



MJAL 2:6 October 2010

ISSN 0974-8741

Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

Distinguishing linguistic and discourse features in ESL students' written performance

Anthony Becker

*Northern Arizona University, Department of English, Liberal Arts Building 18,
Room 140, PO Box 6032, Flagstaff, AZ 86011, USA*

Tel.: +1 215 280 5168.

Email address: anthony.becker@nau.edu

Author's Profile

Anthony Becker is a PhD candidate in applied linguistics in the Department of English at Northern Arizona University. His research and teaching interests include second language assessment and second language writing.



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

Abstract

Prior L2 studies investigating the feature differences identified in performance-based writing have shown that some features of writing help to distinguish the performance of learners with varying levels of proficiency. However, the results of these studies have varied considerably, creating some uncertainty about what distinguishes lower-level writers from more-advanced writers. This article reports a study carried out on a sample of end-of-semester essays written by 43 English as a second language (ESL) students attending an Intensive English Program (IEP) in the United States. Specifically, the aim of the study was to find out if 10 features of written discourse generated from a writing test task differed among three levels of students at the IEP. The results indicate that some writing features (e.g., words per T-unit, clauses per T-unit, and text length) clearly distinguish ESL students' writing at different levels of proficiency, while other features (e.g., coherence, cohesion, and lexical density) do not appear to distinguish students' writing at all.

Keywords: English as a second language; performance-based assessment; writing



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

1. Introduction

The ability to write effectively is an important part of the overall success at universities in the United States. Writing can be especially difficult for English as a Second Language (ESL) learners, since they must contend with learning the language and meeting the same writing standards that their native-speaking peers are expected to meet. As a result, many U.S. universities offer an Intensive English Program (IEP) to help prepare ESL learners for the [writing] demands of mainstream university courses. To distinguish students and measure their writing development, performance-based writing assessments, such as extended-response essays (see Weigle, 2002), are often utilized. However, little is understood about which (if any) linguistic and discourse features distinguish lower-level writers from more-advanced writers. In order to better understand L2 writing development, studies that investigate the feature differences of ESL students' performance-based writing assessments can prove to be useful.

2. Literature Review

Performance-based assessments can be used to indicate ESL students' writing progress and/or development. Performance assessments have gained acceptance as the primary means for assessing second language learners' writing abilities (Hamp-Lyons, 1991; McNamara, 1997; Shaw and Weir, 2007; Weigle, 2002), since performance assessments attempt to make use of tasks that closely reflect what learners will encounter in real-world contexts. For example, in IEPs, this type of assessment appears to provide evidence that second language learners are achieving the skills necessary for writing in university courses. As a result, performance



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

assessments are seen as useful sources of evidence about a student's writing development and are included in both large-scale tests (e.g., TOEFL) and smaller classroom-based tests (e.g., achievement tests).

Some research has looked at the use of writing performance assessments in large-scale tests, in order to compare the feature differences of students' writing. Such studies have focused on identifying the linguistic (and discourse) features that distinguish different levels of writing ability, as determined by scores received using holistic rubrics (e.g., Cumming, Kantor, Baba, Erdosy, Eouanzoui, and James, 2004, 2005; Frase, MacDonald, and Keenan, 1985; Hinkel, 2003; Shaw and Liu, 1998). For instance, Grant and Ginther (2000) looked at a sample of 90 Test of Written English (TWE) essays in order to identify differences in L2 students' writing. For their study, the TWE essays were separated into three levels (i.e., scores 3, 4, or 5) of proficiency. Using a computer-tagging program (for a description of the computer program see Biber, 1988), the linguistic features of the essays were identified and compared between the three levels of proficiency. The analysis included various linguistic features:

1. General features (e.g., essay length, word length, type/token ratio)
2. Lexical features (e.g., conjuncts, hedges, demonstratives, downtoners)
3. Grammatical features (e.g., nouns, nominalizations, personal pronouns, verbs, modals)
4. Clause-level features (e.g., subordination, relative clauses, passives)

The results of the study revealed that the frequency of numerous features (e.g., word count, type/token ratio, average word length, amplifiers, emphatics, demonstratives, subordination, relative clauses, and passives) increased as the level of proficiency increased.



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

Overall, TWE essays that received scores of 5 generally had higher means for most of the features than did essays that received scores of 4 or 3. In addition, TWE essays that received scores of 4 generally had higher means for most of the features than did essays that received scores of 3. The results of the study help to shed light on the potential usefulness of analyzing linguistic differences to distinguish levels of L2 writing performance.

Recently, Cumming et al., (2004, 2005) has been assessed how the discourse written for TOEFL[®] integrated writing tasks differed from the discourse written for TOEFL independent writing tasks. In their study, the authors selected 216 compositions written for six TOEFL tasks by 36 examinees at three different proficiency levels. Then, they coded each essay for a variety of discourse features (e.g., lexical complexity, syntactic complexity, rhetoric, and pragmatics) and used a non-parametric multivariate analysis of variance (MANOVA) to compare how these features occurred in the independent and integrated writing tasks. The results indicated that there were significant differences between the discourse written for independent and integrated writing tasks, in terms of word length, clauses per T-unit, the frequency of propositions, and the proportion of declarations and paraphrases. This study highlights the fact that performance-based tasks, such as independent and integrated writing tasks, can elicit different types of written discourse. Furthermore, this study also identifies variables that can help to distinguish examinees at different score levels.

Both the Grant and Ginther (2000) and Cumming et al. (2004, 2005) studies highlight how linguistic features typically included in scoring rubrics can be used to distinguish different levels



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

of writing ability. Specifically, these two studies looked at linguistic features that are typically used to analyze large-scale proficiency tests, such as the Next Generation TOEFL and TOEFL iBT. These two studies offer helpful methodologies for analyzing linguistic features in performance-based writing assessments, such as writing achievement tests or formative writing assignments.

In order to better understand the development of writing for ESL learners, more studies could aim to investigate the linguistic and discourse features of ESL learners' writing, using similar methodologies employed by Grant and Ginther (2000) and Cumming et al. (2004, 2005), as these studies have clear procedures for distinguishing levels of writing proficiency and the features that relate to the writing at those levels. Additional studies are likely to shed light on which features are indicative of ESL students' writing ability, which can help to improve writing instruction, as well as the assessment of writing performance. Borrowing from these two studies, the present study investigated the characteristics of writing performance for ESL students at different levels of an IEP curriculum. Specifically, this study attempted to address the following research question:

How does the written discourse generated from a writing test task differ among different levels of students at the IEP?

3. Methodology

3.1. Participants

Participants included ESL students (N = 43) enrolled in IEP courses at a medium-sized university in the southwestern United States. These were first-semester students taking English



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

classes in order to gain admission into undergraduate courses at the university. The participants included students from three different IEP levels: Level 1 (low-intermediate), Level 2 (high-intermediate), and Level 3 (advanced). The participants' level of English-language proficiency was determined by scores from the IEP placement test, which was administered to incoming students at the beginning of the fall 2008 semester. As shown in Table 1, both males ($n = 25$) and females ($n = 18$) participated in this study. Several different L1 language backgrounds were included, however, L1 Chinese speakers ($n = 35$) represented approximately 80 percent of the entire sample for this study. None of the participants had spent more than a total of one year living or visiting in the United States, but the participants had received some formal instruction ($M = 6.53$ years; $SD = 2.48$) for learning English in their home country.

[Insert Table 1 here]

3.2. Procedures

Near the end of the fall 2008 semester, the IEP administered week 15 achievement tests for all three learner levels. The achievement tests were intended to measure students' mastery of content that was taught throughout a five-week segment and included the assessment of all four language skills (i.e., listening, reading, speaking, and writing). The writing section for each achievement test consisted of various tasks, including editing, writing, and revising an essay. For the independent essay writing, students were given 30 minutes to complete a writing task, using grammatical structures, vocabulary, and writing conventions that were taught. With the



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

independent writing tasks, students were asked to respond to a given writing prompt (e.g., write about why it is important to save indigenous cultures).

Once the essays were completed, two trained IEP instructors then scored each essay and assigned a holistic score. For each of the three levels, in-house, level-specific scoring rubrics were used to rate the written essays. The scoring rubric used for each level included both an analytic and a holistic scoring scale. The non-numeric, analytic scale focused on three aspects (i.e., organization, content, and language use) of students' writing and was used to derive a numeric score on the 6-point holistic scale. The holistic score from each rater was combined to form a total score, ranging from 0-10 points. In addition, since the IEP wanted to assign more weight to the students' writing, this combined, holistic score was then transformed into a converted score, ranging from 0-57 possible points.

4. Design

This study incorporated a cross-sectional, between-groups design. A cross-sectional design is used when participants and/or variables are observed or measured at a single point in time (Creswell, 2002). Students' end-of-semester writing tests were only collected, resulting in the evaluation of one essay per participant.

5. Analysis

Following Cumming et al. (2004, 2005), this study coded for many of the same linguistic and discourse features in IEP students' written essays. Specifically, the variables and their operationalizations included:



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

1. *Text length*: the total number of words written in a composition within the time allocated for a particular task (Chenoweth and Hayes, 2001).
2. *Lexical sophistication* (operationalized in three ways):
 - 2.1. *average word length*: the total number of characters divided by the total number of words per composition. (Engber, 1995; Frase, Faletti, Ginther, and Grant, 1999);
 - 2.2. *type/token ratio*: the number of different lexical words divided by the total number of words per composition (Cumming and Mellow, 1996; Grant and Ginther, 2000);
 - 2.3. *lexical density*: the total number of content words divided by the total number of words per composition (Cobb, 2009).
3. *Syntactic complexity* (operationalized three ways):
 - 3.1. *number of T-units¹*: the total number of T-units per composition (Ortega, 2003).
 - 3.2. *clauses per T-unit*: the total number of clauses divided by the total number of T-units (Ishikawa, 1995; Wolfe-Quintero, Inagaki, and Kim, 1998);
 - 3.3. *words per T-unit*: the total number of words divided by the total number of T-units (Polio, 1997).
4. *Grammatical accuracy*: a holistic rating of errors as either 1 [many errors (e.g., more than 3 per T-unit) in grammar, punctuation, spelling, and word choice; errors often affect comprehensibility], 2 [some errors (e.g., 2 or 3 per T-unit) in grammar, punctuation, spelling, and word choice; passage is mostly comprehensible to reader],

¹ A *T-unit* is defined as an independent clause with all of its dependent clauses (Ortega, 2003).



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

or 3 [few errors (e.g., about 1 per T-unit or less) in grammar, punctuation, spelling, and word choice; comprehensibility seldom obscured for reader] (Cumming et al. 2004; Hamp-Lyons and Henning, 1991).

5. *Organization*: the effectiveness and appropriateness in the organization of ideas above the sentence level and among different parts of the text (adapted from Chiang, 1999). Organization was operationalized in two ways:
 - 5.1. *coherence*: the discourse-level features that are responsible for giving prose its overall organization;
 - 5.2. *cohesion*: the degree of compactness or efficiency with which ideas within individual sentences are related to each other.

Coherence and cohesion were measured according to the degree of agreement with statements related to the two subscales. The statements were scored using a 5-point Likert scale: 5 (Strongly Agree), 4 (Agree), 3 (Undecided), 2 (Disagree), and 1 (Strongly Disagree). For the coherence subscale, the minimum possible score was 10 and the maximum possible score was 50. For the cohesion subscale, the minimum possible score was 7 and the maximum possible score was 35. A high score for either subscale suggested that an essay displayed a high degree of coherence and/or cohesion.

To establish inter-coder agreement, a second coder reviewed the essays. Simple-agreement percentages were determined for four variables: T-units (42/43 = 98%), clauses per T-unit (38/43 = 89%), words per T-unit (42/43 = 98%), and grammatical accuracy (30/43 = 70%). While the agreement percentages for the three syntactic complexity measures were satisfactory, the inter-



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

coder agreement for grammatical accuracy was somewhat low. This seemingly low agreement rate was considered to be acceptable, as the researcher and coder did not have any grammatical accuracy scores that differed more than one point. In addition, inter-coder reliability was determined for the organization measures, coherence ($P_r = .72$) and cohesion ($P_r = .76$). The reliability coefficients for these measures were considered acceptable for the present study (see Wasserman and Bracken, 2004). Inter-coder agreement for text length and lexical sophistication was not calculated, since these measures were determined using an online program, *Vocab Profile* (Cobb, 2009).

Additionally, since the text length of the essays varied considerably for each learner level (the average essay lengths ranged from 155-340 words), six of the measures (i.e., average word length, clauses per T-unit, lexical density, number of T-units, type/token ratio, and words per T-unit) were normalized to 200 words. The normalization process makes frequencies from samples of markedly different sizes comparable by bringing them to a common base (McEnery, Xiao, and Tono, 2006). As a result of normalizing, more accurate comparisons for these measures could be made, as they would not be influenced by text length.

6. Results

A One-way Analysis of Variance (ANOVA) was used to examine whether the 10 measures were different between the three learner levels. The independent variable represented :

[Insert Table 2 here]

The three learner levels: 1) Level 1 (low-intermediate); 2) Level 2 (high-intermediate); and 3)



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

Level 3 (advanced). The dependent variables included the scores for the 10 writing variables described earlier. See Table 2 for the means and standard deviations for each of the 10 measures for the three learner groups.

The tests for normality, standardized skewness and Shapiro-Wilks, indicated that the data were statistically normal (see Tabachnick and Fidell, 2007). The one-way ANOVA for scores on the 10 measures (see Table 3) revealed statistically significant main effects for seven measures: average word length [$F(2, 40) = 20.13, p < .05$], clauses per T-unit [$F(2, 40) = 18.09, p < .05$], lexical density [$F(2, 40) = 6.62, p < .05$], number of T-units [$F(2, 40) = 20.92, p < .05$], text length [$F(2, 40) = 20.13, p < .05$], type/token ratio [$F(2, 40) = 7.26, p < .05$], and words per T-unit [$F(2, 40) = 34.91, p < .05$]. The Omega-squared (ω^2) values indicate that these particular measures each contributed a significant amount to the total variance in distinguishing students' writing.

Post hoc comparisons, using the Tukey post hoc procedure, were conducted to determine which pairs of the three learner levels differed significantly. These results are presented in Table 4 and indicate several significant differences between learner levels. Students belonging to

[Insert Tables 3 and 4 here]

Levels 1 and 3 used significantly longer words than students belonging to Level 2. The effect sizes² for these two differences were 1.93 (Level 1 vs. Level 2) and 3.07 (Level 2 vs. Level 3). Those students belonging to Levels 2 and 3 used significantly more clauses per T-unit than students belonging to Level 1. The effect sizes for these two significant effects were 1.78 (Level

² All effect sizes for the post hoc comparisons in this study can be interpreted as being large (see Cohen, 1988).



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

1 vs. Level 2) and 1.89 (Level 1 vs. Level 3). Students belonging to Levels 1 and 2 used significantly more T-units than students in Level 3. The effect sizes for these two significant effects were 1.38 (Level 1 vs. Level 3) and 3.85 (Level 2 vs. Level 3).

Those students belonging to Level 2 had significantly longer text length than students in Levels 1 and 3; the effect sizes for these two significant effects were 1.86 (Level 1 vs. Level 2) and 2.90 (Level 2 vs. Level 3). Also, a significant difference was found between Levels 2 and 3 for type/token ratio. The effect size for this significant effect was 1.66, respectively. Finally, students belonging to Levels 2 and 3 used significantly more words per T-unit than students in Level 1, with effect sizes being 1.37 and 1.89 respectively; the difference between Levels 2 and 3 was also significant, with an effect size of 1.78.

7. Discussion

In terms of the written features analyzed in this study, the IEP writing rubric appeared to distinguish between the different levels of proficiency for many of the measures. For example, two of the syntactic complexity measures (i.e., clauses per T-unit and words per T-unit) appeared to distinguish written performance between Level 1 writers and writers from Levels 2 and 3. In addition, words per T-unit seemed to distinguish the performance between Level 2 and Level 3 writers. Increase as the learner level increased. These findings reinforce some research (e.g., Cumming et al., 2004; Frase et al., 1999) that claims that some measures of syntactic complexity, such as clauses per T-unit and words per T-unit, can be effective for distinguishing different levels of academic writing.



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

However, the third syntactic complexity measure (i.e., number of T-units) did not appear to distinguish written performance. For this measure, Level 2 and Level 1 essays contained significantly more T-units than did Level 3 essays, suggesting that more advanced [Level 3] writers in this study seemed to condense more information into individual T-units, while less-skilled writers (belonging to Levels 1 and 2) appeared to use more T-units that were short and included noticeably fewer clauses. The fact that students in Levels 1 and 2, on average, included substantially more words in their essays than did students in Level 3 also implies that lower-level writers expressed their ideas in extended (and wordier) prose, while more advanced writers seemed to express their ideas succinctly and efficiently.

While the present study suggests that text length is not necessarily indicative of better writing performance, these findings appear to run contrary to earlier studies which have shown that increased text length is associated with higher scores by raters and higher levels of proficiency (Grant and Ginther, 2000; Oijima, 2006). A possible explanation for the differences between Level 2 and Level 3 writers could have been related to the writing instruction utilized in the different levels. For example, in the Level 2 classroom, students are typically instructed to expand their ideas and write as many details (and examples) as possible. Meanwhile, Level 3 students are typically encouraged to write more efficiently and with greater complexity. As Biber, Johansson, Leech, Conrad, and Finegan (1999) indicate, more sophisticated levels of writing are typically earmarked by linguistically-efficient features (e.g., noun-modifying prepositional phrases), which enable a writer to present information in a compact, yet meaningful manner.



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

Finally, the two indicators of lexical sophistication (i.e., average word length and type/token ratio) only appeared to somewhat distinguish students' writing performance. For instance, Level 2 writers had substantially shorter word lengths than did writers belonging to Levels 1 and 3. In terms of type/token ratio, Level 3 writers appeared to use substantially different lexical words than Level 2 writers, but not Level 1 writers. While these findings cannot be readily explained, it could be the fact that Level 2 students in this particular IEP are not getting the same type of lexical exposure as their Level 1 and Level 3 counterparts. Specifically, the textbooks used for the Level 2 curriculum at this IEP, do not include vocabulary learning as an extensive component in each thematic unit (IEP director, personal communication, April 15, 2009). Therefore, students in Level 2 are likely not getting the same frequency of exposure to new and/or recycled lexical items (as are students in Levels 1 and 3) and are likely relying on limited lexical resources.

8. Conclusion

The aim of the present study was to find out if the written discourse generated from a writing test task differs among three levels of students at an IEP. While some of the features that were focused on for this study seemed to distinguish between different learner levels (e.g., words per T-unit, clauses per T-unit, and text length), other features did not appear to clearly distinguish among the three levels (e.g., coherence, cohesion, and lexical density). Although this study was conducted using a small sample size, the results indicate that additional research needs to be done to find out if (and which) specific features of writing are robust in distinguishing different



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

levels of L2 writing development. Furthermore, studies such as this, which look at features across different levels of writing proficiency, could be replicated in different writing contexts, in order to gain a better understanding of L2 writing development.

References

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999. Standards for educational and psychological testing. American Psychological Association, Washington, DC.
- Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., 1999. Longman grammar of spoken and written English. Pearson Education Ltd, Essex.
- Bachman, L.F., 2002. Alternative interpretations of alternative assessments: Some validity issues in educational performance assessments. Educational Measurement: Issues and Practice, 21, 5-18.
- Chenoweth, A., Hayes, J.R., 2001. Fluency in writing: Generating text in L1 and L2. Written Communication, 18, 90-98.
- Chiang, S.Y., 1999. Assessing grammatical and textual features in L2 writing samples: The case of French as a foreign language. The Modern Language Journal, 83, 219-232.
- Cobb, T., 2009. VocabProfile [Computer software]. Retrieved January 10, 2009, from <http://www.lex tutor.ca/vp>.
- Cohen, J., 1988. Statistical power analysis for the behavioral sciences, second ed. Lawrence Earlbaum Associates, Hillsdale, New Jersey.
- Creswell, J.W., 2002. Educational research: Planning, conducting, and evaluating quantitative



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

- and qualitative research. Pearson Education, Inc., Upper Saddle River, New Jersey.
- Cumming, A., 1997. The testing of writing in a second language. In Clapham, C., Corson, D. (Eds.), *Encyclopedia of language and education: Language testing and assessment*. Kluwer Academic Publishers, Norwell, Massachusetts, pp. 131-139.
- Cumming, A., Mellow, D., 1996. An investigation into the validity of written indicators of second language proficiency. In Cumming, A., Berwick, R. (Eds.), *Validation in language testing*. Clevedon, Multilingual Matters, United Kingdom, pp. 72-93.
- Cumming, A. et al., 2004. Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for Next Generation TOEFL. Educational Testing Service, Princeton, New Jersey.
- Cumming, A. et al., 2005. Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, 10, 5-43.
- Engber, C., 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4, 139-155.
- Frase, L., MacDonald, N., Keenan, S., 1985. Intuitions, algorithms, and a science of text design. In Duffy, T., Waller, R. (Eds.), *Designing Usable Texts*. Academic Press, New York, pp. 97-112.
- Frase, L. et al., 1999. Computer analysis of the TOEFL test of written English. Educational Testing Service, Princeton, New Jersey.
- Grant, L., Ginther, A., 2000. Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, 9, 123-145.



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

- Hamp-Lyons, L., 1991. Scoring procedures for ESL contexts. In Hamp-Lyons, L. (Ed.), *Assessing second language writing in academic contexts*. Ablex, Norwood, New Jersey, pp. 241–276.
- 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.
- Hinkel, E., 2003. Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, 37, 275-301.
- Ishikawa, S., 1995. Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing*, 4, 51-69.
- Leung, C., Lewkowicz, J., 2006. Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly*, 40, 211-234.
- McEnery, T., Xiao, R., Tono, Y., 2006. Corpus-based language studies: An advanced resource book. Routledge, New York.
- McNamara, T. F., 1997. Performance testing. In Clapham, C., Corson, D. (Eds.), *Encyclopedia of language and education: Language testing and assessment*. Kluwer Academic Publishers, Norwell, Massachusetts, pp. 131-139.
- Oijima, M., 2006. Concept mapping as pre-task planning: A case study of three Japanese ESL writers. *System*, 34, 566-585.
- Ortega, L., 2003. Syntactic complexity measures and their relationship to L2 proficiency: A



Distinguishing linguistic and discourse features in ESL students' written performance by Anthony Becker

- research synthesis of college-level L2 writing. Applied Linguistics, 24, 492-518.
- Polio, C., 1997. Measures of linguistic accuracy in second language writing research. Language Learning, 47, 101-143.
- Sakyi, A.A., 2003. A study of the holistic scoring behaviours of experienced and novice ESL Instructors. Unpublished doctoral dissertation, University of Toronto, Canada.
- Shaw, S.D., Weir, C.J., 2007. Examining writing: Research and practice in assessing second language writing. Cambridge University Press, Cambridge.
- Shaw, P., Liu, E. T.-K., 1998. What develops in the development of second language writing. Applied Linguistics, 19, 225-254.
- Tabachnick, B.G., Fidell, L.S., 2007. Using multivariate statistics, fifth ed. Pearson Education, Inc, New York.
- Wasserman, J.D., Bracken, B.A., 2004. Psychometric characteristics of assessment procedures. In Graham, J.R., Naglieri, J.A. (Eds.), Handbook of psychology: Assessment psychology. John Wiley and Sons, Inc, New Jersey, pp. 43-66.
- Weigle, S. C., 2002. Assessing writing. Cambridge University Press, New York.
- Wolfe-Quintero, K., Inagaki, S., Kim, H.-Y., 1998. Second language development in writing: Measures of fluency, accuracy, and complexity. University of Hawai'i at Manoa, Honolulu, Hawai'i.