An Updated Visual Representation for Writing Assessment Research

Beverly Baker

This article offers a suggestion for a visual representation of writing assessment and its related research enterprise. In today's world of visual data and multiliteracies, representation through images is recognized as a powerful tool of inquiry that can transform our understandings and our research practices (Sanders-Bustle, 2003). Although previous visual representations in the field of language assessment have been useful, they have been predicated on two assumptions: that all writing assessment consists of formal tests, and that assessment processes are linear rather than recursive. This new proposed visual representation has potential to aid in reconceptualizing the interrelated elements of writing assessment, as well as revealing new relationships among elements to explore. Such awareness can benefit anyone involved in the writing assessment enterprise, from writing assessment scholars to classroom teachers of writing.

Cet article offre une suggestion de représentation visuelle de l'évaluation de l'écrit et de la recherche qui en découle. Dans notre monde de données visuelles et de littératies multiples, la représentation par les images est reconnue comme un outil d'enquête puissant qui peut transformer notre connaissance et nos pratiques de recherche (Sanders-Bustle, 2003). Si les représentations visuelles antérieures dans le domaine de l'évaluation des compétences linguistiques ont été utiles, elles ont aussi été basées sur deux hypothèses : toute évaluation de l'écrit consiste en des épreuves formelles et les processus d'évaluation sont linéaires plutôt que récursifs. Cette nouvelle représentation visuelle a le potentiel d'appuyer la reconceptualisation des éléments interreliés de l'évaluation de l'écrit et d'exposer de nouveaux rapports entre les composantes de l'écriture. Toutes les personnes impliquées dans l'évaluation de l'écrit pourraient en profiter, qu'elles soient des professeurs en évaluation ou des enseignants de l'écriture.

The Benefits of Visual Representation

The following is a suggestion for a visual representation of the elements of writing assessment. This visual representation is intended to be a tool for researchers, as well as for future and current teachers of writing of all levels, to aid in recognizing the complexity of the process we are involved in. Representation through images is recognized as a powerful research tool, capable of transforming our understandings of complex phenomena (Sanders-Bustle, 2003). From the burgeoning field of information visualization (see Card,

Mackinlay, & Schneiderman, 1999) we are gaining insight into the positive cognitive effects of being presented with information in a graphic format—visual displays of complex systems and data reveal relationships that would not be salient with textual presentation alone.

Researchers in the social sciences have long recognized the value of creating visual representations of their research domains. For example, scholars working in the area of grounded theory (Glaser & Strauss, 1967; Strauss & Corbin, 1998) discuss the creation of a "theoretical explanatory scheme" (Strauss & Corbin, 1998, p. 11) of an ongoing research enterprise, consisting of conceptually organized categories and of all the relationships and interactions among them. The grounded theorist Adele Clarke (2003) discusses the benefit of representing such a scheme visually: she recommends the use of graphic "situational maps" that "lay out the major human, non-human, discursive, and other elements in the research situation of concern and provoke analyses of the relations among them" (p. 559).

Visualization in Writing Assessment

Several visual schemes are available that relate specifically to writing assessment. Notable examples include Shaw and Weir (2007; adapted from Weir, 2005) and Weigle (2002), who based her depiction on McNamara (1996).

Weigle (2002), who discusses "factors in writing assessment," outlines a scheme in which the candidate (the writer or the person being assessed) works through a particular instrument to produce a performance, and a rater examines the performance with the use of a rating scale to arrive at a score (see Figure 1). The context, pictured around the outside of the scheme, could be interpreted as affecting all the other factors.

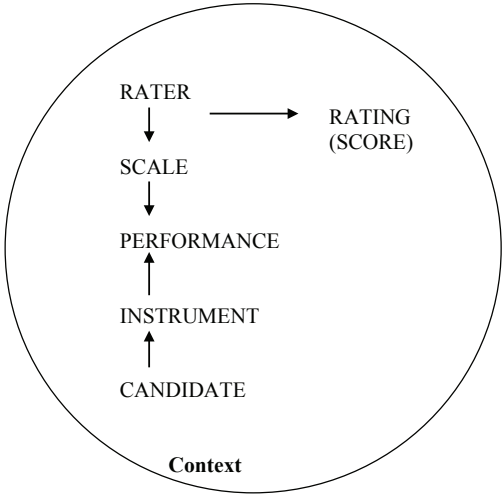

Figure 1. Factors in writing assessment (Weigle, 2002, p. 61).

Shaw and Weir (2007) present a visual comprehensive sociocognitive framework for conceptualizing writing test performance in a "temporal frame" (p. 3); that is, relating the collection of validity evidence to each stage of test development, administration, and beyond to test score use and test consequences.

A Suggestion for a New Visual Representation

The visual representation or scheme presented below is proposed as a useful complement to previous schemes for several reasons. First, it can be applied to all types of writing assessment, not only formal tests. Assessment in writing is widely acknowledged to take many forms, such as classroom-based formative feedback (provided in writing or orally, through conferencing for example). This proposed scheme also resists a linear temporal conceptualization of the assessment process. This may initially seem counterintuitive: After all, assessments must be created and then provided to candidates before scores or other judgements are made—an intuitively chronological series of steps. However, a great deal can potentially be gained by adopting a visual representation that allows for a more holistic conceptualization—less a linear process and more a complex adaptive system, where all elements conceivably have direct relationships with all other elements, and assessment activities are more recursive than linear. If one takes the writing classroom as an example, a linear view would suggest that a decision to make a test comes first, followed by the use of scores to report writing proficiency. However, the decision to create a test might come from outside the classroom—in response to a policy or a societal requirement to verify language competence in a given population. This means that the social context comes first, with the future use of scores perhaps even informing the content of the test itself. Relationships among elements are also bidirectional: researchers examine the effects of student characteristics (such as motivation) on test scores, for example, but it is equally interesting and productive to imagine the effects of test scores on levels of motivation.

This proposed representation does not assume any particular research orientation. It is intended to be democratic in that it can be appropriate for researchers operating from a classical testing perspective, who view all elements besides the score as sources of bias in arriving at the score, as well as for those who take a "construct-based" view (Bachman, 2002; Norris, 2002); that is, where the focus of research is not the score but how information taken from an assessment informs a theoretical construct of language competence. This scheme would also be compatible with writing assessment research that takes a more critical perspective, such as work with the social contexts of assessments, assessment consequences, and other issues of test or assessment use (see McNamara & Roever, 2006).

This proposed visual representation is presented as Figure 2, followed by explanations of the names of the elements and a justification for their organization.

The Elements of the Representation

The choice of terms to describe the elements of this representation is open to debate, but a justification for their selection, in no particular order, is provided below.

The term *performance* refers very broadly to the written output of a candidate in response to assessment requirements (*task demands*). This term includes rhetorical, discourse, and linguistic characteristics of the text produced as well as the content knowledge represented in the text.

The term *task demands* refers to the requirements of the performance to be produced by the candidate during the assessment (given as a written prompt or oral instructions by a teacher, either formally or informally). These include linguistic demands, rhetorical and discourse demands, requirements related to the format and mode of the performance, and demands related to the assessment situation (such as time limits or use of resources). The word *task*

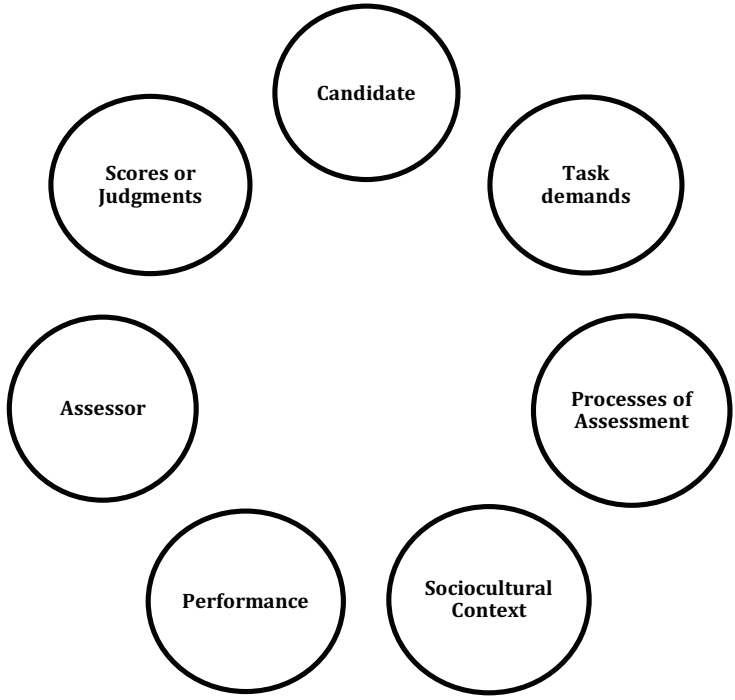

Figure 2. The proposed scheme to visually represent the elements of writing assessment.

here is used following the broad definition of task-based assessment of Norris (2002): "a confluence of ideas, concerns, and recommendations that address the validity of interpretations and actions based on certain uses for language performance assessment" (p. 338).

The term *assessor* was chosen to be more inclusive than *rater*, which seems to imply that a rating scale or rating guidelines are being used in attributing a score (which is not applicable to all writing assessment). In the writing classroom, the writing instructor is usually also the assessor. This element includes the background characteristics of people assessing the performance (such as gender, age, cognitive characteristics, experience, and disciplinary background) as well as assessor expectations.

The term *candidate* is used for the writer or the person being assessed (following Weigle, 2002), rather than *test taker,* to acknowledge that assessment includes more than tests. This categorization includes consideration of the "attributes of the individual" (Bachman & Palmer, 2010), such as age, gender, and native language, as well as content and strategic knowledge and individual cognitive characteristics. The term also encompasses candidates' perceptions about how they are expected to write and how their texts will be interpreted.

The *assessment process* includes all activities related to score attribution, such as the rating scale or guidelines that test developers or assessors have created; practical conditions under which the texts are rated (such as time constraints, location, etc.); and other elements of the scoring process such as rater training and score adjudication procedures. In addition, this element includes processes related to assessment not involving scores or formal tests, such as decisions by writing teachers on the nature of formative feedback to be provided to students.

Score or judgement is an element that represents the tangible result of the assessment process—whether it be a score, grade, comment, or observation—as well as any decisions leading from these results. This element therefore includes intended and unintended consequences of an assessment on all stakeholders, and the perceptions of the assessment by wider society. Included here could also be the effects of assessment on teaching and learning (washback). This factor includes all users of test scores not directly involved in the assessment process, such as policy makers, program administrators, and other decision makers, as well as school officials and parents.

The *sociocultural context* for assessment has to do with factors such as the impetus for and goals of an assessment, the societal values embedded within an assessment, and the perceptions of the assessment by all stakeholders as well as the wider society. Test developers—some of whom may consider themselves to be apart from the sociocultural context of a test—are not neutral entities. As they bring their own perspective and values to their work, so should they be included here.

Each of these elements is complex enough to be the subject of extensive empirical inquiry—and indeed has been. In addition, each element can conceivably be connected to any others with bidirectional lines to create any number of paths, with the path representing the flow of research inquiry that emerges in a given study. For example, Figure 3 represents a hypothetical study of the effects of prompt characteristics and test-taker expectations on the genre of a text produced in a classroom context.

Elements not included in a given study can be kept in the representation: While they are not the focus of research, they still exist as part of the assessment situation.

Table 1 shows a classification of a sample of recent research studies in writing assessment in terms of the elements of the visual representation that they have included. This list is limited to empirical studies only of writing tests or other formal English writing tasks, as it represents the vast majority of work in the field. The summary of Table 1 is not intended to be exhaustive; however, this exercise serves to demonstrate that (a) all the elements of the visual representation are currently seen in multiple combinations in current research, with some relationships explored more greatly than others; and (b) each of these studies could be visually depicted as in Figure 3.

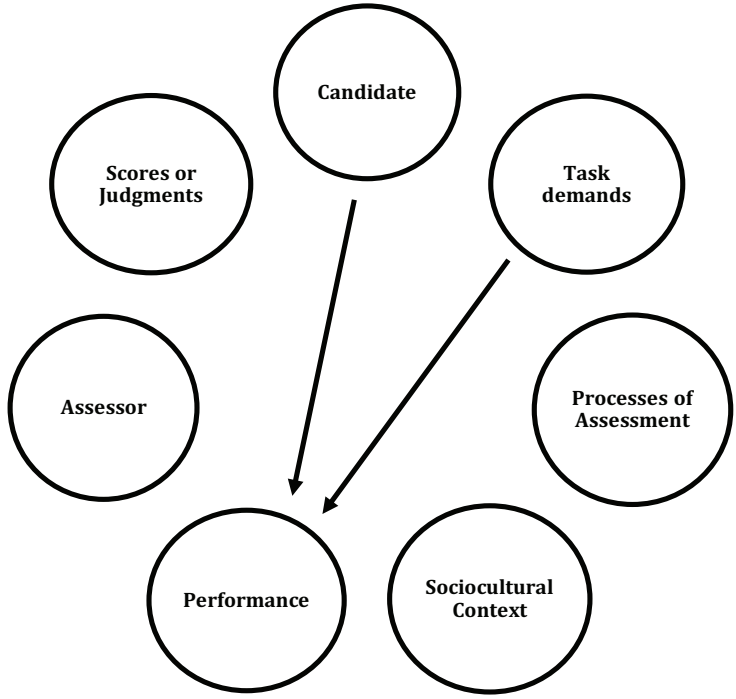

Figure 3. Example of a research path for an empirical study in writing assessment.

Note that even if two studies have focused on the same elements, this does not necessarily mean they have examined the same types of relationships among elements—there are many different research paths possible. For example, Luce-Kapler and Klinger (2005) and Yu, Rea-Dickins, and Kiely (2007) both focus on the elements of task demands and test characteristics, but Luce-Kapler and Klinger (2005) examine test-taker-reported impressions of task demands, while Yu et al. (2007) examine the cognitive processes of test takers on a specific task during the testing process. Of course, the authors of these studies themselves may indeed describe their work in varied ways in terms of the relationships among these elements. I use the terms of the proposed visual scheme in the left column while I keep the terms used by the authors in the summaries on the right.

Table 1
Recent Studies as Related to Elements of the Visual Representation

Combinations of elements of the visual representation	*Examples of empirical studies addressing connections among these elements*
Task demands, score/judgement	• Prompt factors and task effects: Weigle, 1999 • Effects of various time limits on scores: Powers & Fowles, 1996
Task demands, performance	• Comparison of genre elements—as elicited by prompts—with textual elements of performance: Beck & Jeffery, 2007 • Comparative discourse analysis of TOEFL tasks: Cumming et al., 2005
Task demands, performance, score/judgement	• Prompt effects (choice of topics) on score and text characteristics: Cho, 2003; Jennings, Fox, Graves, & Shohamy, 1999
Task demands, assessor, score/judgement, performance	• Effects of task and rater on ratings, and effects of task demands on discourse: Upshur & Turner, 1999
Task demands, performance, process of assessment, sociocultural context	• Analysis of prompts, texts, scoring guides, and interviews with stakeholders: Braine, 2000 • Comparing test design and washback function: Qi, 2005
Candidate, score/judgement	• Comparing scores of native and non-native English test takers: Ruetten, 1994 • Effects of computer familiarity on test-taker scores: Taylor, Jamieson, Eignor, & Kirsch, 1998
Candidate, process of assessment	• Students' attitudes toward the criteria by which they are assessed in university courses: Morozov, 2011
Candidate, performance, score/judgement	•-Effects of handwritten vs. typed tasks by different test-taker language groups on scores: Wolfe & Manalo, 2005

(continued on next page)

Combinations of elements of the visual representation	*Examples of empirical studies addressing connections among these elements*
Candidate, task demands	• Test-taker impressions of task demands: Luce-Kapler & Klinger, 2005 • How test-takers choose prompts: Polio & Glew, 1996 • Cognitive processes of test-takers on an IELTS task: Yu et al., 2007
Candidate, task demands, score/judgement	• Comparing success rates of different student groups on various writing task types: Cheng, Klinger, & Zheng, 2007 • Prompt difficulty and gender effects on scores: Breland, Lee, Najarian, & Muraki, 2004
Candidate, score/judgement, performance	• Individual and paired test-takers compared on text characteristics: Wigglesworth & Storch, 2009
Candidate, score/judgement, sociocultural context	• Test-taker and teacher perceptions of students' language ability compared to TOEFL scores: Johnson, Jordan & Poehner, 2005
Candidate, score/judgement, process of assessment	• Evaluation of a diagnostic procedure for academic skills, with scores compared by ethnic group: Erling & Richardson, 2010
Candidate, assessor, task demands, performance, sociocultural context	• Comparisons made of perceived genre requirements of the task by test-takers, raters, and others who make decisions based on test results; genre analysis of performances: Baker, 2009
Performance, score/judgement	• Influence of handwritten vs. typed texts on scores: Powers, Fowles, Farnum, & Ramsey, 1994; Russell & Tao, 2004
Assessor, score/judgement	• Rater background characteristics: Brown, 1995; Erdosy, 2004; Johnson & Lim, 2009; Kim 2009; Lumley & McNamara, 1995; Shi, 2001
Assessor, process of assessment	• Rater classification in terms of the importance attached to rating criteria: Eckes, 2008
Assessor, process of assessment, score/judgement	• Rater training effects on scores: Barrett, 2001; Elder, Knoch, Barkhuizen, & von Randow, 2005; Hoyt & Kerns, 1999; Kondo-Brown, 2002; McNamara, 1996; Weigle, 1998; Weir, 2005 • Comparing scores by raters of different backgrounds with different scale types: Shohamy, Gordon, & Kraemer, 1992 • Cognitive processes of raters with differing experience during the rating process with different rating scales: Barkaoui, 2010 • Cognitive processes of raters (what raters attend to while rating): Cumming, Kantor, & Powers, 2002; Pollitt & Murray, 1996 • Scores on different rating scales, compared to rater perceptions of scales: Knoch, 2009 • Rater bias patterns with a rating scale: Schaefer, 2008

(continued on next page)

Combinations of elements of the visual representation	*Examples of empirical studies addressing connections among these elements*
Assessor, process of assessment, performance	• Raters' use of different rating scales to assess academic style in texts: Knoch, 2007; 2008 • Raters' assessment of cohesion/coherence on an IELTS task: Wilson & Cotton, 2008
Assessor, processes of assessment, performance, score/judgement	• Examining rater impressions and textual features in assigning scores with a rating scale: Lumley, 2002 • Human vs. electronic scores, compared on text characteristics: Chodorow & Burstein, 2004 • Raters' interpretation of a scale in the assessment of grammatical ability: Neumann, 2010
Assessor, task demands, score/judgement	• Expert raters' judgements of prompt difficulty and effects on score: Hamp-Lyons & Prochnow-Mathias, 1994 • Reliability in scoring by expert vs. lay readers, by rating task (Schoonen, Vergeer, & Eiting, 1997)
Assessor, sociocultural context, score/judgement	• Comparing rater scoring in high- and low-stakes situations; incorporating rater perceptions of scoring (Baker, 2010)
Process of assessment, score/judgement	• Impact of rater discussion on scores: Johnson, Penny, Gordon, Shumate, & Fisher, 2005 • Comparing rater scoring with different rating scales: Barkaoui, 2007; Song & Caruso, 1996
Process of assessment, task demands, score/judgement	• Generalizability analysis comparing writing-only tasks and integrated tasks, with two different rating procedures (same raters for all tasks types vs. different raters for each task type): Gebril, 2010
Processes of assessment, performance, score/judgement	• Rating essays on paper vs. a computer screen: Coniam, 2009
Assessor, performance, process of assessment, score/judgement, sociocultural context	• Multiple activities related to a teacher-verification study of prototype tasks for the new TOEFL: Cumming, Grant, Mulcahy-Ernt, & Powers, 2004

Conclusion

This visual representation is potentially beneficial because, without supposing a specific theoretical orientation to the research act, it presents an environment to critically conceptualize the writing assessment enterprise in all its complexity and variety. A heightened awareness of the complexity of the writing assessment process would benefit anyone involved in the teaching or assessment of writing. This may be particularly helpful for new researchers who are finding their way into this increasingly diverse body of work, as well as writing teachers interested in engaging in action research. Certainly, teachers negotiate and question many of these elements and relationships every day.

We are becoming increasingly aware that visual displays of complex systems reveal relationships that would not be salient with textual presentation alone—revealing underexplored relationships among elements and favouring the cultivation of innovative research questions that may incorporate more elements than would have been considered otherwise. For example, aside from studies of (mostly classroom-based) washback (e.g., Qi, 2005), researchers have identified a lack of consideration of the sociocultural context in empirical studies of writing assessment (see Barkaoui, 2007, 2010; Weigle, 2002). Research lacunae such as this might become more prominent through this visual representation: if all the studies in Table 1 were depicted visually, it would become starkly apparent that the element of sociocultural context is rarely part of any path of inquiry.

This proposal for a visual representation is tentative at best: it is an organic scheme that continues to allow for questioning of the nature of each of the elements within it. The imperative is to continue to engage with visual representations such as these, which have the power to "engage deeply, evoke experiences, and … offer imaginary, constructive, and even transformative meaning" (Sanders-Bustle, 2003, p. x).

The Author

Beverly Baker is Assistant Professor at the Official Languages and Bilingualism Institute at the University of Ottawa, where she also serves as Director, Language Assessment. She currently serves as Vice President of the Canadian Association of Language Assessment, and has also been an ESL/EFL teacher for 20 years.

References

Bachman, L. F. (2002). Some reflections on task-based performance assessment. *Language Testing, 19*(4), 453–476.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their place in the real world.* Oxford, UK: Oxford University Press.

Baker, B. A. (2009). Conflicting genre expectations in a high stakes writing test for teacher certification in Quebec. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7*, 1–19.

Baker, B. A. (2010): Playing with the stakes: A consideration of an aspect of the social context in a gatekeeping writing assessment. *Assessing Writing, 15*(3), 133–153.

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

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74.

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal, 2*(1), 49–58.

Beck, S. W., & Jeffery, J. V. (2007). Genres of high-stakes writing assessments and the construct of writing competence. *Assessing Writing, 12,* 60–79.

Braine, G. (2000). When an exit test fails. *System, 29*, 221–234.

Breland, H., Lee, Y-W., Najarian, M., & Muraki, E. (2004). *An analysis of TOEFL-CBT writing prompt difficulty and comparability for different gender groups* (TOEFL Research Report 76). Princeton, NJ: Educational Testing Service.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12*(1), 54–71.
Card, S. K., Mackinlay, J. D., & Schneiderman, B. (1999). *Readings in information visualization: Using vision to think.* Burlington, MA: Morgan Kaufmann.
Cheng, L., Klinger, D. A., & Zheng, Y. (2007). The challenges of the Ontario Secondary School Literacy Test for second language students. *Language Testing, 24*(2), 185–208.
Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing, 8*(3), 165–191.
Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (TOEFL Research Report 73). Princeton, NJ: Educational Testing Service.
Clarke, A. E. (2003). Situational analyses: Grounded theory mapping after the postmodern turn. *Symbolic Interaction, 26*(4), 553–576.
Coniam, D. (2009). Discrepancy essays: Natural phenomenon or problem to be solved? *Melbourne Papers in Language Testing, 14*(2), 1–31.
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,* 5–43.
Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal, 86*(1), 67–96.
Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155–185.
Elder, C., Knoch, U., Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175–196.
Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions* (TOEFL Research Report 70). Princeton, NJ: Educational Testing Service.
Erling, E. J., & Richardson, J. T. E. (2010). Measuring the academic skills of university students: Evaluation of a diagnostic procedure. *Assessing Writing, 15,* 177–193.
Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15,* 100–117.
Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research.* Chicago, IL: Aldine.
Hamp-Lyons, L., & Prochnow-Mathias, S. (1994). Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing, 3*(1), 49–68.
Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods, 4*(4), 403–424.
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.
Johnson, K. E., Jordan, S. R. & Poehner, M. E. (2005). The TOEFL trump card: An investigation of test impact in an ESL classroom. *Critical Inquiry in Language Studies: An International Journal, 2*(2), 71–94.
Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26*(4), 485–505.
Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve accuracy of scores? *Language Assessment Quarterly, 2*(2), 117–146.
Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187–217.
Knoch, U. (2007). "Little coherence, considerable strain for reader": A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12,* 108–128.

Knoch, U. (2008). The assessment of academic style in EAP writing: The case of the rating scale. *Melbourne Papers in Language Testing, 13*(1), 34–67.
Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275–304.
Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(3), 1–31.
Luce-Kapler, R., & Klinger, D. (2005). Uneasy writing: The defining moments of high-stakes literacy testing. *Assessing Writing, 10*, 157–173.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276.
Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54–71.
McNamara, T. (1996). *Measuring second language performance.* New York, NY: Longman.
McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell.
Morozov, A. (2011). Student attitudes toward the assessment criteria in writing-intensive college courses. *Assessing Writing, 16,* 6–31.
Neumann, H. (2010). *What's in a grade? A mixed methods investigation of teacher assessment of grammatical ability in L2 academic writing* (Unpublished doctoral dissertation). McGill University, Montreal, QC.
Norris, J. M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing, 19*(4), 337–346.
Polio, C., & Glew, M. (1996). ESL writing prompts: How students choose. *Journal of Second Language Writing, 5*(1), 35–49.
Pollitt, A., & Murray, N. L. (1996). What raters *really* pay attention to. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition, and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem.* Cambridge, UK: CUP.
Powers, D. E., & Fowles, M. E. (1996). Effects of applying different time limits to a proposed GRE writing test. *Journal of Educational Measurement, 33*(4), 433–452.
Powers, D., Fowles, M., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31*(3), 220–233.
Qi, L. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142–173.
Ruetten, M. K. (1994). Evaluating students' performance on proficiency exams. *Journal of Second Language Writing, 3*(2), 85–96.
Russell, M., & Tao, W. (2004). The influence of computer-print on rater scores. *Practical Assessment, Research & Evaluation, 9*(10). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=10
Sanders-Bustle, L. (2003). Preface. In L. Sanders-Bustle (Ed.), *Image, inquiry, and transformative practice: Engaging learners in creative and critical inquiry through visual representation* (pp. ix–xxii). New York, NY: Peter Lang.
Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465–493.
Schoonen, R., Vergeer, M., & Eiting, M. (1997). The assessment of writing ability: Expert readers vs. lay readers. *Language Testing, 14*(2), 157–184.
Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge, UK: Cambridge University Press.
Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18,* 303–325.
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.
Song, 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*(2), 163–182.

Strauss, A., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory.* Thousand Oaks, CA: Sage.
Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks* (TOEFL Research Report No. 61). Princeton, NJ: Educational Testing Service.
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*(1), 82–111.
Weigle, S. C. (1998). Using facets to model rater training effects. *Language Testing, 15*(2), 263–287.
Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantative and qualitative approaches. *Assessing Writing, 6,* 145–178.
Weigle, S. C. (2002). *Assessing writing.* Cambridge, UK: Cambridge University Press.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* New York, NY: Palgrave MacMillan.
Wigglesworth, G., & Storch, N. (2009). Pair versus individual writing: Effects on fluency, complexity and accuracy. *Language Testing, 26*, 445–466.
Wilson, K., & Cotton, F. (2008). *An investigation of examiner rating of coherence and cohesion in the IELTS Writing Task 2* (IELTS Research Reports Round 14). London, UK: British Council/ UCLES.
Wolfe, E. W., & Manalo, J. R. (2005). *An investigation of the impact of composition medium on the quality of scores from the TOEFL writing section: A report from the broad-based study* (TOEFL Research Report No. 72).). Princeton, NJ: Educational Testing Service.
Yu, G., Rea-Dickins, P., & Kiely, R. (2007). *The cognitive processes of taking IELTS academic Writing Task 1* (IELTS Research Reports Round 13). London, UK: British Council/UCLES.