JER | Journal of ELT Research
Vol. 3, No. 2, 2018, 142-151 DOI: 10.22236/JER_Vol3Issue2

Raters' Decision Making Variations in Scoring Writing Samples

Asma Dabiri*

Shiraz University, Iran

DOI: 10.22236/JER_Vol3Issue2pp142-151

This study examined raters' decision making variations in a writing assessment task focusing on individual differences in decision-making style (DMS). The participants' of the study were six TEFL instructors. A rating scale obtained from Turner and Upshur (2002) and a General Decision Making Style Inventory questionnaire, GDMSI, obtained from Scott and Bruce (1995) were administered to raters. The results showed the raters' behaviors were not equally the same in the same rating situations. These discrepancies suggested individual socio-cognitive differences in accounting for some rater's variability in scoring. In addition, characteristics of the texts (not just individual cognitive characteristics) favored certain decision-making behaviors. Accordingly, a re-visioning of the one-size-fits-all approach that is currently the norm in the training of raters for scoring writing assessments is needed. Further, a more individualized approach to rater training is needed. If the individual decision-making style to a great extent is dependent on basic cognitive abilities that are stable and not easily changed, then the decision support systems need to be flexible in order to match the needs of the individual decision makers.

Keywords: decision-making variations, subjective scoring, writing assessment

Studi ini meneliti beragam pembuat keputusan nilai pada tes menulis yang difokuskan pada perbedaan individu dalam gaya pengambilan keputusan. Enam instruktur bahasa Inggris berpartisipasi di studi ini. Skala pengukuran dari Turner dan Upshur (2002) dan kwesioner inventori gaya pengambilan keputusan dari Scott dan Bruce (1995) digunakan oleh penilai. Hasil studi ini menunjukan bahwa perilaku penilai tidak selalu sama dalam setiap situasi. Ini menunjukan perbedaan sosio-kognitif dalam menghitung. Disamping itu, karakteristik teks mempengaruhi perilaku pengambilan keputusan tertentu. Oleh karena itu perlu dilakukan perubahan terhadap pandangan yang mengatakan satu pendekatan cocok untuk semua. Pendekatan peniai yang lebih personal juga diperlukan. Jika gaya pengambilan keputusan tergantung pada kemampuan kognitif dasar yang stabil dan tidak berubah, sistem pendukung keputusan harus fleksibel agar cocok dengan kebutuhan pembuat.

* Corresponding author. Email: dabiri_asma@yahoo.com

ISSN: 2502-292X, e-ISSN 2527-7448.

DOI: 10.22236/JER_Vol3Issue2

INTRODUCTION

Making judgments and decisions include cognitive processes involved in retrieving information from memory and in making use of a stimulus input. These processes are affected by memory and other cognitive constraints, prior knowledge and experience with similar situations, construct-irrelevant factors, and other cognitive factors (Newell, Lagnado & Shanks, 2007). Raters in subjective scoring make use of a stimulus input (such as a composition or a speaking sample) to obtain a sufficient amount of information to compare to a rating scale. Raters are not able to retrieve or retain all relevant information, and they are affected by their prior experiences and personal backgrounds as they select, weigh, and integrate information into a final judgment.

Studies have been done to investigate the processes which raters go through in applying a rating scale in scoring writing assessments. A great deal of this work has focused on accounting for systematic variability in rater scoring. Research on rater training has so far suggested that training is useful in increasing rater consistency (Jang, Wagner & Park, 2014; McNamara, 1996; Plakans & Gebril, 2013; Weigle, 1999, 2002; Weir, 2005), but there continues to be unexplained variability that resists training (Crossley, Kyle & MacNamara, 2016; Hoyt & Kerns, 1999; Plakans & Gebril, 2013; Shrestha & Coffin, 2012).

Works on rater variability have mostly focused on raters' differing academic or disciplinary background (Barkaoui, 2010; Brown, 1995; Cumming, Kantor & Powers, 2002; Erdosy, 2004; Lumley & McNamara, 1995; Shi, 2001; Song & Caruso, 1996; Weigle, 1999), or on their language background (Johnson, 2009; Kim, 2009). Another focus of studies of rater behavior has been the differing importance raters attached to particular rating scale criteria or to particular elements of the students' performance (Barkaoui, 2010; Cumming et al., 2002; Eckes, 2008; Orr, 2002; Pollitt & Murray, 1996).

Still, not all rater variability is accounted for, and that at least some of this rater variability seems to be attributable to individual socio-cognitive characteristics of raters. What has not been examined is the impact that individual differences in cognitive style may have on rater behavior. Rater personality characteristics may have an effect on scoring behavior; however, they are less studied than other rater effects (Crossley et al., 2016; Lumley, 2002; Roohani & Taheri, 2015). Accordingly, this study examined raters' decision making variations in a writing assessment task focusing on individual differences in decision-making styles in an attempt to account for unexplained rater variability in scoring.

Literature on Raters' Decision Making

Cooksey, Freebody and Wyatt-Smith (2007) made use of judgment analysis (combined with think-aloud protocols) in studying teachers' assessment of written texts, while Wolfe, Kao and Ranney (1998) made use of the literature in expert judgment to examine cognitive differences in decision-making between experts and novices—what they called "individual differences in scorer cognition" (p. 465) or "stylistic differences in cognitive style that could affect decision-making" (Thunholm, 2004, p. 932). Dörnyei (2006) defines individual differences as "enduring personal characteristics that are assumed to apply to everybody and on which people differ by degree" (p. 42).

Milanovic, Saville and Shuhong (1996) developed a model for decision-making behavior of raters. In addition, in three exploratory studies, Cumming et al. (2002) created a descriptive framework for decision-making processes in rating TOEFL essays. Other work characterizing rating as a decision-making process included Lumley (2002), Orr (2002) and Kondo-Brown (2002). The major concern of these studies have not been the individual cognitive characteristics of decision makers. However, there has recently been increased

interest in the role of the decision maker in the decision-making process and, in particular, individual differences in cognitive style, defined by Messick (1984) as "characteristics of self-consistencies in information processing that develop in congenial ways around underlying personality trends" (p. 61). Individual differences in decision-making style can be viewed as "stylistic differences in cognitive style that could affect decision-making" (Thunholm, 2004, p. 932). Scott and Bruce (1995) defined decision making style (DMS) as "the learned, habitual response pattern exhibited by an individual when confronted with a decision situation" (p. 820). Thunholm (2004) questioned the characterization of decision-making style as a habit, suggesting that individual differences also exist in stable underlying cognitive abilities. Spicer and Sadler-Smith (2005) suggested that "decision making style is in fact a 'surface' manifestation of more stable underlying dimensions, which individuals are able to adapt or change" (p. 146).

To refer to the necessity for "a conceptually consistent and psychometrically sound measure of decision-making style" (Scott and Bruce, 1995, p. 818), they created a general decision-making style inventory (GDMSI), finding evidence for five distinct styles: rational, dependent, intuitive, avoidant, and spontaneous. Each type of DMS, as synthesized from the previous DMS literature (Scott & Bruce, 1995; Spicer & Sadler-Smith, 2005) was defined as:

- Rational DMS: preference for the systematic collection, evaluation, or weighing of information.
- Intuitive DMS: preference for relying on feelings, hunches, and impressions that cannot be put into words when making decisions.
- Dependent DMS: preference for drawing on the opinions or support of others; on receiving second opinions or advice.
- Avoidant DMS: preference for delaying decision-making, hesitating, or making attempts to avoid decision making altogether.
- Spontaneous DMS: preference for coming to a decision immediately or as early as possible.

Gambetti, Fabbri, Bensi and Tonetti (2007) found this five-factor model to provide a good fit for their own sample (Italian university students). In a validation study in two U.K. samples, Spicer and Sadler-Smith (2005) used factor analysis to compare two-, three-, and four-factor alternate models to Scott and Bruce's five-factor model. Their analysis indicated that Scott and Bruce's five-factor model had the best fit to their data.

The results in DMS studies can be summarized as follows: (a) there is sufficient evidence for the existence of a construct of DMS related to individual cognitive style, (b) whereas each of the five DMS types is conceptually independent, decision makers rely on a combination of them in their decision making, (c) the context for decision making also contributes to decision behaviors, and (d) further research has yet to be done regarding the application of the GDMSI to authentic decision making. It is acknowledged that the construct validity of the GDMSI is still relatively unproven and has low internal consistency in some of the scales (regarding rational DMS in particular). This is in addition to the known limitations with Likert-type response items. Second, there are additional limitations associated with the use of self-report scales in measuring psychological constructs (Crossley et al., 2016; Roohani & Taheri, 2015). McDonald (2008) discussed the overreliance on self-report questionnaires in personality research; commenting more generally on how more than simply self-reports are required to obtain a greater understanding of psychological constructs. One reason for this is the possibility of response bias in self-report measures. Shrestha and Coffin (2012) discussed

the response bias brought about by "faking," or "impression management" whereby respondents respond to questions in a way that portrays them in the best possible light. Paulhus (1991) referred to this phenomenon as "socially desirable responding". In terms of DMS, one may prefer to be viewed as a rational decision maker rather than as someone who attempts to avoid decision-making. Additionally, it should be stated that an inherently distorted self-view can cause response bias (McDonald, 2008). If this self-distortion bias is present, simply encouraging respondents to be truthful cannot correct it.

As McDonald (2008) advises, "When there is some doubt about whether the construct under investigation can be represented to its fullest extent, the use of multiple methods should definitely be considered" (p. 12). Therefore, although the GDMSI may be a useful tool, as a self-report measure it should be combined with other measures that may provide insight into individual differences in DMS. It was further explained that ideally, this would include the combination of self-report measures with observational measures of behavior. So, researches on decision-making styles needs to investigate the link between decision-making styles as measured by GDMS and behavior shown in realistic decision-making tasks. Consequently, the underlying research questions were posed in this study to investigate the link between decision-making styles as measured by GDMS and behavior shown in realistic decision-making tasks: (1) Can different sources of evidence be meaningfully combined to create decision-making style profiles of the raters in a writing assessment? (2) What is the relationship between decision-making behaviors and the texts being rated?

METHOD

Participants who voluntarily took part in the study were six TEFL instructors: three female (F) and three male (M). They had experience in teaching writing and grading writing assessments as well as teaching other subjects in TEFL. Participants' information is summarized in the following table.

Table 1. Participants

Rater	Gender	Experience in TEFL
1	F	9
2	F	5
3	F	8
4	M	4
5	M	12
6	M	6

A rating scale was a 4-point focused-holistic rating scale developed by Turner and Upshur (2002), and validation studies indicated that the scale functioned acceptably for the intended purpose in subjectively scoring writing tasks (Crossly, Jang, Wagner & Park, 2014; Kyle & McNamara, 2016; Plakans & Gebril, 2013; Shrestha & Coffin, 2012). To estimate inter-and intra-coder reliability, five pages of the coded data were randomly chosen. To gain the intra-coder reliability, they were coded again by the researcher ten days after the first coding, and the intra-coder reliability was found to be 0.98. For the inter-coder reliability, an assessor in TEFL not participating in the study coded them, and the inter-coder reliability was found to be 0.88.

A GDMSI (a general decision-making style inventory) questionnaire from Scott and Bruce (1995) was used. The questionnaire was a 5-point Likert-type scale. The face and content validity of the questionnaire was checked by two content specialists. They were asked to examine questionnaire items in order to specify whether the instrument adequately represented the mentioned content and objectives. In addition, in the pilot study, the internal consistency of the questionnaire was checked by Cronbach alpha (Coefficient alpha formula) and it was estimated to be 0.76.

Data Collection Procedures

The Training Session

For raters, training sessions preceded scoring (rating) writing assessments and consisted of familiarization with and discussion of the rating scale and of several model papers. It was intended to provide an opportunity to clarify the language of the scale descriptors, and to practice grading authentic papers with subsequent whole-group discussion.

Collection of Raters' DMS

Raters were then provided with the same set of 10 exam samples to rate. Exam samples were in the form of a letter (300-500 words) which 25 EFL students were supposed to write to their parents and informed them of the school initiative program and invited them to take part in the school meeting. Writing samples were obtained from a writing lesson intended for a writing course at the academic level. 25 samples were randomly selected from the exam papers. The rater-participants were told that they would grade as they normally would and that their grades would still count as "real" grades. Also, they were asked to write their feelings at the exact moment they made their score decision. The names of test-takers were omitted from the papers to eliminate the effects of prior familiarity of test-takers and raters. Participants were encouraged to write honestly and were reassured that their comments were going to be held anonymously. This procedure was held for each rater independently. Also, during rating, raters were directed to mark down two scores if they found themselves doubtful between two score levels (indicating doubled scores).

Two weeks after the grading period, raters were contacted and asked to complete an online version of GDMSI questionnaire. They were asked to what extent they agreed with statements regarding their decision-making preferences for important decisions. Responses were given a 5-point Likert scale: strongly disagree, disagree, neutral, agree, and strongly disagree.

Data Analysis Procedures

Analysis of Raters' Decision Making Style

Following techniques associated with grounded theory: the first step was open coding, where a general description was made of each rater's comment, related to feelings, processes, or needs. Unclassifiable comments were those that did not refer to decision making, such as irrelevant personal comments (e.g., about how tired they felt). This initial open coding was followed by axial coding (a division into subcategories). Examples of these subcategories were included, for example, hesitation, what others may think, and gut feelings. These subcategories were then linked to each of Scott and Bruce's (1995) five decision-making styles—hesitation, for example, was linked to avoidant DMS; what others may think, to dependent DMS; and gut feelings, to intuitive DMS. Then, all comments were organized by writing samples, to see if characteristics of the texts themselves brought about any patterns in rater comments. It was concluded, for example, that comments about needing help in making a decision provided

evidence for a tendency toward dependent DMS in a given rater. However, if many raters made similar comments about this particular text, the difficulty in deciding on a grade for it might suggest an idiosyncrasy with this particular text rather than any characteristic of the rater. This consideration of the interaction between text characteristics and the rater's individual characteristics could help to guard against untruthful generalizations about the raters' DMS

Analysis of GDMSI Questionnaire

The data obtained through the questionnaire were analyzed based on the analysis of Likert type questionnaire items. According to the Likert scale, to score the scale, the response categories were weighted: *definitely agree, mostly agree, neither agree nor disagree, mostly disagree,* and *definitely disagree* were scored 5, 4, 3, 2, and 1, respectively. In doing so, the responses of the participants for each item were listed. Then, they were calculated in percentage according to the total number of the participants.

FINDINGS AND DISCUSSION

Findings

Raters' types of DMS

Collection of raters' scoring and their comments revealed that rational comments (31) and intuitive comments (23) were the most numerous, but there were examples of each style, with (13) classified as spontaneous, (15) as dependent, and (4) as avoidant. In table2, the results are tabulated:

Table 2. Each type of DMS

Each type of DMS	Referred by raters
1. Rational DMS	31
2. Intuitive DMS	23
3. Spontaneous DMS	13
4. Dependent DMS	15
5. Avoidant DMS	4

Raters' Doubled Scores

Calculating the percentage of using doubled scores for each rater showed that two of the raters did not write doubled scores at all, whereas 20% of all of rater 2's scores were doubled. Other raters who made use of doubled scores were rater 1 (18% of all scores), rater 3 (12% of all scores), rater 4(2% of all scores).

Table 3. The percentage of doubled scores

Raters	Percentages of doubled scores
1	18
2	20
3	12
4	2
5	-

6	-

Individual Text Analysis Results

Three out of 10 samples showed similar response patterns among the raters. The scores and comments on these letters were examined and a close reading of the letters themselves was undertaken to determine whether there was anything salient that could explain these patterns. It was shown that characteristics of the texts themselves (not just the raters) favored certain decision-making behaviors. Three papers seemed to have solicited a common pattern of comments by the raters. First, the existence of the missing of some key information in the text and raters had to make the decision as to whether this missing information was serious enough to merit awarding or a failing grade. In other words, the paper had characteristics of both levels 1 and 2, making the score decision difficult with this particular text. This characteristic of the text itself could explain the avoidant comments made by the raters. Second, it may be related to text length: one letter was under the word limit at about 260 words. The three of the raters alluded to the text as being too short or lacking in sufficient detail, and three did not allude to this issue at all. The other letter had 540 words. The sheer length of this paper meant perhaps a more systematic treatment was necessary to locate the information needed to decide upon the score. Therefore, these characteristics of the texts themselves could explain the overwhelmingly rational comments made by the raters.

The GDMSI Questionnaire Results

All raters generally agreed with comments associated with intuitive and rational DMS and generally disagreed with comments associated with spontaneous, dependent and avoidant DMS. There were, however, some exceptions: two raters, for example, generally agreed more than they disagreed with comments associated with dependent DMS, and one rater generally agreed with comments associated with spontaneous DMS.

Creation of DMS Profiles for Raters

Patterns of DMS were combined as a way of collecting evidence for individual DMS profiles of raters. Table 4 summarizes the combination of results to create the DMS profiles of each rater.

Table 3. DMS profiles for each rater

Rater	Write-Aloud Comments	Doubled Scores	GDMSI	Dominant DMS for This Context and This Rater
1	R; I; D; A; S	D; A	D	D; A
2	R; I; A	D; A	n/a	A
3	R; I; S	D; A	I; D; S	I; D; S
4	R; I	n/a	R	R
5	R; I; S	n/a	n/a	R; I
6	R; I; S	n/a	R; I; D	R; I

Note: DMS = decision-making style; GDMSI = general decision-making style inventory; R = rational; I = intuitive; D = dependent; A = avoidant; S = spontaneous; n/a = Insufficient evidence found for any style.

Discussion

The results showed the raters' behaviors were not equally the same by all raters in the same rating situations. Likewise, the decision-making behaviors identified by Crossley et al. (2016) were not equal by all raters; raters' decision-making styles indicated a tendency for engaging in some behaviors more than others. For example, one rater indicated an intuitive first reading followed by a systematic second reading. This rating behavior is referred to by Jang et al. (2014) as "a pragmatic two-scan read". The patterns of comments also provided evidence that raters were sometimes coming to their decision as a result of a "pragmatic two-scan read" where an initial scan could not lead to a grade, leading to a second more systematic treatment. In addition, Jang et al. (2014) identified a "provisional mark" technique where some raters (but not all) initially decided upon a grade early in reading a text and continued reading to confirm the provisional mark. Perhaps this technique was used more by raters with a tendency to spontaneous decision making. Raters made use of the "provisional mark" approach—their concerns about coming to a decision about a grade too early. This decision-making behavior might be an artifact of the four-level rating scale, which inherently led to a set of two dichotomous decisions rather than a decision along a continuum of responses (Preston & Colman, 2000). It was found that certain elements of the texts themselves could influence the comments raters made regarding their decision-making behaviors. This pattern was also seen in studies which investigated text characteristics as mentioned by (Plakans & Gebril, 2013; Crossly, Kyle & MacNamara, 2016; Hoyt & Kerns, 1999; Shrestha & Coffin, 2012).

CONCLUSION

Related to the first research question of the study, the results showed there were discrepancies in the DMS of the raters' behaviors. There were not equally the same by all raters in the same rating situations. Although score effects of DMS have yet to be established, it was concluded that despite the exploratory nature of this study, there is potential for the consideration of individual socio-cognitive differences in accounting for some rater variability in scoring. Related to the second question of the study, it was shown that characteristics of the texts themselves (not just the raters) favored certain decision-making behaviors. Accordingly, bringing awareness of the existence of varied individual styles can help raters to try strategies that encourage the expression of certain styles over others, just as individual strategy-use instruction is tied to learning style in language acquisition (Dörnyei, 2005, 2006). In addition, a more individualized approach to rater training is needed. If the individual decision-making style to a great extent is dependent on basic cognitive abilities that are stable and not easily changed, then the decision support systems need to be flexible in order to match the needs of the individual decision makers.

REFERENCES

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

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1–15.

Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Analyzing how teachers evaluate students' writing. *Educational Research and Evaluation*, *13*, 401–434.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality. *Journal of Second Language Writing, 32*, 1-16.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL⁄EFL

writing tasks: A descriptive framework. *The Modern Language Journal*, *86*(1), 67–96.
Dörnyei, Z. (2005). *The psychology of the language learner: Individual differences in second language acquisition*. Mahwah, NJ: Erlbaum.
Dörnyei, Z. (2006). Individual differences in second language acquisition. *AILA Review, 19*, 42–68.
Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, *25*, 155–185.
Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions*. Princeton, NJ: Educational Testing Service.
Gambetti, E., Fabbri, M., Bensi, L., & Tonetti, L. (2007). A contribution to the Italian validation of the General Decision-Making Style Inventory. *Personality and Individual Differences*, *44*, 842–852.
Hoyt, W. T., & Kerns, M. D. (1999). Magnitude and moderators of bias in observer ratings: A meta-analysis. *Psychological Methods*, *4*, 403–424.
Jang, E. E., Wagner, M., & Park. G. (2014). Mixed methods research in language testing and assessment. *Annual Review of Applied Linguistics, 34,*123-153.
Johnson, J. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*, 485–505.
Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing*, *26*, 187–217.
Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(3), 1–31.
Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*, 246–276.
Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*, 54–71.
McDonald, J. D. (2008). Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioral assessments. *Enquire*, *1*, 1–18.
McNamara, T. (1996). *Measuring second language performance*. New York, NY: Longman.
Messick, S. (1984). The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist*, *19*, 59–74.
Milanovic, M., Saville, N., & Shuhong, S. (1996). A study of the behavior of composition markers. In M. Milanovic & N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Annual Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* (pp. 92–114). Cambridge, UK: Cambridge University Press.
Newell, B. R., Lagnado, D. A., & Shanks, D. R. (2007). *Straight choices: The psychology of decision making*. New York, NY: Psychology Press.
Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System*, *30*, 143–154.
Paulhus, D. P. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). San Diego, CA: Academic Press.
Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing, 22*(3), 217-230.
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: Cambridge University Press.
Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica*, *104*, 1–15.
Roohani, A., & Taheri, F. (2015). The effect of portfolio assessment on EFL learners' expository writing ability. *Iranian Journal of Language Testing, 5*(1), 217-230.
Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement*, *55*, 818–831.
Shi, L. (2001). Native-and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing, 18,* 303–325.
Shrestha, P., & Coffin, C. (2012). Dynamic assessment, tutor mediation and academic writing development. *Assessing Writing, 17*(1), 55-70.
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*, 163–182.
Spicer, D. P., & Sadler-Smith, E. (2005). An examination of the general decision making style questionnaire in two UK samples. *Journal of Managerial Psychology*, *20*, 137–149.
Thunholm, P. (2004). Decision-making style: Habit, style, or both? *Personality and Individual Differences*, *36*, 931–944.
Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*, 49–70.
Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Qualitative 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.
Wolfe, E. W., Kao, C-W., & Ranney, M. (1998). Cognitive differences in proficient and non-proficient essay scorers. *Written Communication*, *15*, 465–492.