# Cohesion, Coherence, and Expert Evaluations of Writing Proficiency

**Scott A. Crossley (sc544@msstate.edu)**
Department of English, Mississippi State University
MS, 39762 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 roles of cohesion and coherence in evaluations of essay quality. Cohesion generally has a facilitative effect on text comprehension and is assumed to be related to essay coherence. By contrast, recent studies of essay writing have demonstrated that computational indices of cohesion are not predictive of evaluations of writing quality. This study investigates expert ratings of individual text features, including coherence, in order to examine their relation to evaluations of holistic essay quality. The results suggest that coherence is an important attribute of overall essay quality, but that expert raters evaluate coherence based on the absence of cohesive cues in the essays rather than their presence. This finding has important implications for text understanding and the role of coherence in writing quality.

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

## Introduction

Writing affords the opportunity to thoroughly articulate ideas and synthesize a variety of perspectives allowing for persuasive communication that transcends both time and space (Crowhurst, 1990). As such, the ability to convey meaning proficiently in written texts is a critical skill for academic and professional success. Indeed, college freshmen' writing skills are among the best predictors of academic success (Geiser & Studley, 2001), and even outside of academia, writing skills continue to be important and are an important attribute of professional competence (Light 2001). As such, developing a better understanding of good and poor writing is an important objective, both for theoretical and applied reasons.

The overarching objective of this study is on the identification of essay features that are predictive of overall writing quality. Our goal is to better understand and model writing proficiency. We are particularly interested in the roles that cohesion and coherence play in writing quality. 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. For example, overlapping words and concepts between sentences indicate that the same ideas are being referred to across sentences. Likewise, connectives such as *because*, *therefore*, and *consequently*, inform the reader that there are relationships between ideas and the nature of those relationships. Whereas cohesion refers to the explicit cues in the text, *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 and reading skill (McNamara, Kintsch, Songer, & Kintsch, 1996; O'Reilly & McNamara, 2007).

There is a strongly held sense that essay quality is highly related to the cohesion and coherence of the essay. This is reflected in the literature about writing (e.g., Collins, 1998; DeVillez, 2003), as well as textbooks that teach students how to write (Golightly & Sanders, 1990). However, there are few studies that have empirically investigated the role of cohesion cues and by consequence, coherence, in essays. Whereas there is a strong assumption that coherence is an important aspect of writing, few studies have documented this assumption or tied the notion of coherence to explicit linguistic features of the essay. Indeed, our own examinations of linguistic features of good and poor essays have turned up no evidence that cohesion cues are positively related to essay quality for either first language writers (McNamara, Crossley, & McCarthy, 2010) or writers for whom English is their second language (Crossley & McNamara, in press). Therefore, the question of whether coherence or cohesion play important roles in essay writing and judgments of essay quality remains open.

In contrast, the role of cohesion in text comprehension is much better understood and there are numerous empirical studies on the topic (for a recent review, see McNamara, Louwerse, McCarthy, & Graesser, 2010). These studies show that increasing the cohesion of a text facilitates and improves text comprehension for many readers (Gernsbacher, 1990) and is particularly crucial for low-knowledge readers (McNamara et al., 1996).

From this literature on text comprehension, we glean two competing hypotheses for the effects of cohesion on estimates of essay quality (i.e., the coherence of the essay in the mind of the essay rater). On the one hand, cohesion underlies coherence, and thus should be important. On the other hand, the effects of cohesion on comprehension depend on the knowledge and reading skill of the reader. Indeed, a reverse cohesion effect, or an advantage for low cohesion text, can occur for high knowledge readers (McNamara, 2001; McNamara et al., 1996; O'Reilly & McNamara, 2007). High-knowledge readers, unlike low-knowledge readers, can successfully make the inferences needed to bridge the conceptual gaps that are in low-cohesion text. In fact, high-knowledge readers may benefit from low cohesion texts because gaps in cohesion force the reader to make connections in text that are not explicitly

available (McNamara, 2001; O'Reilly & McNamara, 2007). Hence, when the material covered in a text is familiar to the reader (as is often the case for narratives), cohesion cues may be unnecessary, and perhaps even distracting. Overall, text comprehension literature leads to the conclusion that cohesion may play an important role in facilitating coherence if the rater of the essay has less knowledge about the topic, but cohesion cues may be inversely related to essay scores if the rater has more knowledge about the topic.

We recently explored this topic by examining the effects of cohesion devices on human evaluations of writing quality. McNamara et al (2010) used linguistic indices of cohesion and language sophistication provided by the computational tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) to analyze a corpus of 120 argumentative essays written by college undergraduate and scored by expert raters using a holistic rubric. The essays were scored on a 1-6 scaled SAT rubric and then categorized into two groups: essays judged as low versus high quality. The results indicated that there were no differences between these two groups according to indices of cohesion (e.g., word overlap, causality, connectives). By contrast, indices related to language sophistication (lexical diversity, word frequency, and syntactic complexity) showed significant differences between the groups. A follow-up discriminant function analysis (DFA) showed that these indices successfully classified the essays into their respective groups at a level well above chance. The results of the McNamara et al. study provide initial indications that text cohesion may not be indicative of essay quality. Instead, expert raters in the McNamara et al. study judged essays as higher quality when they were more difficult to process (less familiar words, more complex syntax).

While McNamara et al. (2010) showed that cohesion cues were not related to the overall scores assigned by essay raters, it did not investigate the role of the raters' judgments of the coherence or cohesion of the essay, nor did it investigate whether cohesion cues are related to raters' judgments of coherence and cohesion. Hence the purpose of the current study is two-fold. First, we examine the assumption that judgments of essay coherence are predictive of the overall score for an essay. While this is a commonly held belief, we are aware of no empirical support for this assumption provided in the literature. Second, we examine whether cohesion cues as measured by Coh-Metrix are related to raters' estimates of an essay's coherence. Whereas McNamara et al. (2010) did not find a relation between indices of cohesion and the overall essay scores, it remains an open question as to whether cohesion indices might be related to more direct ratings of an essay's coherence.

## Method

Our method of inquiry involves an analysis of argumentative essays by expert scorers on atomistic features of essay quality (i.e., introductions, thesis statement, topic sentences, relevance, coherence) as well as a holistic evaluation of essay quality. Thus, unlike McNamara et al. (2010), we do not rely solely on computational indices to model overall essay quality, but instead concentrate on the evaluation of human judgments of individual text features in relation to overall text quality. Included in the individual text features evaluated by human experts are two measures of coherence. If the ratings of coherence are predictive of overall essay quality, we will also use computational indices of cohesion to model these human ratings. We can, thus, examine the importance of cohesion and coherence in writing quality and examine which cohesive devises may be predictive of human ratings of coherence Such an analysis will also afford the opportunity to examine whether indices of cohesion correlate with human ratings of coherence, providing us with an opportunity to gain a better understanding of the role cohesion plays in high-knowledge readers (i.e., the expert raters in our study).

### Corpus

As in McNamara et al. (2010), our analyses were conducted using a corpus of essays collected from undergraduate students at Mississippi State University (MSU). The MSU corpus was designed to account for learner variables such as age (adult students) and learning context (freshman college composition class). The corpus was also designed to consider task variables such as medium (writing), first language (English), genre (argumentative essays), essay length (between 500 and 1,000 words), and topics (3 prompts on equality, television, and creativity). The final corpus consisted of 184 essays. The essays were untimed and written outside of the classroom. Thus, referencing of outside sources was allowed, but was not required. Students were allowed to select the essay prompt. Therefore, there are an unequal number of essays per prompt. Although 100 of the essays used in our current analysis were also used in the McNamara et al. study, these 100 essays were evaluated by different raters in the current study. The raters used both an atomistic and holistic survey instrument.

### Rating Rubric

The essay-rating rubric used in this analysis was designed to parallel the rubric used initially by Breetvelt, van den Bergh, and Rijlaarsdam (1994) and later adapted with a focus on structure and argumentation by Sanders and Schilperoord (2006). Three experts in language processing with Ph.D.s in either linguistics or cognitive psychology developed the rubric. It was then subjected to usability tests by expert raters with at least three years experience in essay scoring. The final version of the survey instrument has three subsections: structure, content, and conclusion. The structure subsection contains questions related to essay structure and continuity. The content subsection contains questions related to the introduction, thesis, coherence, topic and evidential sentences, relevance, register use, and mechanics. The conclusion subsection contained questions related to the conclusion type, conclusion summary, and closing. In addition, the survey instrument included a holistic grading scale based on a standardized rubric

commonly used in assessing Scholastic Achievement Test (SAT) essays. This holistic scale was the same scale used by McNamara and colleagues (2010). The holistic scale and all of the rubric items had a minimum score of 1 and a maximum score of 6. The atomistic rubric ratings included the following:

**Structure:** Clarity of division into introductions, argumentation, and conclusion.
**Continuity:** Strength of connection of ideas and themes within and between the essays' paragraphs (cohesion).
**Introduction:** Presence of a clear, introductory sentence.
**Thesis Statement:** Strength of the thesis statement and its attached arguments.
**Reader Orientation:** Overall coherence and ease of understanding.
**Topic Sentences:** Presence of identifiable topic sentences in argumentative paragraphs.
**Evidential Sentences:** Use of evidential sentences in the argumentative paragraphs that support the topic sentence or paragraph purpose.
**Relevance:** Degree to which argumentation in the paper contained only relevant information.
**Appropriate Registers:** Degree to which the vocabulary in the essays followed the expected register.
**Grammar, Spelling, and Punctuation:** Accuracy of grammar, spelling, and punctuation.
**Conclusion:** Clarity of the conclusion.
**Conclusion Type***:* Identifiable conclusion type.
**Conclusion Summary**: Presence of summary within the conclusion including arguments and the thesis of the essay.
**Closing**. Clarity of closing statements within the essay.

### Essay Evaluation

Two expert raters with master's degrees in English and at least 3 years experience teaching composition classes at a large university rated the 184 essays from the corpus using the rubric. 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 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 protocol for the holistic score, but were expected to reach an $r >= .70$.

After the raters had reached an inter-rater reliability of at least $r = .50$ ($r = .70$ for the holistic score), each rater then evaluated the 184 essays that comprise the corpus used in this study. 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 continuity and the highest correlations for essay structure.

Table 1: Pearson Correlations between Raters

| Item | *r* |
|---|---|
| Structure | 0.647 |
| Continuity | 0.307 |
| Introduction | 0.330 |
| Thesis Statement | 0.513 |
| Reader Orientation | 0.367 |
| Topic Sentences | 0.510 |
| Evidential Sentences | 0.404 |
| Relevance | 0.306 |
| Appropriate Registers | 0.394 |
| Grammar, Spelling, Punctuation | 0.599 |
| Conclusion | 0.596 |
| Conclusion Type | 0.355 |
| Conclusion Summary | 0.525 |
| Closing | 0.445 |
| Holistic Score | 0.533 |

## Results

We used a multiple regression analysis to examine the predictive strength of the atomistic writing features in explaining the scoring variance in the holistic scores assigned to the essays. We used a training set to generate a model to examine the amount of variance explained by each writing feature. 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 ($n = 123$) and a test set ($n = 61$). The training set was used to identify which of the atomistic features most highly correlated with the holistic scores assigned to the essays. These features 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 15 observations to 1 predictor, which is standard for analyses of this kind (Field, 2005). Given that the training set contained 123 essays, we determined that we could include eight features in our regression analysis.

### Pearson Correlations

All features on the rubric correlated significantly with the holistic scores assigned to the essays in the training set. The strongest correlations were for Reader Orientation (coherence), Relevance, and Continuity (cohesion). The weakest correlations were for Thesis, Conclusion, and Introduction. All the features along with their *r* values are presented in Table 2 (all $p < .001$).

Table 2: Pearson Correlations Atomistic to Holistic Scores

| Variable | *r* value |
|---|---|
| Reader Orientation | 0.803 |
| Relevance | 0.710 |
| Continuity | 0.650 |
| Conclusion Type | 0.640 |
| Structure | 0.633 |
| Evidential Sentences | 0.629 |
| Grammar, Spelling, & Punctuation | 0.590 |
| Appropriate Registers | 0.589 |
| Topic Sentences | 0.583 |
| Closing | 0.578 |
| Conclusion Summary | 0.551 |
| Thesis Statement | 0.548 |
| Conclusion | 0.526 |
| Introduction | 0.389 |

## Collinearity

The features Structure and Conclusion were both highly correlated (> .70) with the feature Conclusion Type. Because both of these features had lower correlations with the holistic score as compared to Conclusion Type, the Structure and Conclusion variables were dropped from the multiple regression analysis. Thus only the variables Reader Orientation, Relevance, Continuity, Conclusion Type, Evidential Sentences, Grammar, Spelling, & Punctuation, Appropriate Registers, and Topic Sentences were included in the regression.

## Multiple Regression Training Set

A linear regression analysis (stepwise) was conducted including the eight variables. These eight variables were regressed onto the raters' holistic evaluations for the 123 writing samples in the training set. The variables were checked for outliers and multicollinearity. 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 multicollinearity (Field, 2005).

Five variables were significant predictors in the regression: Reader Orientation ($t$ = 6.668, $p$ < .001) Conclusion Types ($t$ = 5.068, $p$ < .001), Evidential Sentences ($t$ = 3.495, $p$ < .001), Topic Sentences ($t$ = 3.180, $p$ < .010), and Appropriate Registers ($t$ = -1.419, $p$ < .050). Three variables were not significant predictors: Relevance ($t$ = 1.841, $p$ > .050), Continuity ($t$ = 1.760, $p$ > .050), and Grammar, Spelling, & Punctuation ($t$ = 1.486, $p$ > .050). The latter variables were left out of the subsequent analysis. The linear regression using the eight variables yielded a significant model, $F(5, 117) = 89.693$, $p < .001$, $r = .891$, $r^2 = .793$, demonstrating that the combination of the five variables accounts for 79% of the variance in the human evaluations essay quality for the 123 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 61 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 $r = .922$, $r^2 = .850$. The results from the test set model demonstrate that the combination of the five variables accounted for 85% of the variance in the evaluation of the 61 essays comprising the test set.

## Linguistic Features Analysis

Our regression analysis demonstrated that text coherence is an important predictor of human judgments of essay quality. Our subsequent goal was to identify which linguistic features are attributable to the coherence construct used by the human raters.

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

| Entry | Variable Added | R | $R^2$ | *B* | B | SE |
|---|---|---|---|---|---|---|
| Entry 1 | Reader Orientation | 0.803 | 0.645 | 0.458 | 0.413 | 0.069 |
| Entry 2 | Conclusion Type | 0.850 | 0.723 | 0.296 | 0.257 | 0.058 |
| Entry 3 | Evidential Sentences | 0.871 | 0.758 | 0.271 | 0.182 | 0.078 |
| Entry 4 | Topic Sentences | 0.882 | 0.778 | 0.222 | 0.160 | 0.070 |
| Entry 5 | Registers | 0.891 | 0.793 | 0.201 | 0.152 | 0.069 |

Notes: Estimated Constant Term is 23.79; *B* is unstandardized Beta; B is standardized Beta; SE is standard error

To accomplish this goal, we conducted an analysis of the Reader Orientation scores using computational indices provided by Coh-Metrix that have theoretical correlates with cohesion features. Our goal in this second analysis is to examine if computational indices related to cohesion can successfully model the human coherence ratings from our essay analysis. We used the same corpus as the principle study, but concentrated solely on the human ratings for the Reader Orientation item (i.e., the coherence feature that was predictive of overall essay quality).

We selected a range of measures related to cohesion from the Coh-Metrix tool. The constructs measured included semantic coreference (LSA indices), causal cohesion, spatial cohesion, temporal cohesion, connectives and logical operators, anaphoric resolution, word overlap, and lexical diversity (see Crossley & McNamara, 2009; Graesser et al., 2004, for an overview of the cohesion indices in Coh-Metrix). Each construct was measured using multiple Coh-Metrix indices.

We first divided the corpus into a training (N = 123) and test set (N= 61). We then conducted Pearson correlations to relationships between the Coh-Metrix Indices and the human ratings of coherence.

**Pearson Correlations.** Among the selected cohesion constructs, only a few reported multiple indices that demonstrated significant correlations with the human ratings of coherence. The constructs that reported multiple significant indices included anaphoric reference (i.e., the proportion of anaphoric references between sentences), causal cohesion (i.e., the incidence of causal verbs and particles), incidence of connectives (i.e., positive temporal connectives, subordinating conjunctions, causative subordinators), and overlap measures (the overlap nouns, stems, and arguments between sentences). However, these correlations were negative (with the exception of Subordinating Conjunctions; i.e. *until, though, since*). Measures for semantic coreference, logical operators, lexical diversity, spatial cohesion, and temporal cohesion did not report significant indices. The indices with the highest correlations from the significant measures are presented in Table 3 along with their *r* and *p* values. The negative correlations indicate that the essays rated high in coherence included fewer cohesion cues.

Table 4: Correlations Coh-Metrix Indices to Raters' Coherence Scores

| Variable | r value | *p* value |
|---|---|---|
| Anaphoric reference | -0.349 | < .001 |
| Ratio of causal particles and verbs | -0.259 | < .010 |
| Incidence of positive temporal connectives | -0.237 | < .010 |
| Subordinating conjunctions | 0.240 | < .010 |
| Causative subordinators | -0.211 | < .050 |
| Content word overlap | -0.187 | < .050 |

## Discussion

This study has demonstrated that human ratings of coherence are an important indicator of holistic evaluations of essay proficiency. However, how human raters construct a coherent mental representation of a text seems opposed to many intuitive notions of coherence. For instance, we might expect that cohesive devices such as word overlap, causal particles and verbs, resolved anaphors, and positive temporal connectives would help the rater to develop a more coherent textual representation. However, in the case of the expert raters used in this study, the opposite was true. The absence of cohesive devices was associated with a more coherent mental representation of the text.

Our results indicate that coherence is an important element of human judgments of essay quality. In fact, overall text coherence is the most predictive feature of holistic essay scores. The coherence of a text (and by extension its understandability) was more predictive of writing quality than conclusion types, the use of evidential sentences, the use of topic sentences, and the use of appropriate registers. The overall coherence of a text was also the primary predictor of essay quality and explained 65% of the variance in the human ratings of writing quality. Human ratings of cohesion (continuity), although not retained in our regression analysis, also significantly correlated with essay quality.

However, our analysis using cohesion indices provided by Coh-Metrix demonstrated that our human judgments of coherence were not positively related to indices related to text cohesion indicating that cohesive devices may not underlie the development of coherent textual representations. Indeed, the majority of cohesive devices negatively correlated with human judgments of coherence. The exception is the use of subordinating conjunctions, which were positively correlated with human ratings of coherence. Yet, subordinating conjunctions also play a syntactic role and, by their nature, create more complex syntactic structures that result in a greater number of words before the main verb. Thus, it is likely that the subordinating conjunction index is actually detecting syntactic complexity, which does positively correlate with estimates of essay quality (McNamara et al., 2010).

So the question becomes: What factors are informing expert raters' mental representations of the text? One conclusion that the results of this study support is that factors important in text comprehension may have similarly important roles when raters evaluate the quality of essays. Specifically, the background knowledge of expert raters may influence text coherence in assessments of essay quality. Expert essay raters tend to be highly educated with advanced degrees and with experience in grading essays and other types of writing. The prompts used in the current study as well as prompts commonly used in essay writing assessments generally deal with topics that are relatively familiar to most educated individuals. As such, we can assume that essay raters will not tend to be low knowledge readers. Low knowledge readers lack sufficient knowledge

to generate inferences to bridge conceptual gaps in text, and, as a result, they tend to benefit from explicit text cohesion (i.e., word overlap, resolved anaphors, causal cohesion, connectives). By contrast, high knowledge readers benefit from texts low in cohesion because the cohesion gaps in the texts induce them to generate appropriate inferences to fill in the conceptual gaps. High knowledge readers can do this successfully because they have sufficient background knowledge to make appropriate inferences. When successful inferences are generated, the coherence of the mental representation can increase due to connections between the new information and their prior knowledge (McNamara, 2001; McNamara & McDaniel, 2004; O'Reilly & McNamara, 2007). Thus, more cohesive devices in essays may produce a less coherent mental representation in expert raters.

## Conclusion

We conclude that coherence is an important attribute of writing quality. Essay raters' evaluations of coherence were highly related to their overall holistic scores for the essays. Nonetheless, we have found here that coherence is not necessarily defined through the use cohesion devices, and in fact may be inversely related to the presence of cohesion cues. Thus, the question becomes: What textual features of an essay lead to higher versus lower estimates of essay coherence? Our results demonstrate that the indices currently available from which to measure cohesion are not strongly linked to human judgments of coherence. However, it is highly unlikely that textual features do not affect coherence. Thus, our task becomes the identification of these features and the derivation of computational algorithms that accurately model them.


### 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 Brad Campbell and Daniel White for their help in scoring the corpus of essays.


## References


Breetvelt, I., Van den Bergh, H., & Rijlaarsdam, G. (1994). Relations between writing processes and text quality: when and how? *Cognition and Instruction, 12*(2), 103-123.

Collins, J.L. (1998). *Strategies for struggling writers*. New York, NY: The Guilford Press.

Crossley, S.A. & McNamara, D.S. (2009). Computationally assessing lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18,* 119-135.

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.

DeVillez, R. (2003). *Writing: Step by step*. Dubuque, IO: Kendall Hunt.

Field, A. (2005). *Discovering statistics using SPSS*. London, English: Sage Publications.

Gernsbacher, M.A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Earlbaum.

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.

Golightly, K. B., & Sanders, G. (2000). *Writing and reading in the disciplines* (2nd Ed.). New Jersey: Pearson Custom Publishing.

Graesser, A.C., McNamara, D.S., & Louwerse, M.M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York, NY: Guilford Publications.

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

Light, R. (2001). *Making the most of college*. Cambridge, MA: Harvard University Press.

McNamara, D.S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51-62.

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

O'Reilly, T. & McNamara, D.S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal*, *44,* 161–196.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43,* 121–152.

Sanders, T., & Schilperoord, J. (2006). Text structure as a window on the cognition of writing. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *The handbook of writing research* (pp. 386 - 402). NY: Guilford Publications.