Samuelson, B. L., Crawford, M., Dyste, S., Youngquist, J., Boehm, D., & Vellenga, H. (2007). Large-scale second-language writing assessment: What's involved? In N. Caplan & C. Pearson (Eds.), *Proceedings of the 2006 Michigan Teachers of English to Speaker of Other Languages (MITESOL) Conference* (pp. 86-97). East Lansing: Michigan State University.

Large-Scale Second Language Writing Assessment: What's Involved?

Beth Lewis Samuelson
Susan Dyste
Mary Ann Crawford
Heidi Vellenga
Judy Youngquist
Diane Boehm

Abstract

This review addresses three areas of significance in the design, implementation and use of large-scale writing tests that impact ESL writers: writing assessment tasks, rater decision-making, and test washback. First, we discuss integrated writing tasks, as exemplified by the TOEFL iBT, in contrast with impromptu or independent writing tasks. Then we examine results from recent think-aloud studies of raters and their decision-making while reading essays. Finally, we review the current research on test washback in several international contexts. Additionally, we offer some research directions and practical considerations, particularly the need to educate teachers and other test users about the needs of ESL writers.

Introduction

This article is the product of an ongoing systematic literature review (Kennedy, 2007) focusing on large-scale assessment of academic second language (L2) writing, particularly for first-year college placement purposes. Although our full study focuses on several issues in second language writing assessment, we devote our attention here to three areas: the development and use of writing tasks, the processes of rater decision-making, and the impact of test washback, or the influence that a test exerts on curriculum, teaching and learning. Given the growing population of international (Institute of International Education, 2007) and U.S.-resident L2 writers (National Clearinghouse on English Language Acquisition, 2006), there is a pressing need for policy-oriented discussion of the current knowledge-base.

Background on Second Language Writing Assessment

The field of second language writing, of which second language writing assessment is a subset, has been growing rapidly. A position statement (Conference on College Composition and Communication, 2001) and articles on second language writing have appeared in major journals and handbooks (Hedgcock, 2005; Leki, 2002; Matsuda, 2006; Silva & Brice, 2004; Silva & Leki, 2004). Just ten years ago, a review of second language writing research identified three major areas of concern: the qualities of L2 texts, the composing processes of second language writers, and the impact of sociocultural contexts (Hamp-Lyons & Kroll, 1996). Much of the resulting research has addressed these areas (e.g. Cumming, 2001; Leki, 1991, 1995; Leki, Cumming, and Silva, 2006), with relatively little attention given to assessment issues.

There is a growing need for more attention to the assessment of second language writing, not only from L2, but also from mainstream and L1 writing assessment researchers (Cumming,

1997). So far, relatively little has been done. For example, the CCCC Statement on Second Language Writing and Writers (2001) includes only a brief paragraph on assessment, offering advice on culturally-sensitive writing prompts and scoring considerations. The National Council of Teachers of English (NCTE) published a leading collection on writing assessment (Cooper & Odell, 1999), which included just two articles addressing the needs of Chinese- (Cai, 1999) and Spanish-speaking (Valdes & Sanders, 1999) second language writers respectively.

This review addresses three areas of second-language writing assessment research where significant work has occurred recently.

Writing Tasks as Assessment Measures

Writing tasks are usually the first to be evaluated as an assessment variable, in part because they are experienced by both teachers and students as a major aspect of the test. We will consider two types of writing assessment tasks: impromptu or independent tasks, which are sometimes called snapshot tasks, and integrated writing tasks, which reflect the priorities of academic literacy and writing-within-wider-academic-competencies approaches (Hamp-Lyons & Kroll, 2001).

Impromptu or Independent Tasks

Snapshot approaches (Hamp-Lyons & Kroll, 2001) involve just what the metaphor suggests: a quick still-image of a student's writing output taken in timed and/or impromptu essay tests of varying lengths. Snapshot writing tasks are easy to administer and are quickly scored using rubrics or scoring guides. As a result, they are favored as placement tools. In contrast, alternative test approaches such as writing portfolios provide a "video" image of a student's writing development over a period of time and are usually inconvenient for use as placement tests because of the time and resources needed to implement them.

Among the biggest problems with impromptu tasks is the concern that students may not be able to show their best effort in only one hastily-written essay. Differences in the cultural and background knowledge of the test-takers and the raters are recurring issues as well (Basham & Kwachka, 1991; Basham, Ray, & Whalley, 1993). Cultural differences in judging the quality of writing can undermine the validity of snapshot approaches (Connor-Linton, 1995), and snapshot prompts often lack clear relationships to the variety of genres, content, and tasks that students will encounter in academic environments. The growth of World Englishes with their culture-specific language norms and rhetorical patterns adds to this complexity (Hamp-Lyons & Zhang, 2001). These issues are briefly addressed in the CCCC Position Statement on Second language Writing and Writers:

> Writing prompts for placement and exit exams should avoid cultural references that are not readily understood by people who come from various cultural backgrounds. To reduce the risk of evaluating students on the basis of their cultural knowledge rather than their writing proficiency, students should be given several writing prompts to choose from when appropriate. (Conference on College Composition and Communication, 2001, p. 671)

While offering choices would seem to be a positive approach to testing, it raises other questions. Which variety of prompts will promote the best possible writing? Are students prepared to make the best decision in choosing their prompt? How can we help students choose wisely? Students do appear to want a choice of prompt (Polio & Glew, 1996), but in order to address the issues these choices raise, we need to have a better understanding of how prompts influence production and how different ethnic, linguistic and cultural backgrounds can affect students' production (Jennings, Fox, Graves, & Shohamy, 1999).

Other persistent issues regarding writing tasks include the costs and benefits of having a limited time for writing and the fairness of using a single writing sample to make a decision about a student's overall writing competence. As a result of these concerns, impromptu tasks generally cannot provide the kind of authentic assessment of a student's academic literacy that test users, such as admissions officers, program administrators, and placement test designers, might want or that integrated approaches can address.

Integrated Tasks

Incorporating reading with writing is a current trend in assessing academic literacy, (e.g., Feak & Dobson, 1996). In response to concerns that the Test of Written English (TWE) (Educational Testing Service, 2006) had negative washback effects due to its "rhetorically formulaic requirements" (Cumming et al., 2005, p. 7), the Next Generation Internet-based Test of English as a Foreign Language (TOEFL iBT), which debuted in 2005, aims to integrate reading, speaking, and listening with writing and includes both integrated and independent tasks. The new Internet-based TOEFL offers tasks that try to recreate real academic tasks that students may encounter in North American universities. For instance, in an integrated task, the test-taker reads a short passage, listens to short lecture on a topic, and then has 20 minutes to write a brief response to a question that requires integration of content from both input sources (Educational Testing Service, 2006). The TOEFL iBT reflects the fact that the type of writing elicited by the TWE was only one form of writing that students needed to master in order to succeed in university courses. Because it reflects the interrelatedness of different language competencies such as discourse, grammatical, sociolinguistic, and genre, this type of test holds the promise of giving insights into international students' ability to function literately in genuine academic situations requiring advanced listening, reading, speaking and writing skills.

A verification study of the new TOEFL iBT demonstrated that the traditional independent or "snapshot" tasks in the paper-based and computer-based TOEFL gave writers opportunities to show that they could produce "extended written arguments" (Cumming et al., 2005, p. 32) that drew from their personal experience. In contrast, the integrated tasks were shown to require writers to summarize ideas that they drew from academic reading or listening activities.

Comparisons of the effects of task type (independent or integrated) examined text length, lexical sophistication (average word length and type/token ratio), syntactic complexity (number of clauses and words per T-unit), holistic rating of grammatical accuracy (1-3 ratings), quality of argument structures (claims, data, warrants, propositions, opposition, and responses to opposition), orientation to source evidence, and functional use of phrases from sources. The researchers found that while the writing produced in response to the integrated tasks tended to be

shorter overall, the writers produced longer, more complex, sentences, a greater variety of longer words, and made use of information sources other than personal experience by repeating, summarizing, and paraphrasing (Cumming et al., 2005). The independent tasks prompted writers to produce more argumentative writing that tended to rely on personal experience as a source of evidence. Both task types appeared to have no effect on the grammatical accuracy of the writing produced.

The researchers noted that the integrated tasks yielded useful information on writer proficiency. The integrated tasks required writers to demonstrate complex literacy skills, cognitive abilities, and language proficiency while also making appropriate use of sources of evidence. More proficient writers tended to summarize source material in their own words, while the less proficient tended to rely more heavily on quotations from the source material. The more proficient writers produced texts that "marked their argument structure by a variety of transition phrases in paragraphs of varying sizes" (Cumming et al., 2005, p. 30), while the less proficient writers used more formulaic patterns. Furthermore, and not unexpectedly, the more proficient writers wrote longer pieces with more and longer clauses, showed greater lexical variation, and demonstrated better grammatical accuracy and argument structure.

Students taking the TOEFL iBT must complete both independent and integrated writing tasks. Given the importance of the test for international students preparing to study in English-speaking countries, two areas of concern need to be addressed. First, because they demand advanced reading and listening skills, quick thinking, and skilled use of evidence, integrated writing tasks will be challenging for less-proficient students who have taken the TOEFL in the past. In response to this situation, academic English instruction and teaching materials are required that will help students prepare realistically. Test developers will have to carefully consider both the materials (lectures, readings) and the content of writing prompts (Cumming, Grant, Mulcahy-Ernt, & Powers, 2004) to ensure that the topics are not too specialized. The second concern is closely linked to the first. More thought needs to be given to the definition of writing that is assumed by the integrated test: what does the result of the test reflect or mean? What vision of writing does the test reflect (Cumming, 2002)? These are questions that cannot be answered by test-developers alone, but require the involvement of all stakeholders.

Raters

We chose to address rater issues because this area is another critical variable in the validity of L2 writing assessment. The CCCC Statement on Second Language Writing and Writers lists some of the issues involved:

> The scoring of second language texts should take into consideration various aspects of writing (e.g., topic development, organization, grammar, word choice), rather than focus only on one or two of these features that stand out as problematic. (Conference on College Composition and Communication, 2001, p. 671)

How and why raters make decisions while reading student essays is crucial to the validity and usability of a large-scale writing test. The think-aloud protocol—a research method in which raters talk aloud about their thoughts and decisions as they read and score papers—has served as

a useful data collection and analysis tool for studying how raters make decisions. Such studies indicate substantial problems.

Smith (2000) noted that almost 50% of raters' comments on essays were not in the rubric given to the raters. Somewhat surprisingly, the raters also struggled with the terminology in the rubric, suggesting varied conceptions about what was rated. In addition, the raters exhibited different reading styles, which may have affected their scoring decisions.

Similarly, Lumley (2002) observed that highly-trained raters use scales in different ways. To study raters' processes of evaluation, Lumley used "misfitting (i.e., unexpected, or surprising)" (p. 252) writing samples from 24 candidates and four trained raters with similar backgrounds. The four raters first read and rated 12 of the 24 samples for norming purposes, and then they rated the remaining 12 samples. They also did think-aloud protocols while reading and scoring. Lumley found that the process of rating essays, even when the rating is done by highly qualified and trained experts, can produce variable results. The more experienced raters did not usually rely on the scale they received, and they were not always influenced by descriptors on the scale. Some raters were frustrated because the features they considered important were not included on the scoring scale. Even when raters used the scale similarly, they appeared to use the descriptors in divergent ways. Lumley concluded that raters may emphasize one part of the scale and inappropriately de-emphasize another, and this tendency may conflict with the rubric training received by the raters.

Rating scales were another area of concern. Lumley (2002) also concluded that essays do not all fit neatly into categories delineated on rating scales. With some of the misfitting texts, the rating scale and descriptors were used primarily as a means of justifying the raters' scores. This suggests that the scales can do little to illuminate the constructs of writing being measured. Lumley offered a salient, if problematic, definition of a writing scale or rubric: "a set of negotiated principles that the raters use as a basis for reliable action, rather than a valid description of language performance" (p. 268). This definition suggests that a strong community-focused orientation towards rubric design and implementation should be a key aspect of writing assessment.

Cumming, Kantor and Powers (2002) used think-aloud protocols to explore the scoring decisions that raters made while scoring TOEFL essays in three different studies. In the first study, ESL/EFL-trained raters scored essays written in response to independent writing tasks. In the second study, raters trained in English L1 composition also scored independent TOEFL essays. In the third, ESL/EFL-trained raters scored TOEFL essays written in response to integrated writing tasks. Although all raters used similar criteria to judge all the writing samples, the ESL/EFL-trained raters tended to use strategies, such as commenting on whether the text was handwritten or typewritten; labeling errors by their categories (e.g., prepositions, relative clauses); mentally editing phrases to improve comprehension; and judging the accuracy of spelling, fluency, lexis, syntax, and punctuation, more heavily than the L1-trained raters. Interestingly, though, the L1-trained raters scored papers faster than the ESL/EFL-trained raters, and often applied more creative criteria to their evaluations. Both groups of raters appeared to rely on their past experience and knowledge when rating writing samples. With the writing samples produced by less proficient writers, the raters looked at specific language choices more

than at rhetorical features. The think-aloud protocols indicated that the raters were also interested in what had been communicated to the test-takers with regard to assessment criteria. Along with knowing how/what to evaluate in the writing, the raters seemed to want to hold the writers accountable for communicating and complying with directions. Based on such findings, the researchers suggested that test-takers should be given clear descriptions about how their writing samples will be assessed. Further, raters felt better informed about test-takers' actual writing abilities when they could assess multiple texts by the same writer.

Handwriting is one additional feature that warrants attention when considering rater factors. Although handwriting is a surface-level concern that has little to do with the qualities of writing typically described in rubrics, it has been shown to affect scoring decisions (Vaughn, 1991). In fact, handwriting ranks as the second most common reason for point deductions, supporting the need for computer-based writing, although differences in students' keyboarding abilities and familiarity with technology may still be a problem.

Test Washback

The existence of washback can be traced as far back as ancient China (ca. 200 B.C.), when the imperial government selected new bureaucrats through the results of arduous written exams (Gipps, 1999). These bureaucratic tests influenced how students learned reading and writing in China for centuries. Washback, or the influence of tests on pedagogy and classroom assessment practices, is an increasingly important topic in both first-language (Hillocks, 2002; Huot, 2002a) and second language writing assessment (Shohamy, 1993, 1996, 1998, 2001a, 2001b; Silva, 1997). So far, there has been little research that has focused specifically on washback of large-scale L2 writing assessment, despite the fact that growing numbers of L2 writers are encountering tests that have far-reaching influences on their classrooms, their teachers, and their school administrators, not to mention their families, communities and selves.

Washback is a consequential aspect of the construct validity of the test (Messick, 1996). Consequential validity connects the vision of writing to be assessed to the values and policies of the test-developers and test-users. These values and policies have an impact on the ways that writing is taught in classrooms where students and teachers are preparing for the test.

Washback is usually considered as negative or positive. Negative washback may include teachers "teaching to the test" and neglecting material not covered in the test. The curriculum may become very restricted because the test sets the standard for the outcomes of learning and teaching. Hillocks (2002) and Shohamy (1993) describe several situations in which the pressure of standardized examinations was linked to decreased student motivation and restricted curricula. Hillocks describes the impact of L1 writing tests in U.S. K-12 contexts; Shohamy reports on educational contexts in Israel.

Positive washback can occur when teaching objectives and/or shared values are promoted in the test, when a test can be used for teaching and learning, and when curriculum can be improved as a result of feedback from the test (Weigle, 2004). Efforts to promote positive washback have been undertaken in EFL contexts. In 1998, Turkey, in order to raise standards of English proficiency, studied the washback of large-scale writing assessment on students, teachers

and curriculum, which resulted in major changes to course syllabi and textbooks (Alderson & Wall, 1993). A study of the washback of a reading and writing test in Sri Lanka resulted in improvement in the content of English lessons as well as a change in the design of in-class tests (Wall & Alderson, 1993).

Planning for positive washback from a test is no guarantee that it will be realized. In the People's Republic of China, the National Matriculation English Test (NMET) was designed with the specific goal of encouraging changes in the instruction of English in secondary schools (Cheng, 2006). In a study examining the reasons why the NMET did not achieve this goal, Qi (2005) found that the test's high-stakes function as a selection instrument for admission to university studies was in direct conflict with its goal of promoting instructional change.

A washback study of the Hong Kong Advanced Supplementary "Use of English" oral examination (Hong Kong Examinations and Assessment Authority, 2006a), however, showed that teachers' attitudes and actions were improved, even though no effect on teaching methodology was noted (Andrews, Fullilove, & Wong, 2002). Also in Hong Kong, a washback study carried out in Hong Kong secondary schools in 1999-2000 suggested that testing was an ineffective way to positively influence teaching approaches and found that teachers encouraged students to take a more active role, which created positive washback (Cheng, 1999; Cheng, Watanabe, & Curtis, 2004). Another more learner-centered Hong Kong study used questionnaires, interviews, and classroom observations to calculate the effect of the Hong Kong Certificate of Education Examination in English (Hong Kong Examinations and Assessment Authority, 2006b) in Hong Kong secondary schools (Cheng, 1997).

One of the perennial problems with studying test washback, whether positive or negative, is the difficulty of linking a specific test with teacher and student actions, curriculum choices and student outcomes. Despite a general understanding that assessment drives pedagogy (e.g., Conference on College Composition and Communication, 1995), it is extremely difficult to prove direct cause and direct effect. How can the effect of large-scale writing tests be defined when many other variables, including students, teachers, curriculum, institutional support and expectations, come in to play? One proposal suggests looking only at evidence showing how "the test influences language teachers and learners to do things that they would not necessarily otherwise do" (Alderson & Wall, 1993, p. 117), yet what are these "things"? And how can they be observed?

Researchers face challenging obstacles to demonstrating how, why, and to which extent washback occurs. Watanabe (2000, 2004) proposed five features to examine when evaluating test washback: specificity, intensity, length, intentionality and value. When evaluating the specificity of a test, test-users should look at whether the test focuses on specific learning strategies and content that might be adopted for emphasis in the curriculum. Indications of the intensity of test washback can be derived from the degree to which certain areas and certain test content produce washback effects. Test-users should also inquire into the length of the washback effect: how long the washback of a specific test can be expected to last. When evaluating for intentionality, test-users should focus on both the planned and unplanned effects of washback. Finally, test-users should scrutinize both positive and negative washback of test, being careful not to overlook any value of the washback effects.

If positive washback is to occur, test designers need to proactively involve stakeholders in determining educational outcomes and achievement; institutional and classroom cultures must be already supportive of teaching and learning; and teachers and administrators must be active collaborators in the process of test materials development. Known factors promoting positive impact on instruction as well as shared values should be incorporated into the test design as much as possible. In order for this to occur, stakeholders and the educational community need to first determine what teaching and learning outcomes they value and then select the factors that reflect these values. For example, test-takers may be asked to engage in writing after reading one or two sources or after listening to a short lecture (Saif, 2006).

Conclusion

In each of the three areas of second language writing assessment reviewed here, a recurring theme is the need for more involvement by all test stakeholders—test-takers, test-users, test-developers, teachers and parents—in determining the purposes and the design of tests and the impact that they should have on curriculum, on teaching, and on students' lives.

Lumley (2002) expressed this theme most directly when he suggested that a strong community-focused orientation towards rubric design and implementation should be a key aspect of writing assessment. Similarly, positive washback appears to be contingent on active involvement by stakeholders. In her discussion of critical testing theory, Shohamy (1998) emphasized that language tests are to be seen as "deeply embedded in cultural, educational, and political arenas where different ideological and social forms struggle for dominance." She encourages test-developers and other stakeholders to "ask themselves what sort of vision of society language tests create and what vision of society tests serve" (p. 332).

From a research perspective, one of the major implications so far of our review has been the need for more ethnographic studies of language-testing contexts, including discourse of discussions that occur as students and raters prepare for their respective roles in the test (e.g., Samuelson, 2005). Additionally, more discourse approaches such as those reviewed by Connor & Mbaye (2002) need to be explored. Integrating discourse analysis with ethnographic procedures may help to identify grammatical, sociolinguistic, discursive, and strategic competencies that will enable evaluation of more than the linguistic criteria of student texts.

From a practical perspective, one of the major recurring concerns is the need to educate teachers and other test users about the needs of ESL writers. ESL writers are by no means a uni-dimensional group, and educators need to be prepared to identify the specific contextual, cultural, and instructional needs of the students. In the case of large-scale tests, this means that educators must be willing to become educated consumers and users of tests so they can be advocates for their ESL students.

Interestingly, the focus on community involvement that we have perceived in our review of L2 writing assessment is exactly the direction that L1 writing assessment appears to be taking, at least in some quarters, with efforts to develop a constructivist paradigm for writing assessment. Some examples include exploring communal writing assessment, in which participants map the criteria by which they will evaluate writing in their program (Broad, 2000,

2003), and exploring discursive connections between teaching and evaluating writing (Huot, 1996, 2002b). While the needs of L2 writers remain distinct from those of L1 writers in many ways, these directions can suggest fruitful and valuable research directions for second language writing assessment.

Author Note

Beth Lewis Samuelson, Assistant Professor of English, Central Michigan University. Susan Dyste, ESL Instructor, English Language Institute, Central Michigan University. Mary Ann Crawford, Director of Basic Writing/Writing Center and Writing Across the Curriculum Programs, Associate Professor of English, Central Michigan University. Heidi Vellenga, Director, English Language Program, Saginaw Valley State University. Judy Youngquist, ESL Specialist, English Language Program, Saginaw Valley State University. Diane Boehm, Director, University Writing Program, Saginaw Valley State University. The authors gratefully acknowledge the Conference on College Composition and Communication for a Research Initiative grant enabling us to undertake this study. Thanks also to Beini Hou, for her efficient and dedicated work as a graduate research assistant on this project. Correspondence concerning this article should be addressed to Beth Lewis Samuelson (beth.samuelson@cmich.edu).

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(4), 115-129.

Andrews, S., Fullilove, J., Wong, Y. (2002). Targeting washback--A case study. *System, 30*(2), 207-223.

Basham, C., & Kwachka, P. (1991). Reading the world differently: A cross-cultural approach to writing assessment. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 37-50). Norwood, NJ: Ablex.

Basham, C., Ray, R., & Whalley, E. (1993). Cross-cultural perspectives on task representation in reading to write. In J. G. Carson & I. Leki (Eds.), *Reading in the composition classroom* (pp. 299-314). Boston, MA: Heinle & Heinle.

Broad, B. (2000). Pulling your hair out: Crises of standardization in communal writing assessment. *Research in the Teaching of English, 35*(2), 213-260.

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

Cai, G. (1999). Texts in contexts: Understanding Chinese students' English compositions. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: The role of teachers' knowledge about text, learning, and culture* (pp. 279-298). Urbana, IL: National Council of Teachers of English.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38-54.

Cheng, L. (1999). Changing assessment: Washback on teacher perceptions and actions. *Teaching and Teacher Education, 15*(3), 253-271.

Cheng, L. (2006). Description and examination of the National Matriculation English Test. *Language Assessment Quarterly, 3*(1), 53-70.

Cheng, L., Watanabe, Y., & Curtis, A. (Eds). (2004). *Washback in language testing: Research contexts and methods*. Mahwah, NJ: Erlbaum.

Conference on College Composition and Communication. (1995). Writing assessment: A position statement. *College Composition and Communication, 46*(3), 430-437.

Conference on College Composition and Communication. (2001). CCCC statement on second language writing and writers. *College Composition and Communication, 52*(4), 669-674.

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*(1), 99-115.

Cooper, C. R., & Odell, L. (Eds.). (1999). *Evaluating writing: Describing, measuring, judging* (2nd ed.). Urbana, IL: National Council of Teachers of English.

Cumming, A. (1997). The testing of writing in second languages. In C. Clapham (Ed.), *Encyclopedia of language and education* (Vol. 7). Boston, MA: Kluwer.

Cumming, A. (2001). Learning to write in a second language: Two decades of research. *International Journal of English Studies, 1*(2), 1-23.

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*(2), 107-145.

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

Educational Testing Service. (2006). TOEFL--Test of English as a Foreign Language. Retrieved December 26, 2006 from http://www.toefl.org.

Feak, C., & Dobson, B. (1996). Building on the impromptu: A source-based academic writing assessment. *College ESL, 6*(1), 73-84.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*, 355-392.

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

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: Erlbaum.

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). New York: Cambridge University Press.

Hedgcock, J. (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: Erlbaum.

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

Hong Kong Examination and Assessment Authority. (2006a). HKALE Exam Syllabuses: Use of English. Retrieved May 23, 2007 from http://eant01.hkea.edu.hk/hkea/new_look_home.asp.

Hong Kong Examination and Assessment Authority. (2006b). HKCEE Exam Syllabuses: English Language. Retrieved May 23, 2007 from http://eant01.hkea.edu.hk/hkea/new_look_home.asp.

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

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

Huot, B. (2002b). Toward a new discourse of writing assessment for the college writing classroom. *College English, 65*(2), 163-180.

Institute of International Education. (2007). Open doors 2006: Report on international educational exchange. New York: The Institute of International Education, Inc.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-taker's choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426-456.

Kennedy, M. (2007). Defining a literature. *Educational Researcher, 36*(3), 139-147.

Leki, I. (1991). Twenty-five years of contrastive rhetoric: Test analysis and writing pedagogies. *TESOL Quarterly, 25*, 123-143.

Leki, I. (1995). Good writing: I know it when I see it. In Belcher, D., & G. Braine (Eds.). *Academic writing in a second language: Essays on research and pedagogy* (pp. 23-46). Norwood, NJ: Ablex.

Leki, I. (2002). Second language writing. In Kaplan, R. (Ed.). *Oxford handbook of applied linguistics* (pp. 60-69). New York: Oxford University Press.

Leki, I., Cumming, A. & Silva, T. (2006). Second language composition teaching and learning. In Smagorinsky, P. (Ed.). *Research on composition: Multiple perspectives on two decades of change* (pp. 141-169). New York: Teachers College Press.

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.

Matsuda, P. (2006). The myth of linguistic homogeneity in U.S. college composition. *College English, 68*(6), 636-651.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-257.

National Clearinghouse on English Language Acquisition. (2006). The growing numbers of Limited English Proficient students, 1994/95 - 2004/05. Retrieved June 1, 2007, from http://www.ncela.gwu.edu/policy/states/reports/statedata/2004LEP/GrowingLEP_0405_Nov06.pdf.

Polio, C., & Glew, M. (1996). ESL writing assessment prompts: How students choose. *Journal of Second Language Writing, 5*(1), 35-49.

Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142-173.

Saif, S. (2006). Aiming for positive washback: A case study of international teaching assistants. *Language Testing, 23*(1), 1-34.

Samuelson, B. (2005). Talk about writing: Mediating knowledge about writing through discussions of student work. (Doctoral Dissertation, University of California, Berkeley, 2004). *Dissertation Abstracts International*, 66/02, 480A.

Shohamy, E. (1993). *The power of tests: The impact of language tests on teaching and learning*. Washington, DC: The National Foreign Language Center at Johns Hopkins University.

Shohamy, E. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*, 298-317.

Shohamy, E. (1998). Critical language testing and beyond. *Studies in Educational Evaluation, 24*(4), 331-345.

Shohamy, E. (2001a). Democratic testing as an alternative. *Language Testing, 18*(4), 373-391.

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

Silva, T. (1997). On the ethical treatment of ESL writers. *TESOL Quarterly, 31*, 359-363.

Silva, T., & Brice, C. (2004). Research in the teaching of writing. *Annual Review of Applied Linguistics, 24*, 70-106.

Silva, T., & Leki, I. (2004). Family matters: The influence of applied linguistics and composition studies on second language writing studies--past, present, and future. *The Modern Language Journal, 88*(1), 1-13.

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In Brindley, G. (Ed.). *Studies in immigrant English language assessment* (pp. 159-189). Sydney, Australia: Macquarie University.

Valdes, G., & Sanders, P. A. (1999). Latino ESL students and the development of writing. In C. R. Cooper & L. Odell (Eds.), *Evaluating writing: The role of teachers' knowledge about text, learning, and culture* (2nd ed., pp. 249-278). Urbana, IL: National Council of Teachers of English.

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

Wall, D. & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41-69.

Watanabe, Y. (2000). Washback effects of the English section of Japanese university entrance examinations on instruction in pre-college level EFL. *Language Testing, 27*, 42-47.

Watanabe, Y. (2004). Methodology in washback studies. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.). *Washback in language testing: Research contexts and methods* (pp. 19-36). Mahwah, NJ: Erlbaum.

Weigle, S. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27-55.