# Text Coherence and Judgments of Essay Quality: Models of Quality and Coherence

**Scott A. Crossley (scrossley@gsu.edu)**
Department of Applied Linguistics/ESL, Georgia State University
Atlanta, GA, 30303 USA

**Danielle S. McNamara (dsmcnamara1@gmail.com)**
Department of Psychology, Institute for Intelligent Systems, The University of Memphis
Memphis TN 38152 USA

**Abstract**

This study investigates the importance of human evaluations of coherence in predicting human judgments of holistic essay quality. Of secondary interest is the potential for computational indices of cohesion and coherence to model human judgments of coherence. The results indicate that human judgments of coherence are the most predictive features of holistic essay scores and that computational indices related to text structure, semantic coherence, lexical sophistication, and grammatical complexity best explain human judgments of text coherence. These findings have important implications for understanding the role of coherence in writing quality.

**Keywords:** Coherence; Writing Quality; Cohesion, Computational Linguistics, Computational Models.

## Introduction

Writing is an important aspect of communication because it provides the opportunity to articulate ideas and synthesize perspectives in a persuasive manner that is independent of time and space constraints (Crowhurst, 1990). Learning how to convey meaning competently in written texts is a crucial skill for academic and professional success. Indeed, the writing skills of college freshmen are among the best predictors of academic success (Geiser & Studley, 2001). The value of writing in academic and professional settings renders the understanding of writing and, particularly, the difference between good and poor writing, an important objective, both for theoretical and applied reasons.

The primary goal of this study is to identify the features of essays that are most predictive of overall writing quality with a specific emphasis on the role text coherence plays in essay quality. Our secondary interest is in modeling human judgments of coherences using new computational indices related to text cohesion and text coherence. Cohesion refers to the presence or absence of explicit cues in the text that allow the reader to make connections between the ideas in the text, whereas coherence refers to the understanding that the reader derives from the text, which may be more or less coherent depending on a number of factors, such as prior knowledge, textual features, and reading skill (McNamara, Kintsch, Songer, & Kintsch, 1996).

There is a general sense that essay quality is highly related to the cohesion of a text, and, by proxy, text coherence. This is reflected in the literature about writing as well as textbooks that teach students how to write. Until recently, there were few studies that had empirically investigated the role of cohesion cues in essays. However, studies by McNamara, Crossley, and McCarthy (2010) and Crossley and McNamara (in press) have found no evidence that cohesion cues and essay quality are related. McNamara et al. (2010) found no differences between high and low proficiency essays according to indices of cohesion. In contrast, indices related to language sophistication did show significant differences between the groups. Crossley and McNamara (2010) also found that linguistic sophistication characterized essays rated as higher quality. In addition they found that an index related to text cohesion (aspect repetition) correlated negatively with human scores of essay quality indicating that more cohesive essays were rated as lower quality.

In a continuation of these studies, Crossley and McNamara (2010) investigated the degree to which analytical rubric scores of essay quality (e.g., essay cohesion, essay coherence, essay structure, strength of thesis, conclusion type) predicted holistic essays scores. This analysis permitted an examination of relations between holistic essay scores and analytic factors to determine the importance of these features in predicting essay quality. They found that human judgments of text coherence were the most informative predictor of human judgments of essay quality, explaining 65% of the variance.

Crossley and McNamara (2010) also examined links between the cohesive devices reported by Coh-Metrix (e.g., semantic coreference, causal cohesion, spatial cohesion, temporal cohesion, connectives and logical operators, anaphoric resolution, word overlap) and human judgments of coherence. Among these variables, only one index (*subordinating conjunctions)* demonstrated positive, significant correlations with the human ratings of coherence; however, this index also had strong links to syntactic complexity. The majority of the cohesion indices correlated negatively to the human ratings, indicating an inverse relation between the selected cohesion variables and the human judgments of coherence. Thus, while the Crossley and McNamara (2010) study indicated that human ratings of coherence were important indicators of holistic evaluations of essay proficiency, how human raters construct a coherent mental representation did not correlate with the cohesive devices provided by Coh-Metrix.

## Method

Our method of analysis is similar to that reported in Crossley and McNamara (2010) in that we examine argumentative essays written by college freshmen and scored by expert raters on analytic features of essay quality (i.e., effective lead, clear purpose, topic sentences, paragraph transitions, organization) as well as a holistic evaluation of essay quality. Our primary goal is to better understand which judgments of individual text features best explain judgments of overall text quality. Like Crossley and McNamara, we are specifically interested in text features related to coherence. However, we improve upon this earlier study by analyzing a larger corpus of essays collected under conditions that better represent high stakes testing. Also, unlike the Crossley & McNamara study, which suffered from low agreement between raters on many of the analytical text features, we use a different set of text features that better represent the organizational and rhetorical characteristics of essays. Using such features, we hope to increase inter-rater reliability between our expert raters and thus provide stronger links to the underlying cognitive construct of interest (i.e., coherence).

In order to assess which semantic features of the text might influence human judgments of coherence, we also report on a range of new computational indices developed to assess coherence. Our secondary goal is to model human judgments of coherence in order to better understand which features of a text help to develop coherent text.

### Corpus

As in Crossley and McNamara (2010), our analyses were conducted using a corpus of essays collected from undergraduate students at Mississippi State University (MSU). However, the essays we collected for this analysis differed in that they were based on SAT prompts and were timed. During the collection process, students were given 25 minutes to write an essay and no outside referencing was allowed. Such an environment better represents high stakes testing (i.e., SAT writing tests). Two SAT prompts were used and students were randomly assigned one prompt to which they responded. All students were native speakers of English and were in either Composition One or Composition Two course (i.e., freshmen composition). In total, 315 students wrote one essay each. Each essay was read and scored by two trained raters using both an analytic and a holistic rubric.

### Rating Rubric

Experts in the field of composition studies developed the analytic rubric used to score the individual features of the essays in this analysis. The rubric was used in the composition program at MSU to evaluate writer proficiency. Minor changes in the rubric were made by trained cognitive scientists and the director of the composition program at MSU to ensure that the construct of interest (coherence) was adequately assessed. The analytic rubric was then subjected to usability tests by expert raters with at least three years experience in essay scoring. The final version of the rubric had four subsections: introduction, body, conclusion, and correctness. The introduction subsection contained questions related to the use of an effective lead, clear purpose, and clear plan. The body subsection addressed the use of topic sentences, paragraph transitions, clear organization, and essay unity. The conclusion subsection included judgments on the strength of summarization and conviction. The correctness subsection identified the proper use of grammar, syntax, and mechanics. Two of these analytic features (Organization and Unity) evaluated semantic based, global cohesion (i.e., structural elements that promote overall comprehension) and thus were classified as measures of text coherence. One of these features (Paragraph Transitions) evaluated explicit cue-based, local cohesion and was classified as a measure of cohesion. A holistic grading scale based on a standardized rubric commonly used in assessing Scholastic Achievement Test (SAT) essays was also included in the rating rubric. This holistic scale was the same scale used by McNamara et al. (2010) and Crossley and McNamara (2010). The holistic scale and all of the rubric items had a minimum score of 1 and a maximum score of 6. The analytic rubric ratings included the following:

**Effective Lead:** The introduction begins with some device to grab the reader's attention and point toward the thesis.
**Clear Purpose:** The introduction provides essential background information and establishes the significance of the discussion.
**Clear Plan:** The introduction ends with a thesis statement that provides a claim and previews the support and organizational principle to be presented in the body.
**Topic Sentences:** Each paragraph includes a sentence that connects with the thesis and makes a comment on one of the points outlined in the introduction.
**Paragraph Transitions:** Each topic sentence is preceded by a phrase, clause, or sentence that links the current paragraph with the previous one.
**Organization:** The body paragraphs follow the plan set up in the introduction.
**Unity:** The details presented throughout the body support the thesis and do not stray from the main idea.
**Perspective:** The writer summarizes the key points that collectively sustain the thesis and stress its significance.
**Conviction:** The author re-establishes the significance of the discussion as it pertains to the thesis.
**Grammar, Syntax, and Mechanics:** The writer employs correct Standard American English.

### Essay Evaluation

Eight expert raters with either master's degrees or Ph.D.s in English and with at least 3 years experience teaching composition classes at the university level rated the 315 essays from the corpus using the analytic and holistic rubrics. The raters were informed that the distance between

each score was equal. Accordingly, a score of 5 is as far above a score of 4 as a score of 2 is above a score of 1. The raters were first trained to use the rubric with 20 essays. A Pearson correlation for each analytic rubric evaluation was conducted between the raters' responses. If the correlations between the raters did not exceed $r = .50$ (which was significant at $p < .05$) on all items, the ratings were reexamined until scores reached the $r = .50$ threshold. Raters followed similar protocols for the holistic score, but were expected to reach an $r >= .70$.

After the raters had reached an inter-rater reliability of at least $r = .50$ for the analytic scores ($r = .70$ for the holistic score), each rater then evaluated a selection of the 315 essays that comprise the corpus used in this study. Each essay was scored by at least two raters. Once final ratings were collected, differences between the raters were calculated. If the difference in ratings on survey feature were less than 2, an average score was computed. If the difference was greater than 2, a third expert rater adjudicated the final rating. Correlations between the raters (before adjudication) are located in Table 1. The raters had the lowest correlations for judgments of unity and the highest correlations for holistic essay scores. All correlations were $r > .65$. The average correlations across all essay feature judgments was $r = .72$. The inter-rater reliability reported here is much higher than that reported by Crossley and McNamara (i.e., $r = .455$, 2010).

Table 1: Pearson Correlations between Raters

| Item | *r* |
|---|---|
| Effective Lead | 0.706 |
| Clear Purpose | 0.693 |
| Clear Plan | 0.684 |
| Topic Sentences | 0.733 |
| Paragraph Transitions | 0.734 |
| Organization | 0.692 |
| Unity | 0.661 |
| Perspective | 0.770 |
| Conviction | 0.762 |
| Grammar, syntax, and mechanics | 0.740 |
| Holistic Score | 0.789 |

## Results

We conducted a multiple regression analysis to examine the predictive strength of the analytic features in explaining the scoring variance in the holistic scores assigned to the essays. We hypothesized that an analytic score representing text coherence would explain the most variance in the holistic scores based on the findings of Crossley and McNamara (2010). We used a training set to generate a model to examine the amount of variance explained by each analytical score. The model was then applied to a test set to calculate the accuracy of the analysis. Accordingly, we randomly divided the corpus into two sets: a training set ($n = 209$) and a test set ($n = 106$). The training set was used to identify which of the analytic scores most highly correlated with the holistic scores assigned to the essays. These analytic scores were later used to predict the holistic scores in the training and test sets using the generated model.

We controlled the number of variables included in the regression analysis in order to reduce the likelihood that the model was over-fitted. If too many variables are used, the model fits not only the signal of the predictors, but also the unwanted noise. The model may, thus, lack accuracy when applied to a new data set. We selected a ratio of 20 observations to 1 predictor, which is standard for analyses of this kind. Given that the training set comprised 209 essays, we determined that we could include 20 features in our regression analysis.

### Pearson Correlations

All features on the analytic rubric correlated significantly with the holistic scores assigned to the essays in the training set. The strongest correlations were for Organization (coherence), Perspective, Unity (coherence), and Conviction. The weakest correlations were for Paragraph Transitions (cohesion), Effective Lead, and Grammar. All the features along with their *r* values are presented in Table 2 (all $p < .001$).

Table 2: Pearson Correlations Analytic to Holistic Scores

| Variable | *r* value |
|---|---|
| Organization | 0.772 |
| Perspective | 0.749 |
| Unity | 0.741 |
| Conviction | 0.719 |
| Topic Sentences | 0.653 |
| Clear Plan | 0.643 |
| Clear Purpose | 0.605 |
| Paragraph Transitions | 0.547 |
| Effective Lead | 0.513 |
| Grammar, syntax, and mechanics | 0.476 |

### Collinearity

Many of the features exhibited multi-collinearity (> .70). Unity and Topic Sentences were highly correlated with Organization. Conviction was highly correlated with Perspective. Clear Purpose was highly correlated with Clear Plan. Because these features had lower correlations with the holistic score as compared to the analytical scores with which they demonstrated multi-collinearity, they were dropped from the multiple regression analysis. Thus, only the variables Organization, Perspective, Clear Plan, Paragraph Transitions, Effective Lead, and Grammar were included in the regression.

### Multiple Regression Training Set

A linear regression analysis (stepwise) was conducted including the six variables. These six variables were

Table 3: Linear Regression Analysis to Predict Holistic Essay Ratings Training Set

| Entry | Variable Added | R | R2 | *B* | B | SE |
|---|---|---|---|---|---|---|
| Entry 1 | Organization | 0.772 | 0.596 | 0.272 | 0.282 | 0.049 |
| Entry 2 | Perspective | 0.840 | 0.705 | 0.324 | 0.359 | 0.039 |
| Entry 3 | Grammar, syntax, and mechanics | 0.871 | 0.759 | 0.200 | 0.225 | 0.030 |
| Entry 4 | Clear Plan | 0.883 | 0.780 | 0.144 | 0.147 | 0.044 |
| Entry 5 | Paragraph Transitions | 0.889 | 0.790 | 0.096 | 0.108 | 0.035 |
| Entry 6 | Effective Lead | 0.892 | 0.795 | 0.097 | 0.093 | 0.041 |

Notes: Estimated Constant Term is -0.621; *B* is unstandardized Beta; B is standardized Beta; SE is standard error

regressed onto the raters' holistic evaluations for the 209 writing samples in the training set. The variables were checked for outliers and multi-collinearity. Coefficients were checked for both variance inflation factors (VIF) values and tolerance. All VIF values were at about 1 and all tolerance levels were well beyond the .2 threshold, indicating that the model data did not suffer from multi-collinearity (Field, 2005).

All six analytic features were significant predictors in the regression: Organization ($t = 5.542$, $p < .001$) Perspective ($t = 8.419$, $p < .001$), Grammar ($t = 6.646$, $p < .001$), Clear Plan ($t = 3.306$, $p < .001$), Paragraph Transitions ($t = -2.701$, $p < .050$), and Effective Lead ($t = 2.371$, $p < .050$). The linear regression using the eight variables yielded a significant model, $F(6, 202) = 130.816$, $p < .001$, $r = .892$, $r^2 = .795$, demonstrating that the combination of the six variables accounted for 80% of the variance in the human evaluations essay quality for the 209 essays examined in the training set. All the features retained in the regression analysis along with their $r$ values, $r^2$ values, unstandardized Beta weights, standardized Beta weights, and standard errors are presented in Table 3.

## Test Set Model

To further support the results from the multiple regression conducted on the training set, we used the B weights and the constant from the training set multiple regression analysis to estimate how well the model would function on an independent data set (the 106 essays and their holistic scores held back in the test set). The model produced an estimated value for each writing sample in the test set. We used this correlation along with its $r^2$ to demonstrate the strength of the model on an independent data set. The model for the test set yielded $r = .899$, $r^2 = .808$. The results from the test set model demonstrate that the combination of the six variables accounted for 81% of the variance in the evaluation of the 106 essays comprising the test set.

## Linguistic Features Analysis

As in Crossley and McNamara (2010), our regression analysis demonstrated that coherence is the most important predictor of human judgments of essay quality. Here, however, coherence was defined more specifically as connections between the claims and supports presented in the introduction and the themes in body paragraphs. Our secondary goal is to identify if computational indices related to text difficulty, test structure, cohesion, and coherence can account for the variance in the coherence ratings produced by the human raters.

To model coherence scores, we conducted an analysis of the Organization scores using computational indices provided by Coh-Metrix and new indices developed for this study. Our analysis was similar to that of our primary study in that we used Pearson Correlations to select variables and check for multi-collinearity. Selected variables were then included within a multiple regression analysis to predict variance in human scores of coherence. Our goal in this second analysis is two-fold: to attempt to model and explain human judgments of coherence using computational indices and examine if indices related to cohesion and coherence are important in this modeling. We used the same corpus as in our first study (including the division into training and test sets), but concentrated solely on the human ratings for the Organization item (i.e., the coherence feature that was most predictive of overall essay quality).

We first chose a selection of measures related to text difficulty that have explained essay quality in previous studies (McNamara et al., 2010; Crossley & McNamara, in press) predicting that text difficulty influences coherence. These measures included lexical sophistication (e.g., frequency, hypernymy, polysemy, concreteness, lexical diversity), syntactic complexity (e.g., part of speech tags, phrase type counts, number of words before the main verb), and essay structure (e.g., number of word types, sentences, and paragraphs). We also selected a range of measures related to cohesion from the Coh-Metrix tool. The constructs measured included semantic coreference (LSA indices), causal cohesion, connectives and logical operators, anaphoric resolution, and word overlap. Each construct was measured using multiple Coh-Metrix indices. (see McNamara & Graesser, in press, for an overview of the indices in Coh-Metrix). We hypothesize, based on the findings of Crossley and McNamara (2010), that none of the cohesion indices found in Coh-Metrix would correlate with human judgments of coherence.

## New Measures of Coherence

We developed new indices of semantic coherence to assess human coherence judgments. These indices measured

lexical and semantic overlap between paragraphs (initial to middle paragraphs, middle paragraphs to final paragraph, and initial paragraph to final paragraph) and between the entire essay and the essay prompt. The indices were designed to evaluate topic coherence at the paragraph and text level, levels previous indices of cohesion did not assess. In the indices assessing middle paragraphs, the middle paragraphs of essays were treated as an entire text segment. Our method of measuring semantic similarity between paragraph types and between the essay and the prompt was through LSA cosine values. Our method of measuring lexical overlap was though key word overlap between paragraph types (initial, middle, and body paragraphs).

Table 4: Correlations between Representative Coh-Metrix Indices and Raters' Organization Scores

| Variable | r value | *p* value |
|---|---|---|
| LSA middle to final paragraph | 0.307 | < .001 |
| LSA initial to middle paragraphs | 0.288 | < .001 |
| LSA initial to final paragraph | 0.143 | < .050 |
| Key word overlap initial to middle paragraphs | 0.139 | < .050 |
| MED content words mean | 0.093 | 0.181 |
| LSA givenness | 0.093 | 0.181 |
| LSA essay to prompt | 0.091 | 0.192 |
| LSA sentence to sentence mean | 0.070 | 0.310 |
| Noun overlap adjacent sentences | 0.027 | 0.702 |
| Key words overlap initial to final paragraph | 0.016 | 0.823 |
| Incidence of all connectives | 0.011 | 0.879 |
| Incidence of causal verbs | -0.034 | 0.628 |
| Incidence of logical operators | -0.035 | 0.613 |
| Adjacent anaphor reference | -0.061 | 0.377 |

## Pearson Correlations

Over 50 computational indices demonstrated significant correlations with the human ratings of coherence in the training set. The majority of the indices were related to text structure (i.e., number of word types, number of sentences, number of paragraphs), and lexical sophistication (i.e., lexical diversity, word frequency, word concreteness). Many of our new indices related to semantic coherence also demonstrated significant (although moderate) correlations with human judgments of coherence. No indices of cohesion from Coh-Metrix demonstrated significant correlations and many were negatively correlated. The correlations for the indices that best represent our coherence and cohesion measures are presented in Table 4. In light of space considerations, we do not present all correlations from this analysis.

## Collinearity

Only LSA initial to middle paragraphs and LSA middle to final paragraph were highly correlated. Because LSA initial to middle paragraphs had lower correlations with the Organization score, it was dropped from the multiple regression analysis. In total, 28 computational indices were available for the regression. To control for overfitting, only the 20 top indices were included in the regression analysis.

## Multiple Regression Training Set

A linear regression analysis (stepwise) was conducted including the 20 computational indices. These 20 variables were regressed onto the raters' Organization evaluations for the 209 writing samples in the training set. The variables were checked for outliers and multi-collinearity. Coefficients were checked for both variance inflation factors (VIF) values and tolerance. All VIF values were at about 1 and all tolerance levels were well beyond the .20 threshold, indicating that the model data did not suffer from multi-collinearity (Field, 2005).

Of these 20 variables, six were significant predictors in the regression: total word types ($t = 4.053$, $p < .001$) LSA middle to final paragraph ($t = 1.851$, $p < .050$), base verb forms (i.e., uninflected, finite verb forms, $t = -3.174$, $p < .010$), word frequency ($t = -5.295$, $p < .001$), lexical diversity ($t = -2.606$, $p < .010$), and number of paragraphs ($t = 2.206$, $p < .050$). The linear regression using the six variables yielded a significant model, $F(6, 201) = 17.840$, $p < .001$, $r = .589$, $r^2 = .347$, demonstrating that the combination of the six variables accounts for 35% of the variance in the human evaluations of coherence for the 209 essays examined in the training set. All the features retained in the regression analysis along with their $r$ values, $r^2$ values, unstandardized Beta weights, standardized Beta weights, and standard errors are presented in Table 5.

## Test Set Model

To further support the results from the multiple regression conducted on the training set, we used the B weights and the constant from the training set multiple regression analysis to estimate how well the model would function on an

Table 5: Linear Regression Analysis to Predict Organization Scores: Training Set

| Entry | Variable Added | R | R2 | *B* | B | SE |
|---|---|---|---|---|---|---|
| Entry 1 | Total word types | 0.403 | 0.162 | 0.158 | 0.361 | 0.002 |
| Entry 2 | LSA middle to final paragraph | 0.463 | 0.214 | 0.207 | 0.140 | 0.350 |
| Entry 3 | Base form incidence score | 0.532 | 0.283 | 0.272 | -0.186 | 0.004 |
| Entry 4 | CELEX frequency content words | 0.552 | 0.305 | 0.291 | -0.342 | 0.386 |
| Entry 5 | Lexical diversity D | 0.564 | 0.318 | 0.301 | -0.236 | 0.004 |
| Entry 6 | Total number of paragraphs | 0.589 | 0.347 | 0.328 | 0.173 | 0.060 |

Notes: Estimated Constant Term is 7.836; *B* is unstandardized Beta; B is standardized Beta; SE is standard error

independent data set (the 106 essays and their Organization scores held back in the test set). The model for the test set yielded $r$ = .619, $r^2$ = .384. The results from the test set model demonstrate that the combination of the six variables accounted for 38% of the variance in the evaluations of Organization for the 106 essays comprising the test set.

## Discussion and Conclusion

This study has provided additional evidence supporting the importance of human judgments of coherence in explaining holistic judgments of essay quality. As in Crossley and McNamara (2010), our top predictor of essay quality was an analytical feature related to coherence, which explained 60% of the variance in holistic essay scores. We built on the Crossley and McNamara study by examining a coherence construct (i.e., text organization) that was better specified and more commonly associated with writing assessment. As a result, the inter-rater reliability between raters for this feature was close to the accepted $r$ = .70 threshold ($r$ = .692). Additionally, because our construct is based on a structural property of text, it provides not only a situated understanding of essay coherence, but should also allow for more appropriately directed student feedback.

An analytical feature related to cohesion (Paragraph Transitions) was also a significant predictor of essay quality, but this feature only explained 1% of the variance in the human judgments. Thus, human judgments of cohesion explain judgments of essay quality to a lesser degree than human judgments of coherence.

The second analysis in this study investigated the potential for computational algorithms to model human judgments of coherence. A variety of indices related to linguistic sophistication and text structure were taken from Coh-Metrix along with new coherence indices developed for this study. Our regression analysis demonstrated that six indices related to text structure, semantic coherence, lexical sophistication, and grammatical complexity explained 38% of the human variance of coherence judgments. The strongest predictor of coherence was the number of types in the text, which overlaps conceptually with word length. Thus, more words in a text likely permit the development of greater representations of text coherence. Our second predictor of coherence was LSA middle to final paragraphs, which demonstrated that the semantic similarity between the body of middle paragraphs and the final paragraph helps develops coherent mental representations. Because a well-written final paragraph also summarizes the plan presented in the introduction and is the last section read, we argue that links between the evidence present in the body paragraphs and the summarization of this evidence in the conclusion affords greater text coherence in the mind of the reader. Our next three predictors of coherence were all related to linguistic sophistication. Coherent texts have more complex verb forms (i.e., inflected or infinitive forms), less frequent words, and a greater diversity of words. The finding likely demonstrates that text coherence for expert raters is a product of linguistic features related to overall writing quality that are more difficult to process and thus force the reader to attend to the text. Our last predictor of coherence was the number of paragraphs with more paragraphs (i.e., greater structure) leading to more coherent texts.

As with previous studies, our analysis also showed that the cohesion indices provided by Coh-Metrix were not positively related to judgments of text coherence, indicating that cohesive devices do not likely underlie the development of coherent textual representations of essay quality.

We conclude that coherence is an important indicator of essay quality and that a significant amount of variance in coherence judgments can be modeled using indices related to text structure, semantic coherence, lexical sophistication, and grammatical complexity. These findings provide a better understanding of text coherence and produce a strong foundation from which to further explore relations among text cohesion, coherence, and judgments of text quality.

## Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES. The authors would like to thank our expert raters.

## References

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. (in press). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*

Crowhurst, M. (1990). Reading/writing relationships: An intervention study. *Canadian Journal of Education, 15*, 155-172.

Geiser, 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.

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

McNamara, D. S., Kintsch, E., Butler-Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14,* 1-43.

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