Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis

Laura K. Varner, Rod D. Roscoe & Danielle S. McNamara

Arizona State University | USA

Abstract: Writing is a necessary skill for success in the classroom and the workplace; yet, many students are failing to develop sufficient skills in this area. One potential problem may stem from a misalignment between students' and teachers' criteria for quality writing. According to the evaluative misalignment hypothesis, students assess their own writing using a different set of criteria from their teachers. In this study, the authors utilize automated textual analyses to examine potential misalignments between students' and teachers' evaluation criteria for writing quality. Specifically, the computational tools Coh-Metrix and Linguistic Inquiry and Word Count (LIWC) are used to examine the relationship between linguistic features and student and teacher ratings of students' prompt-based essays. The study included 126 students who wrote timed, SAT-style essays and assessed their own writing on a scale of 1-6. Teachers also evaluated the essays using the SAT rubric on a scale of 1-6. The results yielded empirical evidence for student-teacher misalignment and advanced our understanding of the nature of students' misalignments. Specifically, teachers were attuned to the linguistic features of the essays at both surface and deep levels of text, whereas students' ratings were related to fewer overall textual features and most closely associated with surface-level features.

Keywords: writing assessment; teacher essay evaluation; self-assessment; computational linguistics; textual analysis

journal of
WRITING RESEARCH

Varner, L.K., Roscoe, R.D., & McNamara, D.S. (2013). Evaluative misalignment of 10th-grade student and teacher criteria for essay quality: An automated textual analysis. *Journal of Writing Research, volume x*(x), 35-59.

Contact: Laura K. Varner, Learning Sciences Institute; Psychology Department; Arizona State University P.O. Box 872111, Tempe, AZ 85287-2111 | U.S.A. – Laura.Varner@asu.edu.

Writing skills are essential for success, both in college and in the workplace (Geiser & Studley, 2001; Powell, 2009; Light, 2001). Unfortunately, national assessments reveal a lack of writing proficiency among high school students. According to the 2007 National Assessment of Educational Progress (NAEP) report, only 33% of United States 8th-graders and 24% of 12th-graders scored at or above the "proficient" level for writing, and only 2% of 8th-graders and 1% of 12th-graders scored at advanced levels for writing. One contributing factor to this problem may be an *evaluative misalignment* between student and teacher writing criteria. Such misalignments have been reported anecdotally in a number of writing studies (Lin, Monroe, & Troia, 2007; Kos & Maslowski, 2001; Hillocks, 1986). According to the evaluative misalignment hypothesis[1] explored here, students do not have an accurate conceptualization of the goals and criteria for quality writing. Their criteria may diverge markedly from the expectations of their instructors. As a result, students produce texts that fail to satisfy the demands of a given genre or assignment and potentially misunderstand teacher evaluations, feedback, and recommendations.

Few studies have sought to explore the phenomenon of evaluative misalignment empirically. A key challenge to such work is the highly subjective nature of writing assessment, including both teachers' assessments of student work and students' own self-assessments. Ratings of essay quality, for example, are influenced by a myriad of linguistic, syntactic, semantic, and rhetorical features of text (Crossley & McNamara, 2011; McNamara, Crossley, & McCarthy, 2010), but human raters may not always be explicitly aware of the impact of such features on their judgments. In this study, we adopt the novel approach of examining the misalignment of student and teacher ratings of essay quality via *automated textual analysis*. Based upon innovations in computational linguistics and computer science, a number of computational tools now exist that enable researchers to quickly and objectively analyze texts at a fine-grained level for diverse textual features. In this research, we use such tools to analyze the underlying text features associated with student and teacher essay ratings. Subsequently, student and teacher misalignment is revealed by the extent to which their writing quality judgments are attuned to overlapping or divergent sets of textual features.

Our analyses utilize two tools: *Coh-Metrix* (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, 2012) and *Linguistic Inquiry and Word Count* software (LIWC; Pennebaker, Booth, & Francis, 2007). Both of these tools provide measures of textual features that include surface-level components of text and measures related to deeper levels of cohesion and comprehension. Coh-Metrix offers a broad analysis of texts, including indices of cohesion, and text difficulty at the lexical, syntactic, structural, and global levels of text (Graesser & McNamara, 2011). LIWC provides several basic text measures (e.g., number of words and paragraphs), along with word-based analyses of lexical, semantic, and thematic properties of text. As Coh-Metrix and LIWC incorporate both overlapping and unique indices, their combination allows us to obtain converging evidence regarding the misalignment of student and teacher

assessments of writing quality. Additionally, these tools may be able to highlight the more nuanced features of students' and teachers' evaluation criteria that may be difficult to detect otherwise.

1. Evaluative Misalignment of Student and Teacher Writing Evaluation Criteria

Students' writing problems cannot be accounted for simply by a lack of instructor expertise or empirical research regarding effective methods for teaching writing. There is ample research on composition instruction, and this work spans a variety of age groups, techniques, first- and second-language proficiencies, and individual differences among students (e.g., Graham & Perin, 2007; Hillocks, 1984; Rogers & Graham, 2008). Exemplifying this research base, Graham and Perin (2007) conducted a meta-analysis of over 120 published studies on writing interventions among students in grades 4-12. Interventions were separated into 11 categories: strategy instruction, summarization, peer assistance, setting product goals, word processing, sentence combining, inquiry, prewriting activities, process writing approach, study of models, and grammar instruction. All intervention types, with the exception of grammar instruction, were found to be beneficial (i.e., moderate to large, positive effect sizes). Rogers and Graham (2008) extended this analysis by conducting a separate meta-analysis of studies involving single-subject design writing interventions. They identified nine writing treatments that were proven to be effective in the writing classroom. In sum, research on the instruction of writing is extensive and has yielded valuable insights and diverse methods for improving the efficacy of writing instruction. Yet, despite the availability of research on the topic, national assessments still reveal that students are struggling to excel in the writing domain.

One caveat for writing instruction is that there is little guarantee that students will internalize or understand the curriculum as intended. Students can ignore key aspects of instruction and fail to develop complete or accurate conceptions of writing goals and criteria. Moreover, when attempting to assess whether their own writing has achieved particular goals, students may judge themselves using limited or faulty criteria. As a result, students may not possess or apply the same evaluation criteria for writing as do their teachers – there may be an *evaluative misalignment* between student and teacher expectations. This misalignment can serve as a barrier to writing instruction, as students may not internalize the same information that teachers are intending to communicate.

Research on the knowledge of writing has observed that many students, particularly struggling writers, indeed display a lack of knowledge and understanding about key writing goals and processes. Wong (1999) describes this metacognitive knowledge as the "awareness of the purpose and process of writing and self-regulation of writing." Such knowledge has been linked to writing proficiency in numerous studies (e.g., Donovan & Smolkin, 2006; Graham, 2006; Wong, 1999). A recurring pattern is that skilled writers are more knowledgeable about writing, particularly the higher-level

principles, such as audience awareness and the development and defense of strong arguments. In contrast, struggling writers tend to think of "better" writing in terms of superficial textual features, such as handwriting, spelling, and punctuation.

For example, Schoonen and de Glopper (1996) investigated the role of writing knowledge in the writing performance of 9th-grade students across three proficiency levels (i.e., low, average, proficient). They instructed older students to write letters to younger students describing the components and features that comprise good writing. Their analysis of these documents indicated that proficient writers focused more on the organization of ideas, whereas less-proficient writers focused on surface-level features (e.g., spelling and grammar). Similarly, Lin, Monroe, and Troia (2007) investigated the development of writing knowledge in 2nd- and 8th-grade struggling and typical writers. In an interview setting, students were asked questions regarding their perspectives on the purposes, processes, and products involved in writing. As above, they found that younger writers and struggling writers placed a strong emphasis on surface-level features of writing (i.e., handwriting, spelling, and sentence structure). However, older writers and successful writers were able to discuss more global aspects of writing, such as audience awareness and communication of meaning. For instance, when asked to describe the purpose of writing, one elementary-aged, struggling writer responded, "Because they want us to learn…" However, a successful middle school student responded, "When we get our job, we need to know how to write and get our recommendations. To [sic] prepared everything for life, you need to write." Similarly, when asked about the processes that good writers employ, an elementary student responded, "They put period," whereas middle school students discussed audience awareness, "They think about who's reading it…Authors have to write so that others can feel like it's the author talking and feel connected." Overall, the results of such research show that as writers develop and expand their knowledge, they move from a shallow and local understanding of writing to a deeper and more global understanding of writing. Nonetheless, across grade levels, a number of students continue to display a misunderstanding of the goals and features of effective writing.

An additional challenge for developing writers resides in the metacognitive processes of self-assessment necessary to judge one's own work. That is, students may display poor writing proficiency because of a failure or inability to accurately assess the quality of their writing. Compared to teachers, who assess students' essays on numerous dimensions and multiple levels of text (e.g., words, organization, meaning, style, and so on), students may apply only a limited set of criteria (e.g., only lexical features) or an incorrect set of criteria (e.g., judging the readability of a text without considering audience needs). More generally, students may struggle with the process of objectively critiquing their own work or with the appreciation of how their outcomes may have fallen short of their intentions.

Research on the metacognitive processes of writing has focused on the benefits and accuracy of students' self-assessment (Andrade & du Boulay, 2003; Andrade, Du, & Wang, 2008; Graham & Perin, 2007; Hillocks, 1986; Ross, Rolheiser, & Hogaboam-

Gray, 1999). For instance, Ross and colleagues (1999) investigated the accuracy and benefits of student self-assessments in writing. They conducted a classroom study that investigated the effects of a training program designed to teach elementary school (4th-6th grades) students how to self-evaluate their narrative writing. Additionally, they explored the direct benefits of this training on students' self-assessment accuracy. As part of the training program, the students helped to define the evaluation criteria for their own writing and were given instructions on how to apply these criteria to their own work. In addition, they were provided with teacher feedback on their self-evaluations. Overall, the treatment condition led to greater gains in the quality of the students' narrative writing than the control condition, particularly for struggling students. Further, the training led to increased precision of the students' self-assessments; in particular, the treatment group was less likely to overestimate the success of their writing performance.

Similarly, Andrade and colleagues (2008) examined the effects of self-assessment on elementary school students' writing performance in a writing class. In their study, students were first presented with a model essay intended to generate classroom discussion about its strengths and weaknesses. Following the discussion, the students collaborated to generate a list of criteria associated with quality stories and essays. Finally, students received a rubric and were asked to self-assess their own writing based on its criteria. The results revealed that the treatment condition led to higher essay scores over the control condition after controlling for prior English ability. Thus, by providing students with explicit instruction of writing criteria, as well as specific guidance on implementing these criteria, students' performance on writing assignments improved. Overall, these studies reveal the important role of the metacognitive process of self-assessment in writing development. Previous research points to improvements in writing performance as a result of students' self-assessments. Nonetheless, little is known about the characteristics and accuracy of these evaluations. Hence, writing research may benefit from an increased understanding not only of the effects of self-evaluation on writing performance, but also of the features in text that influence students' criterion for quality writing.

Although not explicitly related to writing, prior research has revealed that students are generally inaccurate in their self-assessments of performance. As these findings have spanned numerous domains, it is probably safe to assume that students exhibit these same inaccuracies when assessing their own writing. In a widespread review of the literature, covering a number of domains in higher education, such as law, medicine, engineering, and psychology, Falchikov and Boud (1989) found that college student and teacher assessments of performance tended to yield only moderate correlations around $r = .39$, and students' predictions of their anticipated grades exceeded teachers' assigned grades about 68% of the time. The accuracy of self-assessments was slightly higher in well-defined domains, such as engineering, and among students taking advanced courses (i.e., higher-performing students). More recently, Tousignant and DesMarchais (2002) evaluated the accuracy of medical students' self-assessments in a

problem-based learning program. Self-assessments for three tests were obtained both prior to an examination and immediately following the examination. Results indicated that students were inaccurate in predicting their performance prior to the tests, with only weak correlations between students' predictions and their actual test scores (*r* ranged from 0.04 to 0.24 for the three tests). Immediately after completing the exam, students' self-assessment accuracy improved slightly, but the correlation with actual scores remained low ($r = 0.26$ to 0.33). In sum, available research suggests that many students apply limited or faulty criteria in the assessment of their performance, and this misalignment likely extends to self-assessments of writing quality.

This misalignment may lead to serious consequences for student writers. First, students may produce texts that fail to meet particular writing goals (e.g., persuasive essays that lack evidence) or that achieve those goals inappropriately (e.g., evidence that is overly subjective and speculative rather than objective and factual). Second, students may find it difficult to understand or apply the feedback received from teachers. For instance, a teacher may urge the student to "pay closer attention to appropriate word choice," with the intention that the student should employ more diverse, precise, and descriptive wording. However, the student may interpret feedback about word choice to mean they should "use bigger words to impress the teacher." In this case, the revised essay would be more likely to contain more multisyllabic words, perhaps used inappropriately, and continue to display problems of word usage. Similarly, misalignment may contribute to students' difficulties in assessing their own writing. Students may make inaccurate or overly positive judgments of their own work, because they are misapplying the criteria needed to make those judgments. Thus, not only may student and teacher misalignment directly contribute to students' poor writing, it may hinder the very communicative and metacognitive processes needed for students to learn and improve.

We propose that a better understanding of students' writing proficiency calls for further research on how students' beliefs and perceptions of "good" writing diverge from teachers' conceptions. It is possible that teachers' perceptions of essays are colored by subtle influences that are underlying a more specific rubric. Thus, even if provided with a scoring rubric, students' perceptions of the rubric components may still diverge markedly from those of their teachers. In this study, we address two principal questions concerning student and teacher evaluative misalignment. First, how are teachers' ratings of essay quality related to the linguistic features of student essays, such as syntax, cohesion, or emotional word use? Second, how do students' self-assessments of their own writing diverge from teachers' ratings? To address these questions, we analyzed student essays using two automated text analysis tools: Coh-Metrix and LIWC. These tools have the power to provide numerous measures of the nuanced text features in students' essay. Thus, the use of these tools may help to highlight some of the less obvious or explicit aspects of students' and teachers' evaluation criteria.

2. Automated Textual Analyses to Explore Student and Teacher Evaluative Misalignment

In this study, we conducted two automated analyses of student essays in relation to student and teacher quality ratings. These analyses use Coh-Metrix and LIWC, both of which have been widely used in previous studies on text and discourse.

Coh-Metrix

Coh-Metrix is a computational tool that analyzes text on a variety of textual dimensions relating to cohesion, and text difficulty (Graesser et al., 2004; McNamara & Graesser, 2012). A sampling of key indices provided by Coh-Metrix is discussed below.

Basic Text Measures

Coh-Metrix assesses fundamental properties of text, such as the total number of words, words per sentence, incidence of parts of speech, number of paragraphs, and so on. Many of these measures have been shown to be important predictors of expert ratings of essay quality, and capable of discriminating essays based on the grade levels of the writers (Crossley & McNamara, 2011; Crossley, Weston, McLain-Sullivan, & McNamara, 2011).

Lexical Indices

Coh-Metrix also assesses text using many word-level measures. Many of these indices are calculated through the WordNet computational lexical database (Fellbaum, 1998), which is organized into lexical networks based upon connections between related concepts. Example Coh-Metrix indices provided by WordNet include *polysemy* (the number of senses attributed to a word) and *hypernymy* (the specificity of a word). Other Coh-Metrix lexical indices are obtained using the MRC Psycholinguistic Database (Wilson, 1988). The MRC database is comprised of over 150,000 words that have been rated along 26 possible linguistic and psycholinguistic dimensions. For example, *age of acquisition* refers to the fact that some words appear in a child's language before others. Additionally, *word familiarity* provides a measure of how familiar printed words seem to a typical person.

Cohesion Indices

Coh-Metrix provides over 50 measures of textual cohesion. For example, one way to establish cohesion is through the use of connective phrases that make conceptual relations explicit (e.g., the phrase *on the other hand* can be used to signal the presentation of counterevidence or opposing viewpoints). Use of connectives can be assessed based on valence (i.e., whether the connective phrases are positive or negative) and based on functional categories (Halliday & Hasan, 1976; Louwerse, 2001). Such categories include clarifying connectives (e.g., *in other words*), additive connectives (e.g., *moreover*), temporal connectives (e.g., *subsequently*), and causal connectives (e.g., *consequently*).

Another Coh-Metrix cohesion index is lexical overlap, which includes *content word overlap*, *argument overlap*, *stem overlap*, and *noun overlap* (see McNamara, Louwerse, McCarthy, & Graesser, 2010, for more detail). Content word overlap measures the proportion of content words shared between two sentences. Argument, stem, and noun overlap are binary measures of the frequency that two sentences share nouns and pronouns, stems, and identical nouns, respectively.

Latent Semantic Analysis (LSA)

Coh-Metrix uses LSA to analyze text cohesion at the semantic level (McNamara, Cai, & Louwerse, 2007). LSA uses a statistical method to reduce a large matrix of word co-occurrences into approximately 100-500 dimensions and is typically used to compute the similarity between sentences or between a sentence and an entire passage (Landauer, McNamara, Dennis, & Kintsch, 2007). Coh-Metrix provides multiple LSA measures, including *sentence-to-paragraph*, *sentence-to-text*, *paragraph-to-paragraph*, and *paragraph-to-text* to measure the semantic co-referentiality of texts. These measures reflect the *semantic* similarities, rather than *surface* similarities, that occur at the sentence, paragraph, and overall text levels.

Validity of Coh-Metrix

Coh-Metrix has been shown to be an informative and reliable text analysis tool in a number of prior studies. One line of studies has focused on examining the linguistic features of high-quality essays (Crossley & McNamara, 2010; Crossley & McNamara, 2011; Crossley & McNamara, 2012; McNamara, et al., 2010). Crossley and McNamara (2011), for instance, used Coh-Metrix to investigate the role of cohesion and coherence in expert evaluations of essay quality. They found that coherence as assessed by expert raters was an important characteristic of high-quality essays, but that coherence was marked by an *absence* of cohesive cues in the text, rather than a presence. Similarly, McNamara and colleagues (2010) used Coh-Metrix to determine the linguistic differences in essays rated *high* and *low* by experts. The indices most predictive of essay quality were *syntactic complexity* (number of words before the main verb in a sentence), *lexical diversity*, and *word frequency*. Their results indicated that expert judgments of essay quality were sensitive to linguistic features associated with text difficulty and a refined use of language. Indeed, these results have emerged across a number of corpora for both first and second language writers (for a review of the studies, see Crossley & McNamara, 2011).

In addition, Coh-Metrix has been also been used to assess student paraphrases (Rus, Lintean, Graesser, & McNamara, 2009), assess paragraph quality in student essays (Roscoe, Crossley, Weston, & McNamara, 2011), and to detect grade level of student writers (Crossley, Weston, et al., 2011). Thus, Coh-Metrix has been established as a useful tool that is capable of detecting subtle differences within student writing.

Linguistic Inquiry and Word Count (LIWC)

LIWC is a text analysis tool that uses categorical word dictionaries to provide information corresponding to thematic and rhetorical language use (Pennebaker, Booth, & Francis, 2007). The categorical dictionaries are hierarchical and each consists of a list of words that are related to a particular concept or theme. LIWC contains approximately 4,500 words and word stems across a number of dictionaries; select indices are described in detail below.

Basic Text Measures

Like Coh-Metrix, LIWC provides measures of basic textual information. For a given text, LIWC calculates word count, incidence of punctuation, frequency count of words containing more than six letters, incidence for some parts of speech, incidence of verb tenses, and so on. These text measures provide information about surface-level text features.

Psychological Processes

The psychological process categories of LIWC relate to social, affective, cognitive, perceptual, and biological processes of humans. Each category is further divided into sub-dictionaries that reflect specific characteristics of these processes. For instance, the *perceptual processes* category contains words related to sensory and perceptual concepts, which are divided into three sub-dictionaries: *see*, *hear*, and *feel*. Each sub-dictionary contains numerous words related to that specific category. For instance, the sub-dictionary *see* contains words, such as 'gaze' and 'bright,' whereas the *hear* sub-dictionary contains words, such as 'ring' and 'whisper.' The *cognitive processes* category contains numerous sub-dictionaries, such as *insight* (e.g., 'aware' and 'notice'), *certainty* (e.g., 'absolute' and 'factual'), and *exclusion words* (e.g., 'if' and 'just'). The more frequent occurrence of words within a category is assumed to reveal information about the semantic content of an essay. For instance, a high incidence of *social* words can reflect a text that relates to broader social or cultural concerns, but a high incidence of *cognitive* words signals a more opinionated and analytical text.

Personal Concerns

LIWC also provides measures of personal issues that reflect the theme of a text. Some examples of these sub-dictionaries are *work* (e.g., 'boss' and 'publish'), *leisure* (e.g., 'bath' and 'read'), and *money* (e.g., 'account' and 'portfolio'). These measures provide an understanding of the specific themes and topics that are being discussed in a given text.

Validity of LIWC

LIWC has been employed in numerous studies to measure the emotional, cognitive, structural, and process components present in a text. Many of the studies have focused on discriminating emotional states and personality features of the speakers within a given text. For example, Hancock, Landrigan, and Silver (2007) investigated the ways

that individuals express emotion during text-based communication. They found that individuals relied on four strategies to express levels of happiness: negations, negative affect terms, punctuation, and verbosity. All indices except verbosity were calculated using the LIWC software. Gill, French, Gergle, and Oberlander (2008) similarly analyzed the emotional language use of authors in blog entries. They collected blog posts of 50 and 200 words that had been previously coded by expert and naïve raters. The results showed that the "angry" authors used more affective language and negative affect words, whereas "happy" authors used more positive affect words. Moreover, they found that the LIWC results were consistent with human ratings.

In addition to text analyses, LIWC has been used in various ways to understand natural speech. Pennebaker, Mayne, and Francis (1997) found that LIWC indices successfully detected improvements in physical and mental health following traumatic events, and Hancock, Curry, Goorha, and Woodworth (2008) used LIWC to analyze the differences between deceptive and truthful conversations in an instant messaging environment. The results of these studies show that LIWC has the potential to detect changes in individuals' language use, as well as distinguish between groups based on language use. In sum, LIWC has been established as a useful tool that provides information about themes, content, and genre within a text.

3. Method

In this study, we investigate the degree to which linguistic features of text, as measured by Coh-Metrix and LIWC, are predictive of student and teacher ratings of essay quality. Through the use of automated text analysis tools, we aim to identify some of the more nuanced features of students' essays that potentially affect student and teacher ratings of essay quality. Throughout our analyses, alignment is defined as the extent to which student ratings are predicted by, or related to, the same features as teacher ratings. Evaluative misalignment, on the other hand, is defined as the extent to which student ratings are predicted by *fewer* or *different* linguistic indices than teacher ratings. In this manner, our analysis potentially reveals both the complexity and the actual content of student and teacher misalignments.

3.1 Participants

Participants were 126 students enrolled in tenth-grade English courses (approximately 15-16 years of age) at a high school in the Washington, DC area. The school enrolled over 2,400 students, with a student population comprised of 49.0% female students, with 22.3% Asian, 4.2% Black, 9.0% Hispanic, and 59.9% White students. Only 7.0% of the students were described as limited English proficiency, and 10.9% qualified for free or reduced-price meals. As students typically begin to prepare for college entrance exams (including writing assessments) around grade 10, this age group provides a representative sample of students who may be strongly affected by evaluative misalignment.

3.2 Essay Corpus

Each participating student wrote a timed (25 minute), prompt-based, argumentative essay. The essays were written in response to an SAT-style prompt selected by the teachers:

A sense of happiness and fulfillment, not personal gain, is the best motivation and reward for one's achievements. Expecting a reward of wealth or recognition for achieving a goal can lead to disappointment and frustration. If we want to be happy in what we do in life, we should not seek achievement for the sake of winning wealth and fame. The personal satisfaction of a job well done is its own reward.

Are people motivated to achieve by personal satisfaction rather than by money or fame? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

3.3 Essay Evaluations

3.3.1 Teacher Evaluations

Several weeks after students completed their essays, teachers from several classrooms exchanged essays such that no teacher graded his or her own students' work. Teachers rated student essays using the scoring rubric published by the SAT and College Board (Camara, 2003), which resulted in a single, holistic score on a 1-to-6 scale. A "1" is the lowest rating that an essay can receive and a "6" is the highest. Teacher scores had an average rating of 3.67 (SD = 1.01) and were normally distributed.

3.3.2 Student Evaluations

Approximately one week after writing their essays, students were asked to predict the score of their essays. As with teachers, students provided a rating on a scale of 1-to-6, with "1" being the lowest score and "6" being the highest. As this was an in-class activity, there was insufficient time to train students on the complete SAT rubric. Thus, students were given a simplified version of the rubric to assess their own essays. The survey provided both qualitative and quantitative choices for student ratings. For example, the highest rating students could choose stated, "My essay was '*Great*' and will get a 6 out of 6 (highest score)" whereas the lowest rating students could choose stated, "My essay was '*Poor*' and will get a 1 out of 6 (lowest score)."

Students' self-assessments were normally distributed and had an average score of 4.04 with a standard deviation of 0.82. They were only moderately and positively correlated with teacher scores (r = .26, p < .01). Relative to teachers, students tended to slightly overestimate their scores; t (125) = 3.86, p < .001, which represents a small to moderate effect size (d = .40). Overall, the pattern of means, and the low correlation between student and teacher scores, suggest a potential misalignment between the students' and teachers' expectations for the essay quality. We used the automated text

analysis tools Coh-Metrix and LIWC to further explore the *characteristics* of this misalignment.

3.3.3 Statistical Analyses

To examine the misalignment between student and teacher ratings of essay quality, we conducted correlation and regression analyses using essay ratings and textual features. Specifically, we examined how and whether specific linguistic text features were correlated with, or predictive of, student and teacher ratings. Analyses were conducted for each set of scores (i.e., teacher and student ratings) and each computational tool (i.e., Coh-Metrix and LIWC) separately. First, correlations were calculated between text indices provided by the automated tool and the essay scores. The pattern of correlations was examined for indices related to scores at the $p < .05$ level, and the variables with the strongest relations to the scores were included in the regression model. To address multicollinearity, when variables correlated with each other above $r = .70$, the variable with the lowest relation to the student and teacher scores was removed. To avoid over-fitting the model, we chose a ratio of 15 essays to 1 predictor, which allowed 8 indices to be entered, given that there were 126 essays included in the analyses.

4. Results

4.1 Coh-Metrix Analyses

4.1.1 Teacher Ratings

Correlations were calculated between the Coh-Metrix indices and the teacher scores. As shown in Table 1, 12 variables were significantly correlated with the teachers' scores. Corroborating past research with expert raters (Crossley & McNamara, 2011; McNamara et al., 2010), these correlations indicate that teachers' ratings were influenced by aspects of essay elaboration, essay organization, skillful use of language, and a lack of cohesion. For example, the *total number of words*, *total number of sentences*, and *total number of paragraphs* indices are indicators of the overall length and structure of essays. In particular, the *total number of words* and *total number of sentences* provide measures of the length of the essay, and longer essays can be indicative of the elaboration of ideas and examples. Similarly, the total *number of paragraphs* broadly measures the organization of an essay. Essays with more paragraphs may possess clearer demarcations between separate ideas, especially when compared to the commonplace "one-paragraph essays" written by novice writers. Overall, teachers rated longer and organized essays more highly, which is to be expected. This finding indicates that teacher ratings were related to the elaboration and organization of ideas in students' essays.

Table 1. Correlations between Teacher Ratings and Coh-Metrix Variables

Coh-Metrix Variable	Correlation with Teacher Ratings
Total number of words	.337**
Familiarity of content words	-.294**
Total number of paragraphs	.240**
Incidence of locational entities	.239**
LSA paragraph to paragraph	-.226*
Age of acquisition of content words	.224*
Lexical diversity (VOCD)	.218*
Polysemy of words	-.207*
Noun incidence	.205*
Content word overlap	-.194*
Total number of sentences	.192*
Frequency of content words	-.192*

Note. *$p < .05$; ** $p < .01$

The teachers' essay ratings were also correlated with lexical features of the essays. The *familiarity and frequency of content words*, *age of acquisition of content words*, and *polysemy of words* indicate more sophisticated language use. Word *familiarity*, word *frequency*, and *age of acquisition of content words* indicate that a given essay is composed of more uncommon and sophisticated words. *Word polysemy* is indicative of the degree of ambiguous language utilized in an essay. Thus, teachers' ratings were related to students' use of specific words, which indicate more precise descriptions of ideas and concepts. Albeit somewhat weak, the correlations with these variables suggest, not surprisingly, that teachers are sensitive to vocabulary use, particularly students' use of less familiar and abstract words. Teachers' scores were also influenced in part by more concrete language, as measured by their relationship to *incidence of locational entities* and *noun incidence*. This suggests that teachers prefer persuasive essays that contain more nouns, particularly those that refer to a specific location (e.g., house, store, Georgia). These measures are indicative of more concrete language, as they refer to specific places and objects.

Finally, variables such as *lexical diversity*, *LSA paragraph-to-paragraph*, and *content word overlap* reveal a benefit of using diverse language and developing low overt text cohesion. The correlation analyses indicate that teachers associate low-cohesion essays with higher overall essay quality (as found in previous research on expert ratings: Crossley & McNamara, 2010; Crossley & McNamara, 2011; Crossley & McNamara, 2012; McNamara, et al., 2010). It is possible that high-cohesion essays are too reliant on repetitive vocabulary and examples to connect ideas. On the other hand, low-cohesion essays may rely on deeper (i.e., not surface-level) arguments structures to develop text coherence. This finding is line with prior research on text comprehension,

which suggests that cohesive devices can support or hinder the development of coherent text representations depending on readers' level of prior knowledge (McNamara & Magliano, 2009; O'Reilly & McNamara, 2007).

Overall, these results confirm, and allow us to document, common intuitions about teachers' criteria. The teachers' ratings are related to numerous aspects of student essays, ranging from lower- to higher-level features. Namely, teachers seem most sensitive to the elaboration and organization of ideas, sophisticated vocabulary and language use, and a reduced cohesion.

A regression analysis was conducted to assess how and whether the above variables predicted teachers' essay ratings. All variables were tested for multicollinearity ($r > .70$) and two variables (*frequency of content words* and *total number of sentences*) were eliminated due to a strong relationship to other variables. The analysis yielded a significant model, $F(8, 125) = 6.89$, $p < .001$; $R^2 = .32$. The significant predictors in the model were *total number of words* ($B = .27$, $p < .01$) and *LSA paragraph-to-paragraph* ($B = -.29$, $p < .001$). Two additional variables in the model were statistically significant, if tested one-sided: *number of paragraphs* ($B = .15$, $p = .097$) and *word polysemy* ($B = -.16$, $p = .07$). These results suggest that the linguistic features most predictive of the teacher ratings in this sample were related to essay elaboration (i.e., length of the essays), followed by less abstract wording and reduced cohesion. In general, the Coh-Metrix analysis reveals that teacher quality ratings are associated with numerous essay components, including lexical, syntactic, and cohesive features.

4.1.2 Student Ratings

As shown in Table 2, seven variables were significantly correlated with the students' scores. Importantly, these correlations reveal that students' ratings were partially associated with *different* features of the essay than were teachers' ratings, indicating some degree of misalignment in the criteria.

Table 2. Correlations between Student Ratings and Coh-Metrix Variables

Coh-Metrix Variable	Correlation with Student Ratings
Second person pronoun incidence score	-.249**
Noun incidence	.226*
Age of acquisition of content words	.226*
Lexical Density	.221*
LSA verb overlap	.180*
Incidence of locational entities	.180*
Average syllables per word	.178*

Note .*$p < .05$; ** $p < .01$

The correlations indicate that students were most sensitive to the level of personalization in their essays along with strong vocabulary and language use. In particular, the *incidence of 2nd person pronouns* is indicative of the level of personal or familiar language in students' essays. Essays that contain a high incidence of second person pronouns often rely too heavily on personal stories and anecdotes as examples and evidence statements. Accordingly, students seemed to be aware that this overly familiar language potentially reduces the quality of their essays. In addition, students rated their essays more positively when they had a higher *mean number of syllables per word* and *mean age of acquisition of words*. These variables represent the length of the words used (i.e., the number of syllables), as well as the sophistication of the words (i.e., the age at which the vocabulary words are typically acquired). As academic and professional writing typically contain more sophisticated vocabulary, students may have perceived their essays to be of higher quality if they incorporated longer and less common words.

Student essay ratings were also positively related to the incidence of concrete language. In particular, *noun incidence*, *lexical density* (proportion of function words in the text), and *incidence of locational entities* represent more concrete language, which provides more examples and facts. Essays with a higher incidence of nouns and a higher proportion of lexical items typically contain more concrete and grounded language, as they are less reliant on function words and verbs. Further, the *incidence of location entities* provides a count of the nouns that refer to a specific location (e.g., Arizona or house). The positive relation between students' ratings and these location nouns indicates a preference for specific (i.e., not abstract) facts and examples. Overall, when assigning quality ratings, students seemed to be attuned to the level of specific, concrete language use.

Finally, students seemed to rate their essays more highly if they were more cohesive, as suggested by the *LSA verb overlap* measure. Specifically, this measure indicates that students assigned higher quality ratings when their essays were more semantically connected. This result is contrary to the teachers' ratings, and suggests that students are unaware of the level of cohesion appropriate for high essay quality. Because students are often taught to develop clearly connected ideas, their evaluation criteria may require explicit cohesive devices in the essays. Thus, they may provide higher overall ratings to their essays when they contain these overt cohesion features. Overall, students' essay ratings were most highly associated with lexical features and concrete language use. This is somewhat in contrast to teacher ratings, which were associated with a wider variety of indices, including organization and elaboration of ideas.

A regression analysis was conducted to assess how and whether the correlated variables predicted students' essay ratings. All variables were assessed for multicollinearity ($r > .70$) and one variable (*average syllables per word*) was eliminated due to a strong relationship with other variables. Students' predicted scores were regressed onto the six remaining variables in a linear regression, yielding a significant

model, $F(6, 125) = 3.47$, $p < .01$; $R^2 = .15$. No single variable was a significant predictor in the model, although three variables approached significance: *age of acquisition of content words* ($B = .18$, p = .05), *LSA verb overlap* ($B = .15$, $p = .095$), and *incidence of locational entities* ($B = .16$, p = .08). The results of the regression suggest that students' ratings of essay quality are less systematically related to the linguistic features of essays than teachers. Thus, they are utilizing an incomplete or different set of criteria when providing self-assessments of essays. While some linguistic variables are associated with students' overall ratings, they are not strongly related to or predictive of student ratings. Thus, students may focus on other aspects of their essays when assigning ratings, such as the theme or content, or even how they felt emotionally while they wrote it.

4.1.3 Summary of Coh-Metrix Analysis

An analysis of textual features related to students' and teachers' ratings of essays revealed that there was, indeed, misalignment in the evaluation criteria. Overall, teacher ratings were more strongly related to the Coh-Metrix variables with an $R^2 = .32$, compared to the student ratings, which reported an $R^2 = .15$. In addition, teachers' ratings were significantly correlated with a larger number of indices than were students' ratings. This is unsurprising, given that teachers necessarily have a broader understanding of how multiple text features interact to produce quality essays. For example, students seemed attuned to word length (e.g., *number of syllables*) as an indicator of lexical sophistication, whereas teachers attended to whether the words were less common and more precise. Indeed teachers have a more thorough understanding of the different features related to essay quality at both superficial and deep levels. One explanation for this low relationship between student ratings and linguistic variables is that students are paying attention to different levels of the content of their essays (e.g., themes or genres). Thus, our subsequent analyses also evaluated teacher and student misalignment using LIWC, which places a stronger emphasis on the textual features that related to thematic or genre content. Because LIWC is a similar, yet more thematic and idea-based tool, this second analysis serves as a triangulation, providing converging evidence for evaluative misalignment.

4.2 LIWC Analyses

4.2.1 Teacher Ratings

As shown in Table 3, 17 LIWC indices were significantly correlated with the teachers' ratings. The results in Table 3 indicate that teachers' ratings were most strongly related to essay elaboration, vocabulary strength, and the skilled use of language in student essays. Not surprisingly, teachers seemed to be most attuned to the length or elaboration of the student essays, as evidenced by the correlation with *word count*. Essays composed of more words often contain more detailed elaborations of arguments

and examples. Similarly, the teachers rated essays more highly when they included longer vocabulary words (*words containing more than 6 letters*). Because strong vocabulary is typically associated with longer words, this correlation indicates that teachers score higher essays with more sophisticated word choices.

Table 3. Correlations between Teacher Ratings and LIWC Variables

LIWC Variable	Correlation with Teacher Ratings
Word count	.333**
Cognitive mechanisms	-.282**
Tentative words	-.278**
Future tense words	-.275**
Present tense words	-.272**
Verbs	-.261**
Certainty words	.243**
Third person plural pronouns	.239**
Human words	-.233**
Exclusion words	-.223*
Insight words	-.217*
Words containing more than 6 letters	.213*
Feeling words	-.209*
Auxiliary verbs	-.203*
Past tense words	.202*
Perception Words	-.181*

Note *$p < .05$; ** $p < .01$

Other correlated variables suggested that teachers were sensitive to the use of objective and fact-based language in student essays. For example, hypothetical language (*exclusion words*), hedging language (*tentative words*), emotional language (*feeling words*), and other subjective words (*insight words*, *perception words*, and *cognitive mechanisms*) had a negative association with teacher ratings. The *exclusion* words and *tentative words* measure uncertain language, as they represent ungrounded and hesitant word choices. For instance, the *tentative words* category includes words such as "might," "possibly," and "could," which establish weaker arguments and examples. On the other hand, objective and confident language (*certainty words* and *third person plural pronouns*) was associated with higher ratings by teachers. Essays with objective and confident word choices may develop stronger and more sophisticated arguments. Overall, as one would expect, teachers are attentive to the strength and objectivity of the language that students used when developing their arguments.

A regression analysis was conducted to examine the extent to which LIWC variables predicted teacher ratings. The indices were checked for multicollinearity ($r > .70$) and two variables (*perception words* and *verbs*) were eliminated due to a high relationship to other variables. The regression yielded a significant model, F (8, 125) = 5.81, $p < .01$; $R^2 = .28$, with two significant predictors: *word count* ($B = .26$, $p < .01$) and *cognitive mechanisms* ($B = -.21$, $p < .05$); and one variable was statistically significant, if tested one-sided: *future tense words* ($B = -.16$, $p = .08$). The results of this analysis suggest that teachers were most concerned with essay elaboration and a more objective use of language; longer essays were most likely to receive a high rating by teachers. Additionally, objective language, as indicated by a lack of subjective words (e.g., *think*, *should*, and *maybe*) was a factor in teachers' assignment of high ratings to student essays.

4.2.2 Student Ratings

Nine variables were significantly correlated with student scores (see Table 4). The results indicate that student ratings were more highly correlated with linguistic variables related to three major factors: objectivity of language, level of confidence expressed in the essays, and vocabulary strength.

Table 4. Correlations between Student Ratings and LIWC Variables

LIWC Variable	Correlation with Student Ratings
Third person plural pronouns	-.270**
Tentative words	-.261**
Second person pronouns	-.251**
Third person singular pronouns	.226*
Certainty words	.200*
Sadness words	-.196*
Function words	-.193*
Words containing more than 6 letters	.187*
Present tense verbs	-.185*

Note. *$p < .05$; ** $p < .01$

For example, students' essay ratings were also related to the objectivity of their language. Objective language was measured by a positive correlation with *third person singular pronouns*, as well as a negative correlation with *second person pronouns*. Second person pronouns indicate a higher incidence of personal and familiar language, in contrast to third person pronouns, which are representative of more objective language. Thus, students were somewhat aligned with teachers in the focus on strong vocabulary and objective language use when rating essays.

Similarly, students' ratings were dependent on the level of confidence expressed in their essays. When arguments were developed with confidence words (*certainty words*) and with a low incidence of hedges (*tentative words*), students were more likely to rate their essays higher. Essays with more confident language typically develop stronger arguments and examples. Thus, students' perceptions of their writing quality are somewhat aligned with teachers' ratings regarding the use of confident language in the essays.

Finally, vocabulary strength (*words containing more than 6 letters*) was related to high student ratings. Thus, students may feel more confident in quality of their writing if they utilize more complex vocabulary. Overall, the LIWC analysis revealed that student ratings were highly associated with the strength of vocabulary, language use, and use of confident language in the essays.

A regression analysis was conducted to assess which LIWC variables, if any, predicted student essay ratings. The variables were assessed for multicollinearity, but no two variables were correlated above the .70 threshold. The regression yielded a significant model, $F(8, 125) = 4.02$, $p < .001$; $R^2 = .22$, with two significant predictors: *third person plural pronouns* ($B = -.21$, $p < .01$) and *second person pronouns* ($B = -.20$, $p < .05$), and one predictor that approached significance: *certainty words* ($B = .15$, $p = .08$). The positive relation to *third person plural pronouns* and the negative relation to *second person pronouns* suggest that students were sensitive to the level of personalization in the essays when providing quality ratings. Additionally, the predictor, *certainty words,* implies that the level of confidence expressed in the essays influenced student self-assessments of writing quality.

4.2.3 Summary of LIWC Analysis

The linguistic features captured by the LIWC measures provide further information about the characteristics of the misalignment between student and teacher evaluation criteria. Overall, the LIWC indices were able to capture approximately one-fourth of the variance in both teacher ($R^2 = .28$) and student ($R^2 = .22$) ratings of the essays. In addition, LIWC analyses revealed a partial alignment between student and teacher evaluation criteria. Similar to the Coh-Metrix analysis, the LIWC analysis suggested that students' and teachers' ratings relied somewhat on sophisticated vocabulary and objective and confident language use. When students expressed their ideas confidently (*certainty words*) and avoided personalized language (*third person plural pronouns; second person pronouns*), both students and teachers assigned higher quality ratings. The LIWC indices, however, also demonstrated areas of misalignment between the students and teachers. Specifically, the teachers were sensitive to deeper issues and strong language use than were the students. For instance, in addition to the relationship to objective and confident language use, teacher ratings were also associated with a lack of hypothetical, emotional, and perceptual language. This suggests that teachers were better able to assess texts based on a larger number of textual features than the

students. The results of the LIWC analyses suggest that students and teachers were, at least partially, misaligned in their criteria for quality essays. As one would expect, teachers have a more expansive conceptualization of the different features that interact to produce quality essays.

5. Discussion

The results of this study indicate that teachers do indeed assess student essays on a variety of linguistic measures at surface- and deep-levels of text. In contrast, students' ratings are associated with a smaller subset of variables, namely surface-level features. Thus, our results are in line with the hypothesis that there is an evaluative misalignment between the criteria of students and teachers. Despite the importance of students understanding the feedback that they receive on their writing, the nature of potential misalignments between students' and teachers' writing evaluation criteria has not been examined in the composition literature. In this study, we explored one assumption of the misalignment hypothesis using a textual analysis of students' and teachers' assessments of SAT-style essays. Specifically, we took the novel approach of investigating how misalignments manifest themselves in terms of linguistic textual properties.

One substantial contribution of this study is the analysis of the linguistic features that most accurately predict teacher ratings of essay quality. Although researchers have investigated the textual features related to *expert* ratings, no study to our knowledge has explored the features that characterize *teacher* ratings. A second contribution is the comparison of teacher ratings to students' self-assessments of their own writing. Specifically, our analyses reveal the areas in which students' and teachers' evaluation criteria are disparate. With this analysis, we were able to establish linguistic features that characterize students' and teachers' evaluations of essays and confirm the presence of student-teacher misalignment in essay evaluation.

5.1 Teacher Ratings

The results of this study provide an extensive analysis of the textual features that are most predictive of teacher ratings of essay quality. Although the results of the teacher analyses are not surprising, they have now been assessed empirically. Through the use of two related, yet different, automated text analysis tools, Coh-Metrix and LIWC, we measured a number of surface- and deeper-level linguistic features of student essays with which teachers' ratings were highly associated. Coh-Metrix provided a broad analysis of the student essays, including indices of cohesion, and text difficulty at the lexical, syntactic, structural, and global levels of text (Graesser & McNamara, 2011). In addition, LIWC offered word-based analyses of the lexical, semantic, and thematic properties of student essays. With these linguistic measures of student essays, we are able to account for a significant amount of variance in teachers' essay assessments.

Although prior work has utilized text analysis tools to investigate expert raters' scores of essay quality, it is unclear whether and how these ratings correspond to classroom teachers' assessments. As classroom teachers and expert raters differ in their goals, training, and context of their scoring, their evaluation criteria also differ. Our analyses revealed that teacher ratings were, indeed, similar to expert ratings of essay quality. Consistent with prior research on expert ratings of essay quality, the teachers' ratings were related to skillful language, text organization, and text elaboration. In addition, a notable similarity between experts and teachers was the negative influence of cohesive devices. Similar to experts, the teachers in the current study did not associate cohesive essays with higher quality ratings (Crossley & McNamara, 2010; Crossley & McNamara, 2011; Crossley & McNamara, 2012; McNamara, et al., 2010). In fact, the teacher ratings were *negatively* correlated with indices of cohesion. Overall, the analyses suggest that classroom teachers, like expert raters, are able to assess student essays on myriad different features, ranging from surface-level lexical features to deep-level properties of text cohesion.

5.2 Student-Teacher Evaluative Misalignment

Beyond our analysis of teacher ratings, we further investigated the degree to which students' and teachers' ratings of essay quality were *misaligned*. Although prior research has investigated the *benefits* and, less commonly, the *accuracy* of students' self-assessments (Andrade & du Boulay, 2003; Andrade, et al., 2008; Graham & Perin, 2007; Hillocks, 1986; Ross et al., 1999), no work has explored the linguistic features that predict students' self-assessments. In our study, we investigated the linguistic features associated with students' self-assessments in order to determine the degree of misalignment between students' and teachers' essay ratings.

The correlation and regression analyses confirmed that there was, indeed, misalignment of student and teacher expectations for writing quality. In line with prior research on students' performance self-assessments (Falchikov & Boud, 1989; Dunning et al., 2003; Kruger & Dunning, 1999), the students participating in this study overestimated the quality of their writing. Teachers gave essays an average score of 3.67, whereas students had an average self-assessment of 4.04. In addition, the scores were weakly correlated ($r = .26$) indicating that while the score means are not vastly different, the essays rated as low quality by the teachers may have been given higher ratings by students, and vice versa.

A potential limitation of this study lies in the differences between the types of rubrics and training used by the teachers and students. Indeed, this factor deserves further attention in future studies, as student-teacher misalignment may be remediated through more specific student rubrics or more extensive evaluation training. Nonetheless, the differences in the linguistic features of the essays that are associated with the scores are less likely to be driven solely by the rubrics. First, our analyses of the linguistic features related to student and teacher essay ratings indicated that student ratings were related to fewer measured variables than the teacher ratings. While

teachers' assessments were influenced by multiple features of the essays, such as sophistication of vocabulary, text organization, and objective and factual language use, students seemed to focus only on a subset of these features. For instance, the LIWC analysis revealed that students and teachers were both sensitive to objective language use, as revealed by the association between essay ratings and pronoun usage. However, teachers' ratings were associated with numerous additional aspects of language use, such as a lack of hypotheticals, hedging words, and emotionally charged language. These findings suggest that, while students can understand the individual aspects of writing quality highlighted in classroom instruction, they likely find it challenging to understand how the effects of some features depend on others. As a result, they focus on fewer or more simplistic aspects of writing quality when assessing essay quality.

Second, in addition to focusing on fewer essay features, students rated essays based on a *different* collection of features than the teachers. For instance, student ratings were positively affected by high semantic overlap (*LSA verb overlap*) within the essays. Thus, the more essays exhibited semantic cohesion among sentences, the more highly students rated their essays. As previously discussed, however, teachers' ratings were *negatively* associated with measures of essay cohesion (*LSA paragraph-to-paragraph* and *content word overlap*). Our analysis, therefore, reveals distinct misalignment between the evaluation criteria of the students and teachers. Specifically, student ratings were based on both fewer and different textual features than teachers' ratings.

We assume that students have yet to develop a complete criterion for evaluating their own essays, and, as a result, are missing the more nuanced and dynamic features that contribute to quality essays. However, students' inaccurate self-assessments are the consequence of numerous factors, including students' knowledge of writing, students' metacognitive monitoring, classroom environment, and pedagogical methods. Of course, this study only addressed part of the evaluative misalignment hypothesis. Here, our goal was to establish the presence of a misalignment and to explore its nature in terms of the linguistic features of the essays that influence the students' and teachers' assessments of essay quality. As such, it is important to note that this type of analysis is complementary, not alternative, to analyses of teachers' explicit criteria for writing. In the future, researchers should investigate the misalignments revealed through students' and teachers' explicit reports of writing quality.

In addition, future studies should explore the sources and causes of student-teacher misalignments. One potential source of the student-teacher misalignment could be the teachers' ability to compare student essays with those from their classmates. That is, because teachers are able to view multiple essays along numerous levels of quality, they are better able to make sophisticated quality judgments than students. Future studies should investigate this question by presenting students with their own essays in conjunction with several other essays varying in quality. Analyses could then investigate whether students' self-ratings are more aligned with teachers' ratings, and whether this peer-review process results in improved performance on subsequent writing tasks. Along these lines, prior research on peer review and writing has revealed

that students benefit from peer review exercises (Nelson & Schunn, 2000; Cho & Schunn, 2007).

Future research should also investigate this evaluative misalignment from a developmental perspective. Because students' aptitude for writing is a result of multiple developing factors, such as interest, motivation, and conceptual competence (Lipstein & Renninger, 2006), it is important to investigate how these variables interact with students' development of accurate evaluation criteria. In the future, the method outlined in this paper should be used to provide important insight into students' development of criteria. In particular, these textual analyses can be applied to longitudinal data of student and teacher essay assessments to determine how misalignment changes over time.

Regardless of their source or cause, misalignments in expectations pose a threat to students' successful writing development. Our results here indicate that students lack stringent evaluation criteria for their essays that consider the myriad of linguistic, rhetorical, and semantic characteristics associated with essay quality. Despite teachers' best approaches and practices in the classroom, students can still misinterpret or misunderstand the writing process, as well as the associated lessons and strategies. Students without systematic criteria to evaluate their writing will have difficulties with certain phases of the writing process (e.g., revision) as well as improving performance on future writing assignments.

Note

[1] Not to be confused with alignment in conversational discourse, *evaluative misalignment* refers to a discrepancy between students' and teachers' criteria for judging essay quality.

References

Andrade, H., & du Boulay, B. (2003). The role of rubric-referenced self-assessment in learning to write. *Journal of Educational Research, 97,* 21-34.

Andrade, H., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practices, 27,* 3-13.

Camara, W. J. (2003). *Scoring the essay on the SAT writing section.* New York: College Entrance Examination Board.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education, 48,* 409-426.

Crossley, S. A. & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life-Long Learning, 21,* 170-191.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35,* 115-136.

Crossley, S. A., Weston, J., McLain-Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28,* 282-311.

Donovan, C. A., & Smolkin, L. B. (2006). Children's understanding of genre and writing development. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 131-143). New York: Guilford.

Dunning, D., Johnson, K., Ehrlinger, J., & Kruger, J. (2003). Why people fail to recognize their own incompetence. *Current Directions in Psychological Science, 12,* 83-86.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59,* 395-430.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database.* Cambridge, MA: MIT Press.

Gesier, S., & Studley, R. (2001). UC and SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California. Oakland, CA: University of California.

Gill, A. J., French, R. M., Gergle, D., & Oberlander, J. (2008). The language of emotion in short blog texts. *Proceedings of CSCW* (pp. 299-302). New York, NY: ACM Press.

Graesser, A. C. & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 2,* 371-398.

Graesser, A. C., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193-202.

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 187-207). New York: Guilford Press.

Graham, S. & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*, 445-476.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* London: Longman.

Hancock, J. T., Curry, L. E., Goorha, S., & Woodworth, M. (2008). On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes, 45,* 1-23.

Hancock, J. T., Landrigan, C., & Silver, C. (2007). Expressing emotion in text-based communication. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 929-932). New York, NY: ACM.

Hillocks, G. (1984). What works in teaching composition: A meta-analysis of experimental treatment studies. *American Journal of Education, 93,* 133-170.

Hillocks, G. (1986). *Research on written composition: New directions for teaching.* Urbana, IL: Eric Clearinghouse on Reading and Communication Skills.

Kos, R., & Maslowski, C. (2001). Second graders' perceptions of what is important in writing. *The Elementary School Journal, 101,* 567-585.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77,* 1121-1134.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *LSA: A road to meaning.* Mahwah, NJ: Erlbaum.

Light, R. J. (2001). *Making the most of college: Students speaking their minds.* Cambridge: Harvard University Press.

Lin, S., Monroe, B., & Troia, G. A. (2007). Development of writing knowledge in grades 2-8: A comparison of typically developing writers and their struggling peers. *Reading and Writing Quarterly, 23,* 207-230.

Lipstein, R., & Renninger, K. A. (2006). "Putting things into words": 12-15 year-old students' interest for writing. In P. Boscolo & S. Hidi (Eds.), *Motivation and writing: Research and school practice.* New York: Kluwer Academic/Plenum.

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics, 12,* 291-315.

McNamara, D. S., Cai, Z., & Louwerse, M. M. (2007). Optimizing LSA measures of cohesion. In T. K. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 379-400). Mahwah, NJ: Erlbaum.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27,* 57-86.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188-205). Hershey, PA: IGI Global.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*, 292-330.

McNamara, D. S., & Magliano, J.P. (2009). Self-explanation and metacognition: the dynamics of reading. In D.J. Hacker, J. Dunlosky, & A.C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 60–81). New York: Routledge.

National Assessment of Educational Progress. (2007). *The Nation's Report Card: Writing 2007*. Retrieved Nov. 5, 2010, nces.ed.gov/nationsreportcard/writing/

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science, 37,* 375-401.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC [computer software]. Austin, TX.

Pennebaker, J. W., Mayne, T. J., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72,* 863-871.

Powell, P. (2009). Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication, 60*, 664-682.

Rogers, L. A., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology, 100,* 879-906.

Roscoe, R. D., Crossley, S. A., Weston, J. L., & McNamara, D. S. (2011). Automated assessment of paragraph quality: Introductions, body, and conclusion paragraphs. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 281-286). Menlo Park, CA: AAAI Press.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing, 6,* 107-132.

Rus, V., Lintean, M., Graesser, A. C., & McNamara, D. S. (2009). Assessing student paraphrases using lexical semantics and word weighting. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C. Graesser (Eds.), *Artificial intelligence in education: Building learning systems that care; From knowledge representation to affective modeling* (pp. 165-172). Amsterdam, The Netherlands: IOS Press.

Schoonen, R., & De Glopper, K. D. (1996). Writing performance and knowledge about writing. In G. Rijlaarsdam, H. Van Den Bergh, & M. Couzijn (Eds.), *Theories, models and methodology in writing research* (pp. 87-107). Amsterdam: Amsterdam University Press.

Tousignant, M., & DesMarchais, J. E. (2002). Accuracy of student self-assessment ability compared to their own performance in a problem-based learning medical program: A correlation study. *Advances in Health Sciences Education, 7,* 19-27.

Wilson, M. D. (1988). The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2. *Behavior Research Methods, Instruments, & Computers, 20,* 6-11.

Wong, B. (1999). Metacognition in writing. In R. Gallimore, L. P. Bernheimer, D. L.MacMillan, D. L. Speech, & S. Vaughn (Eds.), *Developmental perspectives on children with high-incidence disabilities* (pp. 183-198). Mahwah, NJ: Erlbaum.