to which readers consider various essay characteristics but first broadly categorize essays on the basis of one salient feature. Raters then use combinations of other features to determine more finely tuned categorizations. Recent studies of rater decision-making (e.g., Cumming et al., 2002; Wolfe et al., 1998), however, have questioned this linear model and argued that essay rating is a complex, iterative process in which 'raters make multiple evaluation decisions, each of which revises the previous one' (Wolfe et al., 1998, p. 469).

Other studies used think-aloud protocols to examine the effects of different essay characteristics on rater behaviour (Cumming, 1990; Cumming et al., 2002; Delaruelle, 1997; Hamp-Lyons & Zhang, 2001; Lumley, 2002, 2005; Milanovic et al., 1996; Smith, 2000; Vaughan, 1991). These studies have identified a variety of textual features that raters mention when evaluating ESL essays. For example, Vaughan found that most of the comments raters made in her study concerned, in a descending order, content, handwriting, tense/verb problems, and punctuation. As Lumley (2005) has cautioned, however, frequency of comment does not necessarily mean that the feature mentioned affects the outcomes of the rating. In addition, several studies have shown

that raters attend to different textual features, depending on such factors as text type, essay proficiency level, and rater background. For example, Delaruelle (1997) found that register was salient for an interpersonal text, while cohesion was mentioned more frequently for a persuasive text; organization, grammar, and task fulfilment were equally salient for both text types. Cumming (1990) and Cumming et al. (2002) found that raters attended more to language features when rating low-proficiency essays but to both rhetoric and language when reading high-proficiency essays (cf. Connor-Linton, 1995; Shi, 2001). Hamp-Lyons and Zhang (2001) showed how raters from different L1 backgrounds react differently to the same essay's rhetorical aspects.

Finally, some studies found that essay characteristics can influence rater reliability (H.K. Lee, 2004; Wolfe & Manalo, 2005) and the ease with which raters are able to mark essays (Weigle, 1999). For example, H.K. Lee found that scores assigned to word-processed essays were more reliable than were scores assigned to handwritten essays, which seems to suggest that handwritten texts yield less consistent decisions from raters.

Text 3: The rating scale

Davies et al. (1999) defined a rating scale as 'a scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged' (p. 153). Three main types of rating scales are widely used in essay tests: holistic, multiple-trait, and primary-trait.⁴ In holistic scoring, the rater considers individual elements of performance but chooses one score to reflect the overall performance. Multiple-trait scoring involves assigning multiple sub-scores to individual traits or dimensions (e.g., grammar, mechanics, text organization). For primary-trait scoring a single score is assigned to an essay according to the degree to which the writer has addressed the specific requirements of the task (Cooper, 1977; Goulden, 1992, 1994; Hamp-Lyons, 1991b; Lloyd-Jones, 1977; Perkins, 1983; Weigle, 2002).

As Weigle (2002) pointed out, the literature is replete with guidelines for developing and validating rating scales (e.g., Brown & Bailey, 1984; Hamp-Lyons & Henning, 1991; Sasaki & Hirose, 1999; Turner, 2000; Turner & Upshur, 1996) as well as arguments for and against different rating methods (e.g., Hamp-Lyons, 1991b, 1995; Hamp-Lyons & Kroll, 1997; Perkins, 1983; Wood, 1991). For instance, Perkins (1983) argued that while holistic scoring is weak in reliability, it has high validity when overall attained writing proficiency is

the construct to be assessed. Multiple-trait scoring, in contrast, enhances reliability but lacks in practicality and is of questionable validity because it isolates text features from context. Wood (1991) also argued that, because rater variability is often large, it is better to use the sum of holistic scores of several readers than to have a single reader assign several different scores to the same essay, which compounds measurement error. Hamp-Lyons (1991b, 1995), on the other hand, contended that while holistic scoring is appropriate for scoring L1 essays, multiple-trait scoring has higher validity and reliability when rating L2 essays, because different learners have different levels of proficiency in different aspects of L2 writing. Furthermore, multiple-trait scoring provides more information on students' performance.

Surprisingly, however, there is little empirical research on the effects of different scoring methods on essay rating processes and scores. Thus, although the major difference between holistic and multiple-trait scoring, for example, concerns their assumptions about the relationship between the parts and whole of the performance or product being assessed and the number of scores assigned to an essay (Cumming, 1997; Goulden, 1992, 1994), it is not clear how these differences affect essay reading and rating, rater severity and self-consistency, and essay scores (Hamp-Lyons & Kroll, 1997; Weigle, 2002). As Table 1 shows, most research has focused on rater and task effects on ESL essay rating and scores, perhaps because most assessment systems include several tasks and raters but only one scoring method. As Schoonen (2005) has argued, however, 'The effects of task and rater are most likely dependent on what has to be scored in a text and how it has to be scored. In other words, trait and scoring procedure will mediate the task and rater effect' (p. 5).

I was able to identify five studies that examined rating scale effects in ESL/EFL essay tests; all of them compared holistic and multiple-trait scoring methods (Bacha, 2001; Carr, 2000; H.K. Lee, 2004; O'Laughlin, 1994; Song & Caruso, 1996).⁵ The findings of these studies are mixed. As discussed above, O'Laughlin found that English-as-L1 and ESL teachers differed significantly in the multiple-trait, but not the holistic, scores they assigned to ESL essays. In addition, holistic ratings achieved a higher inter-rater reliability. In contrast, Song and Caruso, also comparing the holistic and multiple-trait ratings of English and ESL teachers, found significant differences in the holistic, but not the multiple-trait, scores. Both studies indicate a significant interaction effect between rater background and scoring method, though in

different directions. O'Laughlin concluded that while it is more reliable, holistic rating is less valid because it seems to conceal important differences between raters of different backgrounds and professional experience. The multiple-trait rating seems more faithful to real dissimilarities between raters. Song and Caruso, on the other hand, concluded that while rater teaching and rating experience may affect their holistic scores, this may not be true for multiple-trait scoring because this method minimizes or controls for the effects of rater factors 'by focusing raters' attention on the same qualitative aspects, or identified features, of a composition' (p. 175).

H.K. Lee (2004) also found an interaction effect between writing mode (handwritten or word-processed) and scoring method. Holistic rating yielded no significant mean differences across writing modes, while multiple-trait rating resulted in significantly higher scores for the word-processed version of an ESL essay test. Lee interpreted this finding to suggest that holistic and multiple-trait ratings measure different constructs. Bacha (2001), on the other hand, comparing the effects of multiple-trait and holistic scoring on student placement in an EFL program, found high correlations between the two sets of scores as well as high inter- and intra-rater reliabilities for both methods. Finally, Carr (2000) examined how multiple-trait and holistic rating scales affect scores in an ESL test that includes a writing component. The results of factor and regression analyses indicated that altering the rating scale changed the interpretation of the writing test, resulting in total test scores that were not comparable because the factor structure of the test itself changed. For the writing component Carr concluded that 'the difference between [multiple-trait] and holistic scales is principally one of focus: holistic scores provide an assessment of a single construct, whereas composite scores from [a multiple-trait] rating scale conflate the information from several constructs' (p. 228).

I am not aware of any qualitative study that has examined the effects of different rating scales on essay rating. Most studies have investigated the decision making and essay aspects that raters focus on when rating essays with no specific rating guidelines (e.g., Cumming et al., 2002), or when using holistic (e.g., Milanovic et al., 1996) or multiple-trait rating scales (e.g., Cumming, 1990; Lumley, 2002; Smith, 2000).⁶ Lumley (2002, 2005) and Smith (2000) may be two exceptions in that, although they did not specifically compare different scoring methods, their findings raise several relevant questions concerning the role of the scoring method in essay rating. Examining the rating processes of four experienced

ESL essay raters, Lumley (2002) found that the raters faced problems reconciling their impression of the text, the specific features of the text, and the wordings of the rating scale. Second, the relationship between scale contents and text quality remained obscure. Third, although they seem to have understood the rating contents similarly in general terms, the raters might have applied the contents of the scale in different ways and emphasized different components of the scale descriptors. Fourth, the raters seem to have formed their judgements independently of the scale wordings but 'somehow managed to refer to the scale content' to articulate and justify their scoring decisions (p. 263). Smith (2000) also reported that the six raters in his study attended to other textual features in addition to those mentioned in the rating scale and that raters with different reading strategies interpreted and applied the rating criteria differently. Clearly, there is a need for more research on whether and how the assumptions, content, organization, and number of rating dimensions and proficiency levels in rating scales affect rater decision making, severity, and self-consistency (Cumming, 1997; Davidson, 1991; Hamp-Lyons & Kroll, 1997; McNamara, 1996; Weigle, 2002).

Implications and conclusions

This review has focused on empirical studies published in English between 1980 and 2006 that explored text and participant variables in ESL/EFL essay tests. While the review highlights the growing number of empirical studies of essay tests, it also draws attention to the conflicting results of these studies, crucial areas still under-researched, and fundamental questions that remain unanswered. In particular, while numerous studies have examined the roles and effects of writing tasks, essay features, and rater characteristics on essay writing and rating performance and scores, little attention has been given to examinee factors, scoring methods, and assessment contexts. This lack seems mainly due to the dominant view of task and rater variability as 'sources of measurement error' that must be identified, estimated, and then eliminated or reduced (Coffman, 1971; Diederich et al., 1961).

However, we need to go beyond this limiting view. More specifically, we need to view the texts, participants, and processes discussed in this paper as 'sources of variability' that contribute to the richness and uniqueness of the contexts within which L2 writing performance and assessment occur. Such a view has the potential

of enhancing our understanding of the roles of these factors in local assessment systems and practices and their effects on the validity of the inferences and fairness of the decisions that educators make about learners, teachers, and programs based on essay test scores (Broad, 2003; Cumming, 1997; Deville & Chalhoub-Deville, 2006; Huot, 2002).

To further understand ESL essay test performance more research is needed on the texts, participants, processes, and contexts involved in essay tests. The following represent some of the areas that need to be further explored.

Participants: More studies are needed on the effects of writer and rater factors, such as language, cultural, and educational background, affective factors (e.g., anxiety, goals, motivation), background knowledge, learning and assessment history, and other individual variables (e.g., personality type), on L2 writing test performance and scores. In particular, how do these factors affect examinees' and raters' readings, interpretations of, and responses to writing tasks? How do they affect examinees' and raters' writing and rating processes and outcomes?

Texts: How do different writing tasks affect the writing, texts, and scores of ESL examinees? How do task characteristics (e.g., discourse mode, purpose, audience, content) relate to task difficulty? How do they affect essay features and scores? How do task and essay characteristics affect essay rating and outcomes? What role does the rating scale play in the rating? How do rating scales, in assumptions, content, and organization, affect rater decision making, severity, and self-consistency?

Processes: What are the writing and test-taking processes that examinees engage in when taking ESL essay tests? How do these processes compare with those they employ under non-test conditions? For example, what are the effects of time pressure on these processes and the textual aspects that examinees attend to? How do examinees read, interpret, choose, and respond to writing tasks? What are the effects of task, examinee, and other contextual factors on these processes? How do raters mentally represent the rating scale and apply it? Do these mental representations differ across raters and contexts?

Interactions and Contexts: What are the effects of the interactions between the variables in authentic assessment on essay writing and rating and outcomes? What are the effects of the broader sociocultural and institutional contexts within which essay tests occur on test performance and scores?

One strategy to enhance understanding of the effects of the factors discussed above and their interactions on L2 writing assessment and outcomes is the *replication* of previous studies in different contexts. One impediment to such an approach, however, is the lack of consistent and clear definitions of these factors and variables. Different studies seem to define, operationalize, and measure different factors (e.g., task discourse mode, rater experience) in different ways, which may explain the mixed findings reported above. Another explanation of these mixed findings, of course, is the variability of the contexts within which these tests and studies are conducted and the complexity of the interactions between the factors in the assessment, as this review has shown.

Another strategy to explore the sources of variability in L2 writing performance and assessment practices is to combine qualitative and quantitative methods in future studies. Such studies need to continue to use rigorous quantitative procedures and methods to address questions and concerns about the technical qualities of L2 writing assessment systems. But they need also to embrace qualitative methods to answer broader questions about the processes involved in writing assessment and the pedagogical, social, and political contexts and implications of L2 writing assessment practices in specific contexts (Cumming, 2004). Finally, and most importantly, because the different factors and interactions discussed above are highly context dependent, we need more local, in-depth studies of writing assessment practices to address the questions raised above (cf. Broad, 2003; Huot, 2002).

This review offers several implications for ESL writing assessment practice as well. First, practitioners must be aware of and take into account the factors that can influence essay test performance and scores when designing essay tests and making inferences and decisions about ESL/EFL learners, teachers, and programs based on such scores. For example, Hamp-Lyons and Mathias's (1994) findings suggest that there are complex interactions between tasks, writers, and readers that may lead to bias in essay tests. As Hamp-Lyons and Kroll (1997) have cautioned, 'Differences among test-takers are both natural and desirable, since the test has a discriminant function, but these differences must be due to real differences in test-takers' [ESL/EFL] writing abilities and not due to either obvious or subtle bias in the test' (p. 21). As we have seen, bias can occur in the writing task, the scoring procedures, and/or the rater. These biases can have serious ethical, political, and social consequences

for individuals, groups, and programs (Hamp-Lyons, 1996, 2003; Shohamy, 2001).

To protect against task bias, practitioners should pay particular attention to the wording and content of writing tasks and how they may affect students' and raters' performance. They need to be aware of and protect against the possible biased interactions between task variables and writer, rater, and other contextual factors when designing or selecting writing assessment tasks (Hamp-Lyons & Kroll, 1997; Hamp-Lyons & Mathias, 1994; Kroll, 1998; Kroll & Reid, 1994). The studies reviewed above indicate also that writing performance is task dependent and that essay tests should include a variety of tasks and contexts to ensure fairness and to obtain precise evaluations of examinees' writing abilities (Cumming et al., 2000; Hamp-Lyons, 2003; Hamp-Lyons & Kroll, 1997). To enhance test fairness, it is also necessary to provide examinees with task choice because, while the findings concerning the effects of task choice on scores are mixed, several studies have shown that providing choice has a positive affective impact on examinees (e.g., Chiste & O'Shea, 1988; Jennings et al., 1999; Polio & Glew, 1996).

Practitioners should also be aware of the effects of test conditions (e.g., time constraints, use of the computer) on students' writing performance when interpreting and using test scores to make decisions about individuals and programs (Hall, 1991; Kroll, 1990; Polio, Fleck, & Leder, 1998). As Hall showed, students may employ different writing processes and produce texts that differ significantly in their quality under test and non-test conditions. Similarly, several studies have shown that the use of the computer may affect the writing processes of ESL examinees, the quality of their texts, and their essay test scores (e.g., H.K. Lee, 2004; Y. Lee, 2002; Li, 2006).

Practitioners must also guard against obvious or subtle bias in rating criteria, raters, or other contextual factors. They should, for example, be aware of and guard against the effects of construct-irrelevant factors, such as essay order, on essay scores. Furthermore, educators must be aware of the diversity of rater factors that may influence essay test scores. As pointed out above, two methods – detailed rating criteria and rater training – are often used to reduce rater-based variability in essay scores (Jacobs et al., 1981; Weigle, 2002). Educators need to keep in mind that these methods have their limitations as well. As several studies have shown (e.g., Kondo-Brown, 2002; Sweedler-Brown, 1985; Weigle, 1998), rating scales and rater training might enhance self-consistency (i.e., intra-rater reliability),

but they do not completely eliminate or reduce idiosyncratic differences (e.g. bias, severity) across raters (i.e., inter-rater reliability). Furthermore, interaction effects may reduce the effectiveness of these methods. Although this may cast doubt about the value of rater training, there is a consensus on the value of this strategy in clarifying the rating criteria and enhancing rater self-consistency (Erdosy, 2004; Lumley, 2005; Shohamy et al., 1992; Weigle, 1998, 2002). Finally, practitioners should be aware that scoring methods and rater-training procedures are also context and task dependent and that varying the assumptions, organization, and content of these procedures can affect scoring decisions significantly.

The scope of the conclusions and implications of this review is, of course, limited by the parameters and limitations of the review itself. The main limitation of this review is that it has treated all the studies equally, regardless of the quality of their designs and conclusions. In order to obtain more informative results, future reviews need to adopt more rigorous approaches (e.g., meta-analysis) and criteria for including, reviewing, and synthesizing studies. They should include and compare findings based on such criteria as the quality and appropriateness of the design of the study, whether the research was well carried out and described, whether the population and sampling procedures are carefully described, and whether the claims and conclusions made are credible and supported by the evidence presented (Christian, 2006).

Another limitation of this review is that it does not consider research on test preparation practices and the ethical, political, and social consequences of the use of essay test scores for L2 learners, teachers, and programs, as well as society. This is a growing area of interest in language testing (see Cumming, 2002; Hamp-Lyons, 1996, 2003; McNamara & Roever, 2006; Shohamy, 2001). Finally, the current review focuses only on essay tests. While this is still the most widely used method in both large-scale and classroom writing assessment, new assessment methods are now being more widely and frequently used to assess L2 writing. Examples of these alternative methods include portfolios (e.g., Hamp-Lyons & Condon, 2000; Weigle, 2002; Wilhelm, 1996) and self-assessment (e.g., Luoma & Tarnanen, 2003; Oscarson, 1997). These alternative methods integrate assessment and learning and seem more compatible with current views of writing and education as cognitive and social activities (Cumming, 2001, 2002; Hamp-Lyons & Condon, 2000). Most of the factors and issues discussed above with reference to essay tests apply to these new forms of assessment as well. However, because these approaches

involve more participants, texts, and processes, they may prove challenging to design, implement, and evaluate (see Brown & Hudson, 1998). This should not deter us from exploring such innovations, however.

Khaled Barkaoui is a PhD student at the Ontario Institute for Studies in Education of the University of Toronto. His research interests include second language assessment, second language writing, research methodology, language program evaluation, and English for academic purposes. His publications have appeared in *Assessing Writing*, *TESL Canada*, *TESL Reporter*, and *Contact*.

Contact: kbarkaoui@oise.utoronto.ca

Notes

- 1 But see Broad (2003), Huot (2002), Purves (1992a, 1992b), and Gorman, Purves, and Degenhart (1988) for research in L1 writing assessment, and Cumming (2001) for research in L2 writing assessment.
- 2 See Schoonen, Vergeer, and Eiting (1997) for similar findings in an L1 writing test.
- 3 See Huot (1993), Pula and Huot (1993), and Wolfe, Kao, and Ranney (1998) for similar findings in L1 assessment research.
- 4 As several authors have noted (e.g., Goulden, 1992; Hamp-Lyons, 1991b), there is some confusion about the terms used to describe rating scales in performance assessment. To address this confusion, we need to distinguish between *evaluation approach* and *scoring method* (Goulden, 1992; Hamp-Lyons, 1991b). There are two main evaluation approaches: *holistic evaluation approaches*, which include 'any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing,' and *analytic evaluation approaches* whereby the reader is 'required to count or tally incidents of the features' (Cooper, 1977, p. 4). Studies that analyze learners' texts in specific textual features (e.g., number of errors, use of metadiscourse markers) adopt an analytic approach (see previous section). The three types of rating scales – multiple-trait, holistic, and primary-trait – discussed in this section represent three different *scoring methods* under the holistic approach as defined by Cooper. Note, however, that several studies reviewed in this article refer to 'multiple-trait scoring' as 'analytic rating' (for detailed discussion of this issue, see Cooper, 1977; Fulcher, 2003; Goulden, 1992; Hamp-Lyons, 1991b; and Lloyd-Jones, 1977).
- 5 For studies in L1 assessment contexts see Freedman (1979), Schoonen (2005), and Swartz et al. (1999).

- 6 See Huot (1993) for a comparison of the rating processes of raters with and without holistic rubrics in an L1 essay test.

Acknowledgement

I would like to thank Alister Cumming and the two anonymous *CMLR* reviewers for their comments and suggestions on an earlier version of this manuscript.

References

- Asterisked items are the 70 empirical studies included in the review.
- *Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, 29, 371–383.
- Bachman, L.F. (1990). *Fundamental considerations in language testing*. Oxford, UK: Oxford University Press.
- *Breland, H., Lee, Y., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts: Response mode analyses* (TOEFL Research Report RR-04-23). Princeton, NJ: Educational Testing Service.
- *Breland, H., Lee, Y., Najarian, M., & Muraki, E. (2004). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups* (TOEFL Research Report RR-04-05). Princeton, NJ: Educational Testing Service.
- Broad, B. (2003). *What we really value: Rubrics in teaching and assessing writing*. Logan, UT: Utah State University Press.
- Brossell, G. (1986). Current research and unanswered questions in writing assessment. In K.L. Greenberg, H.S. Weiner & R.S. Donovan (Eds.), *Writing assessment: Issues and strategies* (pp. 168–182). New York: Longman.
- *Brown, J.D. (1991). Do English and ESL faculties rate writing samples differently?. *TESOL Quarterly*, 25, 587–603.
- Brown, J.D., & Bailey, K.M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, 34, 21–42.
- Brown, J.D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32, 653–675.
- Burstein, J., & Chodorow, M. (2002). Directions in automated essay analysis. In Kaplan Robert (Ed.), *The Oxford handbook of applied linguistics* (pp. 487–497). New York: Oxford 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). Creskill, NJ: Hampton.
- *Campbell, C. (1990). Writing with others' words: Using background reading text in academic compositions. In B. Kroll (Ed.), *Second language writing:*

Research insights for the classroom (pp. 211–230). Cambridge, UK: Cambridge University Press.

- *Carr, N. (2000). A comparison of the effects of analytic and holistic composition in the context of composition tests. *Issues in Applied Linguistics*, 11, 207–241.
- *Chiste, K.B., & O’Shea, J. (1988). Patterns of question selection and writing performance of ESL students. *TESOL Quarterly*, 22, 681–684.
- *Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater’s performance on TOEFL essays* (TOEFL Research Report RR-04-04). Princeton, NJ: Educational Testing Service.
- Christian, D. (2006). Introduction. In F. Genesee, K. Lindholm-Leary, W.M. Saunders & D. Christian (Eds.), *Educating English language learners: A synthesis of research evidence* (pp. 1–13). Cambridge, UK: Cambridge University Press.
- Clachar, A. (1999). It’s not just cognition: The effect of emotion on multiple-level discourse processing in second-language writing. *Language Sciences*, 21, 31–60.
- Coffman, W.E. (1971). Essay examinations. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed.), (pp. 271–302). Washington, DC: American Council on Education.
- *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.L. (1993). The interpretation of the tasks by writers and readers in holistically rated direct assessment of writing. In J.G. Carson & I. Leki (Eds.), *Reading in the composition classroom: Second language perspectives* (pp. 141–160). Boston, MA: Heine & Heine.
- *Connor, U.M., & Kramer, M.G. (1995). Writing from sources: Case studies of graduate students in business management. In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 155–182). Norwood, NJ: Ablex.
- Connor, U., & Mbaye, A. (2002). Discourse approaches to writing assessment. *Annual Review of Applied Linguistics*, 22, 263–278.
- *Connor-Linton, J. (1995). Crosscultural comparison of writing standards: American ESL and Japanese EFL. *World Englishes*, 14, 99–115.
- 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 language proficiency. *Language Learning*, 39, 81–141.
- *Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing*, 7, 31–51.

- Cumming, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language Testing and Assessment* (pp. 51–63). Dordrecht, Netherlands: Kluwer.
- Cumming, A. (2001). ESL/EFL instructors' practices for writing assessment: Specific purposes or general purposes? *Language Testing*, 18, 207–224.
- Cumming, A. (2002). Assessing L2 writing: Alternative constructs and ethical dilemmas. *Assessing Writing*, 8, 73–83.
- Cumming, A. (2004). Broadening, deepening, and consolidating. *Language Assessment Quarterly*, 1, 5–18.
- *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, 5–43.
- *Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67–96.
- Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL Monograph Series 18). Princeton, NJ: Educational Testing Service.
- Daly, J.A., & Dickson-Markman, F. (1982). Contrast effects in evaluating essays. *Journal of Educational Measurement*, 19, 309–316.
- Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 155–164). Norwood, NJ: Ablex.
- Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Dictionary of language testing*. Cambridge, UK: Cambridge University Press.
- DeGrujter, D.N.M. (1980). The essay examination. In L.J.T. vander Kamp, W. F. Langerak & D.N.M. Gruijter (Eds.), *Psychometrics for educational debates* (pp. 245–262). New York: Willey.
- *Delaruelle, S. (1997). Text type and rater decision-making in the writing module. In G. Brindley & G. Wigglesworth (Eds.), *Access: Issues in English language test design and delivery* (pp. 215–242). Sydney, Australia: National Center for English Language Teaching and Research, Macquarie University.
- *DeRemer, M. (1998). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing*, 5, 7–29.
- Deville, G., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C.A. Chapelle & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 9–25). Amsterdam: Benjamins.
- Diederich, P.B., French, J.W., & Carlton, S.T. (1961). *Factors in the judgment of writing quality*. Princeton, NJ: Educational Testing Service.

- *Engber, C. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4, 139–155.
- *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 RR-03-17). Princeton, NJ: Educational Testing Service.
- *Frase, L.T., Faletti, J., Ginther, A., & Grant, L. (1999). *Computer analysis of the TOEFL test of written English* (TOEFL Research Report N 64). Princeton, NJ: Educational Testing Service.
- Freedman, S.W., & Calfee, R.C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In P. Mosenthal, L. Tamor & S.A. Walmsley (Eds.), *Research on writing: Principles and methods* (pp. 75–98). New York: Longman.
- Freedman, S.W. (1979). How characteristics of student essays influence teachers' evaluation. *Journal of Educational Psychology*, 71, 328–338.
- Fulcher, G. (2003). *Testing second language speaking*. New York: Longman.
- *Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System*, 28, 31–53.
- Gorman, T.P., Purves, A.C., & Degenhart, R.E. (Eds.). (1988). *The IEA study of written composition I: The international writing tasks and scoring scales*. New York: Pergamon.
- Goulden, N.R. (1992). Theory and vocabulary for communication assessments. *Communication Education*, 41, 258–269.
- Goulden, N.R. (1994). Relationship of analytic and holistic methods to raters' scores for speeches. *Journal of Research and Development in Education*, 27, 73–82.
- *Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9, 123–145.
- Groot, P.J.M. (1990). Language testing in research and education: The need for standards. *AILA Review*, 7, 9–23.
- Hall, E. (1991). Variations in composing behaviors of academic ESL writers in test and non-test situations. *TESL Canada Journal*, 8, 9–33.
- *Hamp-Lyons, L. (1989). Raters respond to rhetoric in writing. In H.W. Dechert & M. Raupach (Eds.), *Interlingual processes* (pp. 229–244). Tübingen, Germany: Gunter Narr Verlag.
- 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, UK: Cambridge University Press.
- Hamp-Lyons, L. (1991a). Basic concepts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 1–15). Norwood, NJ: Ablex.

- Hamp-Lyons, L. (1991b). 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. (1995). Rating non-native writing: The trouble with holistic scoring. *TESOL Quarterly*, 29, 759–62.
- Hamp-Lyons, L. (1996). Applying ethical standards to portfolio assessment of writing in English as a second language. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium* (pp. 151–164). Cambridge, UK: Cambridge University Press.
- 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., & Condon, W. (2000). *Assessing the portfolio: Principles for practice, theory and research*. Cresskill, NJ: Hampton.
- Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning*, 41, 337–373.
- Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL*, 6, 52–72.
- Hamp-Lyons, L., & Kroll, B. (1997). *TOEFL 2000–writing: Composition, community and assessment*. (TOEFL Monograph Series N 5). Princeton, NJ: Educational Testing Service.
- *Hamp-Lyons, L., & Mathias, S.P. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing*, 3, 85–96.
- *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.
- Henning, G. (1987). *A guide to language testing: Development, evaluation and research*. New York: Newbury House.
- *Hill, K. (1997). Who should be the judge? The use of non-native speakers as raters on a test of English as an international language. In A. Huhta, V. Kohonen, L. Kurki-Suonio & S. Luoma (Eds.), *Current developments and alternatives in language assessment: Proceedings of LTRC 96* (pp. 275–290). Jyväskylä, Finland: University of Jyväskylä.
- *Hinkel, E. (2002). *Second language writers' text: Linguistic and rhetorical features*. Mahwah, NJ: Erlbaum.
- *Homburg, T.J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively?. *TESOL Quarterly*, 18, 87–107.

- Huot, B. (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). Creskill, NJ: Hampton.
- Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Logan, UT: Utah University Press.
- Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F., & Hughey, J.B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.
- *Janopoulos, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of Second Language Writing, 1*, 109–121.
- *Jarvis, S., Grant, L., Bikowskia, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12*, 377–403.
- *Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16*, 426–456.
- *Khalil, A. (1985). Communicative error evaluation: Native speakers' evaluation and interpretation of written errors of Arab EFL learners. *TESOL Quarterly, 19*, 335–351.
- *Kobayashi, H., & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: Cultural rhetorical pattern and readers' background. *Language Learning, 46*, 397–437.
- *Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly, 26*, 81–112.
- Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*, 3–31.
- 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. 140–54). Cambridge, UK: Cambridge University Press.
- Kroll, B. (1998). Assessing writing abilities. *Annual Review of Applied Linguistics, 18*, 219–240.
- Kroll, B., & Reid, J. (1994). Designing and assessing effective classroom writing assignments for NES and ESL students. *Journal of Second Language Writing, 4*, 17–41.
- *Kunnan, A.J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL Quarterly, 24*, 741–746.
- *Lee, H.K. (2004). A comparative study of ESL writers' performance in a paper-based and a computer-delivered writing test. *Assessing Writing, 9*, 4–26.

- *Lee, Y. (2002). A comparison of composing processes and written products in timed-essay tests across paper-and-pencil and computer modes. *Assessing Writing*, 8, 135–157.
- *Lee, Y., Breland, H., & Muraki, E. (2004). *Comparability of TOEFL CBT writing prompts for different native language groups* (TOEFL Research Report RR-04-24). Princeton, NJ: Educational Testing Service.
- *Leki, I. (1995). Good writing: I know it when I see it. In D. Belcher & G. Braine (Eds.), *Academic writing in a second language: Essays on research and pedagogy* (pp. 23–46). Norwood, NJ: Ablex.
- *Li, J. (2006). The mediation of technology in ESL writing and its implications for writing assessment. *Assessing Writing*, 11, 5–21.
- Lloyd-Jones, R. (1977). Primary trait scoring. In C.R. Cooper & L. Odell (Eds.), *Evaluating writing: Describing, measuring, judging* (pp. 33–66). Urbana, IL: NCTE.
- *Lukmani, Y. (1996). Linguistic accuracy versus coherence in assessing examination answers in content subjects. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 130–50). Cambridge, UK: Cambridge University Press.
- *Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, 19, 246–76.
- Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York: Lang.
- Luoma, S., & Tarnanen, M. (2003). Creating a self-rating instrument for second language writing: From idea to implementation. *Language Testing*, 20, 440–465.
- McNamara, T. (1996). *Measuring second language performance*. London, UK: Longman.
- McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford: Blackwell.
- *Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organization in ESL compositions. *TESL Canada Journal*, 5, 9–26.
- *Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp. 92–114). Cambridge, UK: Cambridge University Press.
- Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62, 229–258.
- Moss, P.A. (1994). Can there be validity without reliability? *Educational Researcher*, 23, 5–12.

- Moss, P.A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions. *Educational Researcher*, 25, 20–28.
- *O’Laughlin, K. (1994). The assessment of writing by English and ESL teachers. *Australian Review of Applied Linguistics*, 17, 23–44.
- Oscarson, M. (1997). Self-assessment of foreign and second language proficiency. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 175–187). Dordrecht: Kluwer.
- *Park, Y.M. (1988). Academic and ethnic background as factors affecting writing performance. In A.C. Purves (Ed.), *Writing across languages and cultures* (pp. 261–272). Beverly Hills, CA: Sage.
- *Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly*, 14, 61–69.
- 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.
- Pilliner, A.E.G. (1968). Subjective and objective testing. In A. Davies (Ed.), *Language testing symposium: A psycholinguistic approach* (pp. 19–35). Oxford, UK: Oxford University Press.
- Polio, C., Fleck, C., & Leder, N. (1998). ‘If I only had more time’: ESL learners’ changes in linguistic accuracy on essay revisions. *Journal of Second Language Writing*, 7, 43–68.
- *Polio, C., & Glew, M. (1996). ESL writing assessment prompts: How students choose. *Journal of Second Language Writing*, 5, 35–49.
- *Porter, D., & O’Sullivan, B. (1999). The effect of audience age on measured written performance. *System*, 27, 65–77.
- Pula, J.J., & Huot, B.A. (1993). A model of background influences on holistic raters. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and Empirical foundations* (pp. 237–265). Creskill, NJ: Hampton.
- Purves, A.C. (1992a). Reflections on research and assessment in written composition. *Research in the Teaching of English*, 26, 108–122.
- Purves, A.C. (Ed.), (1992b). *The IEA study of written composition II: Education and performance in fourteen countries*. New York: Pergamon.
- Raimes, A. (1987). Language proficiency, writing ability, and composing strategies: A study of ESL student writers. *Language Learning*, 37, 439–469.
- *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). Cambridge, UK: Cambridge University Press.

- *Rinnert, C., & Kobayashi, H. (2001). Differing perceptions of EFL writing among readers in Japan. *Modern Language Journal*, 85, 189–209.
- Ruth, L., & Murphy, S. (1988). *Designing writing tasks for the assessment of writing*. Norwood, NJ: Ablex.
- *Santos, T. (1988). Professors' reactions to the academic writing of non-native-speaking students. *TESOL Quarterly*, 22, 69–90.
- Sasaki, M., & Hirose, K. (1999). Development of an analytic rating scale for Japanese L1 writing. *Language Testing*, 16, 457–478.
- *Scarcella, R.C. (1984). How writers orient their readers in expository essays: A comparative study of native and non-native English writers. *TESOL Quarterly*, 18, 671–88.
- *Schneider, M., & Connor, U. (1990). Analyzing topical structure in ESL essays: Not all topics are equal. *Studies in Second Language Acquisition*, 12, 411–427.
- Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22, 1–30.
- Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers versus lay readers. *Language Testing*, 14, 157–184.
- *Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, 18, 303–325.
- Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. New York: Longman.
- *Shohamy, E., Gordon, C.M., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal*, 76, 27–33.
- *Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment: Vol. 1* (pp. 159–189). Sydney: Macquarie University Press.
- Smith, W.L. (1993). Assessing the reliability and adequacy of using holistic scoring of essays as a college composition placement technique. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 142–205). Creskill, NJ: Hampton.
- *Song, C.B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing*, 5, 163–182.
- *Spaan, M. (1993). The effect of prompt on essay examinations. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 98–122). Alexandria, VA: TESOL.

- Swain, M. (1993). Second language testing and second language acquisition: Is there a conflict with traditional psychometrics? *Language Testing*, 10, 193–207.
- Swartz, C.W., Hooper, S.R., Montgomery, J.W., Wakely, M.B., De Kruif, R.E.L., Reed, M., Brown, T.T., Levine, M.D., & White, K.P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement*, 59, 492–506.
- Sweedler-Brown, C.O. (1985). The influence of training and experience on holistic essay evaluation. *English Journal*, 74, 49–55.
- *Sweedler-Brown, C.O. (1993a). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing*, 2, 3–17.
- *Sweedler-Brown, C.O. (1993b). The effects of ESL errors on holistic scores assigned by English composition faculty. *College ESL*, 3, 53–69.
- *Tedick, D.J. (1990). ESL writing assessment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes*, 9, 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.
- Turner, C., & Upshur, J. (1996). Developing rating scales for the assessment of second language performance. In G. Wigglesworth & C. Elder (Eds.), *The language testing cycle: From inception to washback* (pp. 55–79). Melbourne: Australian Review of Applied Linguistics.
- Turner, C.E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *The Canadian Modern Language Review*, 56, 555–584.
- Underhill, N. (1982). The great reliability validity trade-off: Problems in assessing the productive skills. In B. Heaton (Ed.), *Language testing* (pp. 17–23). Hayes, UK: Modern English Publications.
- Upshur, J.A., & Turner, C.E. (1999). Systematic effects in the rating of second-language speaking ability: Test method and learner discourse. *Language Testing*, 16, 82–111.
- *Vann, R.J., Lorenz, F.O., & Meyer, D.M. (1991). Error gravity: Faculty response to errors in the written discourse of nonnative speakers of English. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 181–195). Norwood, NJ: Ablex.
- *Vann, R.J., Meyer, D.M., & Lorenz, F.O. (1984). Error gravity: A study of faculty opinions of ESL errors. *TESOL Quarterly*, 18, 427–440.

- *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.
- Weigle, S.C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263–87.
- *Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*, 145–178.
- Weigle, S.C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
- White, E.M. (1993). Holistic scoring: Past triumphs, future challenges. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 79–108). Creskill, NJ: Hampton.
- Wilhelm, K.H. (1996). Combined assessment model for EAP writing workshop: Portfolio decision-making, criterion-referenced grading, and contract negotiation. *TESL Canada Journal, 14*, 21–33.
- Williamson, M.M. (1993). An introduction to holistic scoring: Historical and theoretical context for writing assessment. In M.M. Williamson & B.A. Huot (Eds.), *Validating holistic scoring for writing assessment: Theoretical and empirical foundations* (pp. 1–43). Creskill, NJ: Hampton.
- Wolfe, E.W., Kao, C., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication, 15*, 465–492.
- *Wolfe, E.W., & Manalo, J.R. (2005). *An investigation of the impact of composition medium on the quality of TOEFL writing scores* (TOEFL Research Report RR-04-29). Princeton, NJ: Educational Testing Service.
- Wood, R. (1991). *Assessment and testing: A survey of research commissioned by the University of Cambridge Local Examination Syndicate*. Cambridge, UK: Cambridge University Press.
- *Yeh, S.S. (1998). Validation of a scheme for assessing argumentative writing of middle school students. *Assessing Writing, 5*, 123–150.
- *Zhang, S. (1987). Cognitive complexity and written production in English as a second language. *Language Learning, 37*, 469–481.

Appendix 1

Journals searched

Annual Review of Applied Linguistics
Assessing Writing
Canadian Modern Language Review
College ESL

APPENDIX 2
List of studies reviewed

Writer	Rater	Task	Rating scale	Essay
Breland, Lee, & Muraki (2004)	Brown (1991)	Breland et al. (2004)	Bacha (2001)	Bacha (2001)
Connor & Kramer (1995)	Connor-Linton (1995)	Campbell (1990)	Carr (2000)	Breland, Lee & Muraki (2004)
Frase et al. (1999)	Cumming (1990)	Chiste & O'Shea (1988)	Lee, H. K. (2004)	Chodorow & Burststein (2004)
Hinkel (2002)	Cumming et al. (2002)	Connor & Carrell (1993)	O'Laughlin (1994)	Connor (1991)
Kunnan (1990)	Delaruelle (1997)	Cumming et al. (2002)	Song & Caruso (1996)	Cumming (1990)
Lee, Breland, & Muraki (2004)	DeRemer (1998)	Cumming et al. (2005)		Cumming et al. (2002)
Park (1988)	Erdosy (2004)	Hamp-Lyons & Mathias (1994)		Cumming et al. (2005)
Pollo & Glew (1996)	Hamp-Lyons & Zhang (2001)	Hinkel (2002)		Delaruelle (1997)
Reid (1990)	Hill (1997)	Jennings et al. (1999)		Engber (1995)
Scarcella (1984)	Kobayashi & Rinnert (1996)	Lee, H.K. (2004)		Frase et al. (1999)
	Kobayashi (1992)	Lee, Y. (2002)		Gamaroff (2000)
	Leki (1995)	Li (2006)		Grant & Ginther (2000)
	Lukmani (1996)	Park (1988)		Hamp-Lyons (1989)
	Mendelsohn & Cumming (1987)	Pollo & Glew (1996)		Hamp-Lyons & Zhang (2001)
	O'Laughlin (1994)	Porter & O'Sullivan (1999)		Homburg (1984)
	Rinnert & Kobayashi (2001)	Reid (1990)		Janopoulos (1992)
	Santos (1988)	Spaan (1993)		Jarvis et al. (2003)
	Shi (2001)	Tedick (1990)		Khalil (1985)
	Shohamy, Gordon, & Kraemer (1992)	Weigle (1999)		Kobayashi & Rinnert (1996)

(Continued)

APPENDIX 2
Continued.

Writer	Rater	Task	Rating scale	Essay
	Song & Caruso (1996)	Wolfe & Manalo (2005)		Lee, H. K. (2004)
	Vann et al. (1984, 1991)	Zhang (1987)		Leki (1995)
	Weigle (1999)			Lukmani (1996)
				Lumley (2002, 2005)
				Mendelsohn & Cumming (1987)
				Miljanovic et al. (1996)
				O'Laughlin (1994)
				Perkins (1980)
				Rinnert & Kobayashi (2001)
				Santos (1988)
				Schneider & Connor (1990)
				Smith (2000)
				Song & Caruso (1996)
				Sweedler-Brown (1993a, 1993b)
				Tedick & Mathison (1995)
				Van et al. (1984, 1991)
				Vaughan (1991)
				Weigle (1999)
				Wolfe & Manalo (2005)
				Yeh (1998)