Optional ERIC Coversheet — Only for Use with U.S. Department of Education Grantee Submissions

This coversheet should be completed by grantees and added to the PDF of your submission if the information required in this form **is not included on the PDF to be submitted**.

INSTRUCTIONS

- Before beginning submission process, download this PDF coversheet if you will need to provide information not on the PDF.
- Fill in all fields—information in this form **must match** the information on the submitted PDF and add missing information.
- Attach completed coversheet to the PDF you will upload to ERIC [use Adobe Acrobat or other program to combine PDF files]—do not upload the coversheet as a separate document.
- Begin completing submission form at https://eric.ed.gov/submit/ and upload the full-text PDF with attached coversheet when indicated. Your full-text PDF will display in ERIC after the 12-month embargo period.

GRANTEE SUBMISSION REQUIRED FIELDS

Title of article, paper, or other content

AUTOMATED ANALYSES OF NATURAL LANGUAGE IN PSYCHOLOGICAL RESEARCH

All author name(s) and affiliations on PDF. If more than 6 names, ERIC will complete the list from the submitted PDF.

Last Name, First Name	Academic/Organizational Affiliation	ORCID ID
Allen, Laura K.	University of Minnesota	
Graesser, Arthur C.	University of Memphis	
McNamara, Danielle S.	Arizona State University	

Publication/Completion Date—(if *In Press,* enter year accepted or completed) 2023

Check type of content being submitted and complete one of the following in the box below:

- ☐ If article: Name of journal, volume, and issue number if available
- ☐ If paper: Name of conference, date of conference, and place of conference
- ☑ If book chapter: Title of book, page range, publisher name and location
- ☐ If book: Publisher name and location
- ☐ If dissertation: Name of institution, type of degree, and department granting degree

APA Handbook of Research Methods in Psychology: Foundations, Planning, Measures and Psychometrics, 36

DOI or URL to published work (if available)

Acknowledgement of Funding— Grantees should check with their grant officer for the preferred wording to acknowledge funding. If the grant officer does not have a preference, grantees can use this suggested wording (adjust wording if multiple grants are to be acknowledged). Fill in Department of Education funding office, grant number, and name of grant recipient institution or organization.

"This work was supported by U.S. Department of Education **[Office name]** National Center for Educat through **[Grant number]** R305A1802 to **Institution]** Arizona State U. The opinions expressed are those of the authors and do not represent views of the **[Office name]** Arizona State University or the U.S. Department of Education.

CHAPTER 17

AUTOMATED ANALYSES OF NATURAL LANGUAGE IN PSYCHOLOGICAL RESEARCH

Laura K. Allen, Arthur C. Graesser, and Danielle S. McNamara

Research in psychology often relies on qualitative and quantitative assessments of natural language in order to better identify and understand the mechanisms underlying complex cognitive tasks (Dowell et al., 2019; Graesser & McNamara, 2011; Johns & Jamieson, 2018; Magliano & Graesser, 2012; Pennebaker et al., 2007; Yan et al., 2020). Essays, open-ended questions, think-aloud protocols, and interviews are often the most robust methods for gleaning detailed insights into individuals' thoughts. They are frequently collected throughout psychological fields, including applied research domains, such as education, discourse processes, cognitive science, social and personality psychology, forensics, and clinical psychology. Despite the wealth of information that can be gleaned from these responses, analyses often rely on time-intensive annotation and scoring on the part of human expert raters, which can hinder progress and deter researchers from collecting these sources of data.

In response to these methodological challenges, researchers and educators have turned to complementary fields in computer science and learning analytics that have made substantial progress in automating the analysis of natural language through natural language processing (NLP; Clark et al., 2013; Hirschberg & Manning, 2015; Jurafsky & Martin, 2008), as well as leveraging big data approaches more broadly (Griffiths, 2015; Jones, 2017). This interdisciplinary work has led to landmark advances in language learning (Kyle, 2021; Kyle & Crossley, 2018), education (Dowell et al., 2020; Litman, 2016; McNamara et al., 2017), discourse processes (McNamara et al., 2014), and automated analyses of discourse cohesion (Dascalu et al., 2018; Graesser & McNamara, 2011). Researchers now have a broad range of options to implement automated textual analyses by taking advantage of development packages with popular programming languages (e.g., spaCy in Python and R; NLTK in Python; tidytext in R) as well as freely available NLP software facilities that render accomplished programming knowledge unnecessary (e.g., Crossley, Kyle, et al., 2019; Kyle et al., 2018; McNamara et al., 2014).

This chapter provides an overview of current approaches to NLP and how they have been applied to research in the psychological domain. We first provide an overview of how NLP techniques are used to aid in the scoring of natural language responses. Second, we describe how

This research was supported in part by IES Grants R305A180261, R305A180144, and R305A190063 as well as the Office of Naval Research (Grants: N00014-17-1-2300, N00014-19-1-2424, N00014-20-1-2627). Opinions, conclusions, or recommendations do not necessarily reflect the view of the Department of Education, IES, or the Office of Naval Research.

https://doi.org/10.1037/0000318-017
APA Handbook of Research Methods in Psychology, Second Edition: Vol. 1. Foundations, Planning, Measures, and Psychometrics, H. Cooper (Editor-in-Chief)

APA Handbook of Research Methods in Psychology: Foundations, Planning, Measures, and Psychometrics, edited by H. Cooper, M. N. Coutanche, L. M. McMullen, A. T. Panter, D. Rindskopf, and K. J. Sher

these same techniques can be used to infer psychological attributes from written responses, such as individual differences and learning processes. Third, we discuss how these analyses of natural language responses have been incorporated into intelligent tutoring systems (ITSs) that provide adaptive instruction to student users. Finally, we conclude with a brief discussion of more recently developed tools and approaches that examine multi-modal approaches to language analysis, with the inclusion of information related to timing, emotional states, and group dynamics.

NATURAL LANGUAGE PROCESSING AS A METHOD OF RESPONSE SCORING

The most common application of NLP to research has been in the automated scoring of written language. A variety of tasks in psychological domains rely on participants to generate written language responses, ranging from brief responses to longer essays. NLP techniques can be leveraged to provide scores on these written responses using features calculated at multiple dimensions of language. For example, NLP techniques can analyze characteristics of the words, sentences, and entire texts, including features related to a broad range of constructs including familiarity, complexity, cohesion, and semantics of language. These computational approaches have several advantages over human coding, which can be expensive and time-consuming. Compared with discourse analysis conducted by humans, computational approaches can provide instantaneous feedback, do not get fatigued, are reliable, and can provide greater detail on a wider number of dimensions (Hirschberg & Manning, 2015; McNamara et al., 2014). In the following sections, we provide examples of how NLP has been used in two different contexts: brief responses to prompts while reading and longer forms of writing (e.g., argumentative essays).

Automated Analyses of Short Natural Language Responses

Short natural language responses are commonly collected in psychological research. These responses may be answers to open-ended questions, contributions in dialogues or multiparty conversations, or think-aloud responses produced during complex tasks, such as reading or problem solving. Computational approaches have been leveraged to analyze these responses along a variety of dimensions, such as *accuracy, relevance*, *style*, *verbosity*, and *coherence.* Depending on the nature of the scoring task, the assessment of these verbal protocols can be simplistic (e.g., assessing the accuracy of students' responses on a short-answer test) or more complex (e.g., assessing students' use of strategies within a think-aloud task).

Constructed responses to texts have commonly been collected to examine cognitive processes underlying complex tasks such as reading or problem solving (Coté & Goldman, 1999; Denton et al., 2015; Magliano & Graesser, 2012; Magliano et al., 2011). There is substantial evidence that such constructed responses are sensitive to the processes involved in comprehending and learning new information (Magliano et al., 1999; Ozuru et al., 2004). For example, open-ended think-aloud protocols are assumed to capture a learner's thoughts and experiences while comprehending material and solving problems (K. A. Ericsson & Simon, 1984), whereas more targeted forms of constructed responses (e.g., self-explanation, question answering) have instructions that are intended to *modify* comprehension and learning (Magliano & Graesser, 2012; McNamara, 2004).

Although these responses have substantial value in the study of comprehension and learning, their use is substantially limited by the labor-intensive nature of protocol analysis (Magliano & Graesser, 2012). Thus, the past 2 decades have seen substantial advances in the application of NLP techniques to support the analyses of constructed responses (Allen et al., 2015; Landauer et al., 2007). These advances have been in the context of computer-based assessments of explanations and think-aloud protocols during reading comprehension (Gilliam et al., 2007; Magliano et al., 2011), the grading of short-answer questions (Leacock & Chodorow, 2003), and ITSs that require students to produce constructed

responses during interactive conversations (Graesser, 2016; Graesser et al., 2020; McCarthy et al., 2020). These automated systems incorporate a variety of NLP tools and algorithms to assess the responses, and make inferences about student comprehension, learning, and problem solving.

As one example, Magliano and colleagues developed the Reading Strategy Assessment Tool (RSAT; Magliano et al., 2011), which asks students to produce open-ended responses to prompts that are intended to engender a think-aloud response or answer to questions designed to tap into comprehension levels (e.g., why and how questions related to a recently read sentence). RSAT uses simple computer algorithms to analyze responses for evidence of comprehension processes, such as paraphrasing, bridging inferences, and elaborative inferences. Assessments of RSAT (Magliano et al., 2011; Millis & Magliano, 2012) report that RSAT does a reasonable job predicting objective comprehension scores and discriminating comprehension strategies.

In the Magliano et al. (2011) study, college students read a set of texts and answered direct and indirect questions while interacting with RSAT. They then completed multiple measures of comprehension, including the Gates-MacGinitie reading test and experimenter-generated, open-ended comprehension assessments. Researchers first examined the correlations between RSAT's overall measure of text comprehension and participants' performance on the two comprehension measures. The RSAT scores were correlated with performance on both the Gates-MacGinitie reading test ($r = 0.52$) and the open-ended comprehension assessments ($r = 0.45$), suggesting that RSAT successfully detected comprehension processes based on participants' responses. Correlations between these RSAT strategy scores (i.e., paraphrasing, bridging, elaborating) and expert human raters' identification of such strategies varied between .46 and .70, indicating that RSAT successfully detected comprehension strategies based on the constructed responses. Students' RSAT strategy scores accounted for approximately 21% of the variance in performance on the open-ended comprehension test. Specifically, higher comprehension was positively related to the generation of bridging and elaborations but negatively associated with paraphrasing. Overall, this work suggests that NLP can be used to identify the use of strategies *during reading* and that those strategies are predictive of individuals' ability to learn from the text.

Automated Analyses of Essay Responses

Beyond the scoring of brief constructed responses, NLP techniques have also been applied to the automated scoring of essays. Indeed, automated essay scoring (AES) has now reached a level of accuracy that the scoring of many classes of essays is as accurate as expert human raters (Attali & Burstein, 2006; McNamara et al., 2015; Shermis et al., 2010; Yan et al., 2020). Typically, AES systems are trained on a corpus of essays that have been rated by expert human raters according to a rubric. The corpus is divided into two sets of essays: a training set (used to train a model) and a testing set (used to examine the extent to which the model generalizes to new essays). Machine learning algorithms are applied to optimally fit the essays in the training set. The developed model is then applied to the essays in the testing set and these scores are compared to the human raters' scores. An AES model is considered successful if the scores between the computer and humans are similarly aligned to the scores between humans.

Shermis et al. (2010) reviewed the performance of the three most successful AES systems: e-rater developed at Educational Testing Service (Attali & Burstein, 2006; Burstein, 2003), Intelligent Essay Assessor developed at Pearson Knowledge Technologies (Landauer et al., 2003), and IntelliMetric developed by Vantage Learning (Elliot, 2003; Rudner et al., 2006). These systems have reported exact agreements with raters as high as the mid-80s, adjacent agreements in the high mid-90s, and correlations as high as the mid-80s. Just as impressive, these performance measures are slightly higher than agreement between trained human raters.

The performance of these AES systems has been sufficiently impressive to scale them for use in

educational applications. They have been used in a scoring process for high-stakes tests, such as the analytic writing assessment of the Graduate Management Admission Test (GMAT). The GMAT includes two 30-minute writing tasks to assess abilities related to critical thinking and communicating ideas. One task involves an analysis of an issue: Test takers receive an issue or opinion and are instructed to explain their point of view by citing relevant reasons or evidence. The second task is an analysis of an argument: Test takers read a brief argument, analyze the reasoning behind it, and critique the argument. The AESs are also used in electronic portfolio systems to help students improve writing by providing feedback on multiple features of their essays, similar to Criterion (Attali & Burstein, 2006) and MY Access (Elliot, 2003).

Although the practical use of AESs is undeniable, critics raise questions that challenge the ubiquitous use of these systems without some human expertise. Some critics voice concerns about aspects of writing that the AES systems are unlikely to capture, the ethics of using computers rather than teachers to teach writing, and differences in the criteria that humans versus the computers use to grade the essays (Calfee, 2000; P. F. Ericsson & Haswell, 2006). There is also a persistent third variable that robustly predicts essay scores, namely, the number of words in the essay. The incremental gain from computational algorithms beyond word count is often not reported or is unspectacular in some evaluations that have controlled for number of words. One barrier to overcoming this challenge is that human raters often base some aspects of their ratings on the number of words, and more words means more content, which results in better essays.

It is beyond the scope of this chapter to give a precise specification of the computational algorithms that have been implemented in AESs, particularly because some are proprietary or the published reports do not reflect the current systems. An edited volume by Shermis and Burstein (2003) provides detailed descriptions of many of the early systems to the extent that the corporations were comfortable in sharing the information. The e-rater AES (Attali & Burstein, 2006) scored essays on six areas of analysis aligned with human scoring criteria: errors in grammar, errors in word usage, errors in mechanics, style, inclusion of organizational segments (e.g., inclusion of a thesis statement or some evidence), and vocabulary content. The IntelliMetric AES (Elliot, 2003; Rudner et al., 2006) matched the words to a vocabulary of over 500,000 unique words, identified more than 500 linguistic and grammatical features that occur in the text, and analyzed this content through a word concept net, which examines similarities amongst words to determine their semantic meaning. These text characteristics were then associated with essays in each level of scoring rubric of the training corpus in order to discover which essay characteristics are most strongly diagnostic of each level.

The Intelligent Essay Assessor AES (Landauer et al., 2003) analyzed the words in the essay using latent semantic analysis (LSA; Landauer et al., 2007) and n-gram analyses (i.e., sequences of words, such as word pairs or triplets). The algorithm computes the similarity of the words and word sequences between the incoming essay and the essays associated with each level of the scoring rubric. LSA is an important method of computing the conceptual similarity between words, sentences, paragraphs, or essays because it considers implicit knowledge. LSA is a mathematical, statistical technique for representing knowledge about words and the world on the basis of a large corpus of texts that attempts to capture the knowledge of a typical test taker. The central intuition of LSA is that the meaning of a word, W, is reflected in the company of other words that surround the word in naturalistic documents (imagine 40,000 texts or 11 million words). Two words are similar in meaning to the extent that they share similar surrounding words. For example, the word "glass" will be highly associated with words of the same functional context, such as *cup*, *liquid, pour, shatter,* and *transparent*. These are not synonyms or antonyms that would occur in a dictionary but, rather, words that are likely to occur in the same documents as the word *glass*. LSA uses a statistical technique

called singular value decomposition (SVD) to condense a very large corpus of texts to 100 to 500 statistical dimensions (Landauer et al., 2007).

More recently, computational methods have been developed to capture better words' contexts, with the assumption that words are embedded in contexts defined by surrounding words. Similar to LSA, Word2Vec represents words as vectors but uses two-layer neural networks (rather than SVD) to train models (Mikolov et al., 2013). Bidirectional Encoder Representations from Transformers (BERT) expands the window of words' contextual embeddings by using deep learning to generate multiple contextual representations for each word (Devlin et al., 2018). Semantic models such as these generally compute the conceptual similarity between text excerpts (e.g., word, clause, sentence, essay) as the geometric cosine (i.e., 0–1) between the values and weighted dimensions of the excerpts.

A holistic grade for an essay has some value to the writer as an overall index of writing quality. However, more specific feedback on different characteristics of writing provides more useful information to the student and instructor. Is there a problem with spelling, vocabulary, syntax, cohesion of the message, missing content, elements of style, and so on? The e-rater AES has provided this feedback on 12 features in support of *Criterion*, an electronic portfolio of the students' writing. The portfolio of writing samples can be collected over time for students or instructors to track progress. Similarly, the LSA modules in the Intelligent Essay Assessor have been used in a system called *Summary Street* (Franzke et al., 2005) that gives feedback to the student on the quality of their summaries of a text. Summary Street identifies sentences that have low LSA relevance scores with other sentences in the text and low scores with expected information in different content categories of an underlying content rubric. An ideal summary would cover the expected content and have sentences that relate to one another conceptually (see Botarleanu et al., 2021, and Crossley, Kim, et al., 2019, for recent summarization algorithms).

Burstein et al. (2003) developed an automated scoring technology for the Criterion system at Educational Testing Service (ETS) that identifies the extent to which an essay contains particular components of an essay. The targeted categories of the essay include the title, the introductory material, a thesis statement, main ideas with respect to the thesis, supporting ideas, conclusions, and irrelevant segments. Trained human judges can identify these sections with kappa agreement scores of approximately 0.80 (between 0.86 and 0.95 on three different essay prompts). Kappa scores correct for guessing, adjust for the distribution of decisions, and vary between 0 (chance) and 1.0 (perfect agreement). Kappa scores have an advantage over correlations, but in practice the performance metrics lead to identical conclusions in this line of research. The kappa scores between the computer algorithms and human raters are respectable, typically above .70.

In addition to kappa and correlations, researchers routinely collect recall, precision, and F-measure scores between the computer decision on specific observations and the decision of a human judge (or alternatively between one judge and another judge). A recall score for a computer system is the proportion of computer decisions that receive the same decision as a human on the occurrence of a particular language/discourse features in an observation. The precision score is the proportion of computer decisions that agree with a human. The F-measure is 2 ∗ recall ∗ precision/(recall + precision), essentially an average between recall and precision scores. Burstein et al. (2003) reported that the scores between computer and human were approximately the same for these three metrics and averaged .76, depending on various parameters and criteria. Agreement between pairs of human judges averaged .91. Although not perfect, these automated systems are clearly making significant progress in identifying components of essays. These categories are important to identify in order to give informative guidance on how students can improve writing.

This push towards more detailed feedback from AES systems has coincided with the development of automated writing evaluation (AWE) systems, which are intended to move beyond simply

providing scores on students' essays. The purpose of AWE systems is to provide an opportunity for students to engage in writing practice and receive summative and formative feedback on their writing (Allen & Perret, 2016). These systems have been successfully integrated into a number of classroom environments and are commonly used in high-stakes writing assessments (Dikli, 2006). Although a substantial amount of research in this area still focuses on evaluating the accuracy of the automated scores (Warschauer & Ware, 2006; Yan et al., 2020), more recent research has also examined other aspects of writing, such as whether students can increase the quality of their essays after receiving system feedback (Roscoe et al., 2015) or whether they can more accurately monitor their own performance (Allen et al., 2015). A primary goal of computer-based writing systems should, therefore, be not only to provide accurate scores for students' performance but also to provide instruction and feedback that can help students to assess their own work more accurately.

Challenges in Automated Writing Evaluation

There are a number of methodological challenges that require attention for those who develop instructional systems designed to track and improve writing over time. One problem is that there are a limited number of standardized tests of writing achievement with norms that afford gauging progress over time. A second problem is that the available norm-referenced standardized tests, such as the Woodcock-Johnson or the Wechsler Individual Achievement Test, cover few writing skills and genres. A third problem is that the writing process is influenced by a number of factors associated with the pragmatic writing context, intended audience, writing prompts, time allotted for writing, mode of writing (handwriting vs. keyboard), choice of topics to write about, and characteristics of the writer (Graham & Perin, 2007).

The time-intensive nature of scoring written essays has traditionally limited teachers from giving a large number of writing assignments. This limitation can of course be circumvented by AES and AWE systems. There are also other methods other than the use of computers. For example, having students assess their own writing performance and development enhances writing skills (Andrade & Boulay, 2003; Graham & Perin, 2007; Ross et al., 1999). Teachers can also have students assess each other's writing. When learners are taught how to assess and provide feedback to their peers, their writing as well as their peers' writing improves (Cho et al., 2006; Graham & Perin, 2007).

Overview of NLP Assessment

Overall, this section illustrates the ways in which NLP techniques can be used to provide automated assessments of natural language across a variety of psychological and educational contexts. To illustrate the feasibility of these approaches, consider two students who are tasked with developing an opinion on whether uniforms should be required by schools. In this hypothetical task, the students would be asked to read multiple texts that provide information about this topic and would periodically be prompted to produce self-explanations of what they have just read. They would then be asked to provide a brief summary of their opinions on the issue. In this example, we have two primary sources of natural language that we can assess using NLP: the self-explanations and the summaries. Next, we illustrate a few ways that we may approach an NLP-based analysis of these self-explanations.

Consider the following excerpts from the self-explanations produced by the two students:

Student 1: I would never want anyone to tell me how to dress for school; that feels like such a violation of my freedom, and I like to be able to express myself creatively.

Student 2: I think this first passage is trying to indicate that one of the benefits of uniforms is that they can reduce perceptions of inequality. I wonder if the other passages will address the consequences as well.

In these two excerpts, we can see that the students are engaging in different types of text processing; Student 1 seems to be drawing on their own personal experiences when discussing

the texts, whereas Student 2 is paraphrasing the purpose of the first text and engaging in metacognitive processing as they anticipate the content of the remaining texts. We may choose to characterize these self-explanations along multiple dimensions to assess these processing differences. Table 17.1 provides an example of some of the metrics that may be calculated (for this analysis we used all of the self-explanations the students produced, not just the excerpts given above) across two categories. We can see from Table 17.1 that NLP analyses can allow us to assess the self-explanations along multiple dimensions.

The descriptive indices indicate that Student 1 generated *more* words in their self-explanations overall, but Student 2 wrote longer words on average. This provides us with some basic information about their verbosity during the task as well as the lexical sophistication of the students. The RSAT indices follow from Magliano et al. (2011) and provide more nuanced information about the specific strategies the students were engaged in during reading. Here, we can see that Student 1 was engaged in more shallow processing of the text than Student 2, as they predominantly engaged in paraphrasing compared with bridging or elaboration. Importantly, the indices shown here are just a small subset of the indices that could be calculated for these self-explanations but are intended to illustrate the power of NLP to assess students' natural language from a more multidimensional perspective compared with more standard holistic scores.

TABLE 17.1

Example Natural Language Processing (NLP) Indices for the Assessment of Self-Explanations

NLP variable	Type of variable	Student 1	Student 2
Number of words	Descriptive	453.00	236.00
Mean letters per word	Descriptive	3.90	5.23
Number of paraphrases	Strategy use	6.00	3.00
Number of bridges	Strategy use	4.00	5.00
Number of elaborations	Strategy use	2.00	5.00

INFERRING PSYCHOLOGICAL ATTRIBUTES AND PROCESSES FROM NATURAL LANGUAGE

Thus far, we have focused on research that examines computational systems' accuracy in scoring individuals' constructed responses. However, there is a growing body of work that examines ways in which NLP techniques can be used to model aspects of individual writers and their behaviors. Such approaches could be used to provide more nuanced information about the contextual factors influencing discourse processing and production. Notably, research has begun to examine whether NLP techniques can be used to model individual differences based on the linguistic features of individuals' produced discourse (e.g., constructed responses, essays). For example, recent work suggests that the cohesion of individuals' constructed responses (e.g., self-explanations, think-aloud responses) during reading is indicative of the coherence of their mental representation (Allen et al., 2016a). For instance, Allen and colleagues (2016) reported that the cohesion of constructed responses was higher when readers were prompted to self-explain compared to paraphrase, and that the cohesion of students' constructed responses increased over the course of self-explanation instruction and practice. Thus, automated analyses of the cohesion of students' constructed responses provides a window into the coherence of readers' mental representations. In turn, we can predict that the individual is a better reader if they produce language that is cohesive and lexically sophisticated (Allen et al., 2016a).

This work has been extended to multiple-document comprehension contexts. Allen et al. (2021) asked participants to generate constructed responses while reading multiple documents and then write an essay to assess integration across documents. The cohesion of the constructed responses within the individual documents was negatively related to essay quality. By contrast, cohesion of the constructed responses across the documents was positively related to essay quality. Further, compared to thinking aloud, strategic

instructions to either self-explain or evaluate sources enhanced across-document integration. As such, the NLP analyses of the cohesion of students' constructed responses provided both theoretical and practical insights into successful comprehension and learning processes and, in particular, strategic comprehension processes lending to more coherent mental representations of text.

Beyond individual differences, work has been conducted to examine emotions and other psychological states from written responses. One tool that has provided substantial advancements in the work in this domain is the Linguistic Inquiry and Word Count (LIWC) tool developed by Pennebaker et al. (2007). LIWC has been used to analyze a wide range of phenomena in psychology and education, far more than any other effort with automated systems. LIWC reports the percentage of words in a given text devoted to grammatical (e.g., "articles," "pronouns," "prepositions"), psychological (e.g., "emotions," "cognitive mechanisms," "social"), or content categories (e.g., "home," "occupation," "religion"). For example, "crying" and "grief" are words in the sad category, whereas "love" and "nice" are words that are assigned the positive emotion category. The mapping between words and word categories is not mutually exclusive because a word can map onto several categories. LIWC provides roughly 80 categories of words but also groups these word categories into broader dimensions, such as psychological constructs (e.g., causations, sadness) and personal constructs (e.g., work, religion). LIWC operates by analyzing a transcript of discourse and counting the number of words that belong to each category. A proportion score for each category is then computed by dividing the number of words in the discourse that belong to that category by the total number of words.

LIWC categories have been shown to be valid and reliable markers of a variety of psychologically meaningful constructs (Chung & Pennebaker, 2007; Pennebaker et al., 2003). The relative frequency of psychological words would obviously map onto relevant psychological constructs, and these references review such trends. However, the more counterintuitive finding that Pennebaker and his colleagues have documented is the role of the linguistic features of words. LIWC provides linguistic features that comprise function words, various types of pronouns, common and auxiliary verbs, different tenses, adverbs, conjunctions, negations, quantifiers, numbers, and swear words. Somewhat surprisingly, function words rather than the content words are diagnostic of many psychological states (Pennebaker, 2011). Function words are difficult for people to deliberately control and, thus, examining their use in text provides a nonreactive way to explore many social and personality processes.

Function word use has been linked to a wide range of individual differences. Function word use can vary as a function of sex, age, and social class (Pennebaker, 2011). For example, pronouns have been linked to psychological states such as depression and suicide in essays, natural conversations, and poetry (Rude et al., 2004; Stirman & Pennebaker, 2001). This work spurred research on language *style* (Pennebaker et al., 2003), represented by the use of function words across varied contexts. Language style has been linked to a number of factors, such as personality (Pennebaker, 2011) and emotional states (Tausczik & Pennebaker, 2010). More recently, researchers have examined how language styles dynamically shift during conversations. For example, Müller-Frommeyer and colleagues (2020) reported that language styles were significantly different in monologues compared with conversations and that this change was greater for conflict-based conversations compared with friendly conversations. Thus, NLP analyses have potential to reveal the nature of interactions in joint conversational contexts.

Inferring Emotions

One important application of LIWC and other similar tools has been the prediction of emotional states based on individuals' language. There are a number of different approaches to analyzing the affective content of text samples. One straightforward approach is to identify a small number of dimensions that underlie expressions of affect

(Samsonovich & Ascoli, 2006). This research was pioneered decades ago by Osgood and colleagues, who analyzed how people in different cultures rated the similarity of various emotion words (Osgood et al., 1975). His analyses converged on *evaluation* (i.e., good or bad), *potency* (i.e., strong or weak), and *activity* (i.e., active or passive) as the critical dimensions. These dimensions are aligned with valence and arousal, which are considered to be the fundamental dimensions of affective experience (Barrett et al., 2007; Russell, 2003).

A second approach is to conduct a more detailed lexical analysis of the text in order to identify words that are predictive of specific affective states of writers or speakers (Cohn et al., 2004; Crossley et al., 2017; Pennebaker et al., 2003). Other researchers have developed lexical databases that provide affective information for common words. For example, WordNet-Affect (Strapparava & Valitutti, 2004) is an extension of WordNet for affective content. Others have gone beyond the words and into a semantic analysis of the text. For example, Gill et al. (2008) analyzed blogs and reported that texts judged by humans as expressing fear and joy were semantically similar to emotional concept words (e.g., "phobia" and "terror" for "fear," but "delight" and "bliss" for "joy"). They used LSA (Landauer et al., 2007) and the hyperspace analogue to language model (Burgess et al., 1998) to automatically compute the semantic similarity between the texts and emotion keywords (e.g., "fear," "joy"). Although this method of semantically aligning text to emotional concept words showed some promise for fear and joy texts, it failed for texts conveying other emotions, such as anger and sadness. D'Mello and colleagues (2008, 2010) predicted student emotions using the language and discourse in tutorial dialogues with AutoTutor. They found that feedback, speech act categories (e.g., indirect hints), cohesion, negations, and other linguistic features successfully predicted student affect states that are frequent during tutoring, such as boredom, frustration, confusion, and engagement.

The fourth and most sophisticated approach to text-based affect sensing involves systems that construct affective models from a large corpora of world knowledge and apply these models to identify the affective tone in texts (Crossley et al., 2017; Pang & Lee, 2008; Wiebe et al., 2005). For example, the word "accident" is typically associated with an undesirable event so the presence of "accident" will increase the assigned negative valence of the sentence "I was held up from an accident on the freeway." This approach is sometimes called *sentiment analysis*, opinion extraction, or *subjectivity analysis* because it focuses on valence of a textual sample, rather than assigning the text to a particular emotion category (e.g., angry, sad).

Overview of NLP as a Tool of Modeling Psychological Processes

Overall, this section extends the section on assessment to reveal how NLP can be used to infer psychological and emotional states from natural language. We can illustrate these approaches by reconsidering Student 1 and Student 2 from the prior section. Above, we focused explicitly on assessments of the quality and types of strategies in which the students were engaged. However, we can also use NLP to infer the specific types of processes in which they are engaged as well as their emotional states during reading.

As shown in Table 17.2, the cohesion of the two students' self-explanations was quite varied. These indices indicated that the self-explanations generated by Student 2 were more cohesive than those written by Student 1, suggesting that

TABLE 17.2

Example Natural Language Processing (NLP) Indices for Inferring Psychological Processes and States From Self-Explanations

NLP variable	Type of variable	Student 1	Student 2
Number of connectives	Cohesion	10.00	15.00
Semantic overlap (LSA)	Cohesion	0.36	0.53
Positive words proportion	Emotion	0.53	0.21
Negative words proportion	Emotion	0.42	0.33

Note. LSA = latent semantic analysis.

Student 2 was potentially engaged in more integrative processes during reading, which has been linked to increased comprehension of the text information. On the other hand, Student 1 seemed to be engaged in more emotional processing of the text, which could have been linked to their focus on it related to their own life experiences. Thus, while this student was less likely to develop connections across the texts, they were more emotional, which could indicate that they were more motivated or engaged during the task. Overall, this example indicates that NLP techniques can be used to move beyond standardized assessments of natural language and provide context that is important to more fully understanding the learning process.

APPLICATION OF NLP TECHNIQUES TO INTELLIGENT TUTORING SYSTEMS

One common application of the work described above is to increase personalization and feedback delivery in educational technologies, such as ITSs. ITSs provide personalized learning through student modeling, which involves computational analyses that track the domain knowledge, strategies, and other psychological states of users (Chrysafiadi & Virvou, 2013; Woolf, 2009). ITSs adaptively respond to users by providing activities and feedback that are sensitive to these states and that advance instructional agendas. The interaction between the ITS and its users follows a large, if not an infinite number of alternative trajectories that attempt to fit constraints of both the student and the instructional goals. Thus, assessments of student responses are essential in any ITS. Such assessments are straightforward when the responses are selections among a fixed set of alternatives, as in the case of multiple-choice questions, true-false questions, ratings, or toggled decisions on a long list of possibilities. Challenges arise, however, when the student is prompted to input natural language within the ITS. In these circumstances, NLP techniques are required to provide scores and automated feedback to users on their responses.

A number of ITSs have been developed that process and respond to students using natural language. Examples include ITSPOKE (Litman et al., 2006), spoken conversational computer (Pon-Barry et al., 2004), tactical language and culture training system (L. W. Johnson & Valente, 2008), and Why-Atlas (VanLehn et al., 2007). In the following section, we describe three language-based ITSs to highlight work in this domain: AutoTutor (Graesser, 2016; Graesser, Lu et al., 2004), iSTART (McCarthy et al., 2020; McNamara et al., 2004), and the Writing Pal (Roscoe & McNamara, 2013).

AutoTutor

AutoTutor is an ITS that provides students with instruction on computer literacy, physics, critical thinking skills, and other technical topics by holding conversations in natural language (Graesser, 2016; Graesser et al., 2020; Graesser, Lu et al., 2004; Nye et al., 2014). AutoTutor shows learning gains of between 0.3 sigma (standard deviation units) and 0.8 sigma (Graesser, Jeon, & Dufty, 2008) compared with pretests or with a condition that has students read a textbook for an equivalent amount of time. The tutorial dialogues are organized around difficult questions and problems that require reasoning and explanations in the answers. For example, AutoTutor might ask, "If a lightweight car and a massive truck have a head-on collision, upon which vehicle is the impact force greater? Which vehicle undergoes the greater change in its motion, and why?" Such questions require the learner to construct approximately three to seven sentences and to exhibit reasoning in their responses.

When asked a question, students typically provide short answers during the first conversational turn, typically ranging from a few words to a couple of sentences. It takes a conversation to glean better insights into what the student knows even when the student has reasonable subject matter knowledge. The dialogue for one of these challenging questions consists of approximately 20 to 100 conversational turns between AutoTutor and the student. AutoTutor

provides feedback based on the student's input (positive or neutral vs. negative feedback), pumps the student for more information ("What else?"), prompts the student to fill in missing words, gives the student hints, fills in missing information with assertions, corrects erroneous ideas and misconceptions, answers the student's questions, and summarizes answers. These responses are important dialogue moves of AutoTutor and lead to the eventually construction of a full answer to the question across the dialogue.

There are many different ways to score the performance of AutoTutor (Graesser et al., 2007, 2020; Jackson & Graesser, 2006; VanLehn et al., 2007). One method is to score the extent to which students' verbal contributions match good answers to the question (called *expectations*) versus bad answers (called *misconceptions*). Students receive higher scores to the extent that they express more of the expectations and fewer of the misconceptions in the tutorial dialogue. Scores of expectation coverage and misconceptions can be computed during the first student turn or after they have finished the conversational dialogue. Students rarely articulate the expectations perfectly because natural language is much too imprecise, fragmentary, vague, ungrammatical, and elliptical. Thus, AutoTutor has used a number of semantic match algorithms to evaluate the extent to the students' verbal responses match any given expectation (Graesser et al., 2020).

Another method of assessing student performance in AutoTutor is to analyze the number and type of dialogue moves by AutoTutor that were selected to extract information from the student during the evolution of the answer. The system periodically identifies a missing expectation during the course of the dialogue and posts the goal of covering the expectation. When an expectation is posted, AutoTutor attempts to induce the student to articulate it by generating hints and prompts that encourage the student to fill in words and propositions. Specific prompts and hints are generated that maximize the student's filling in this content and boosting the match score above threshold.

A student's level of performance in AutoTutor can be measured by computing the number of AutoTutor pumps, hints, and prompts it requires for the student to generate an answer to a question. This was assessed in an analysis of four dialogue move categories that attempt to cover the content of particular expectations: pumps, hints, prompts, and assertions (Jackson & Graesser, 2006). The proportion of dialogue moves in these categories should be sensitive to student knowledge of physics (as measured by a pretest of physics with multiple-choice questions similar to the Force Concept Inventory; Hestenes et al., 1992). There is a continuum from the student supplying information to the tutor supplying information as we move from pumps to hints to prompts to assertions. The correlations with student knowledge reflected this continuum perfectly, with correlations of .49, .24, –.19, and –.40. For students with more knowledge of physics, AutoTutor can get by with pumps and hints, thereby encouraging the student to articulate the expectations. For students with less knowledge of physics, AutoTutor needs to generate prompts that elicit specific words or to assert the correct information, thereby extracting knowledge piecemeal or merely telling the student the correct information.

These analyses of student verbal responses through AutoTutor support a number of claims. First, there are several automated algorithms that can score whether particular sentences are covered in verbal responses that evolve in conversational turns over the course of a conversation. Second, the computer scores for sentential content matches have a moderate but unspectacular level of accuracy, at least compared with the scoring of lengthy essays. There is less content in a sentence than an essay, so this second conclusion is quite expected. On the other hand, the scoring of verbal responses is extremely high when the expectation unit is a single word, intermediate when it is a sentence, and high when it is an essay. Third, the scoring of verbal responses with AutoTutor requires an analysis of expected content and an assessment of the extent to which verbal responses match

the expected content. It is beyond the scope of AutoTutor to analyze content that is not on the radar of these expectations.

iSTART (Interactive Strategy Trainer for Automated Reading and Thinking)

iSTART (Levinstein et al., 2007; McNamara et al., 2004) is an ITS that helps high school, college, and adult literacy students learn and practice comprehension strategies to improve their comprehension of challenging expository text. iSTART has been shown to improve self-explanation quality, comprehension strategy use, and reading comprehension for readers from middle school through adulthood (Magliano et al., 2005; McCarthy et al., 2018; McNamara et al., 2007). iSTART is particularly effective in helping low knowledge and less skilled readers better understand challenging text.

iSTART includes modules for students to learn three macrostrategies: self-explanation (McNamara et al., 2017), question-asking (Ruseti et al., 2018), and summarization (Botarleanu et al., 2021; Crossley et al., 2019). Each module comprises brief lessons that provide the student with information on how to use the strategies, as well as microstrategies to facilitate students' application of the strategies. Students practice the strategies using natural language responses, such as generating self-explanations or summaries. A crucial aspect of iSTART's effectiveness is the feedback provided to students by a pedagogical agent as they type in responses to text using the comprehension strategies. Automated NLP algorithms detects the quality of the responses so that adaptive feedback can be provided to the student.

iSTART also includes two types of game-based practice (Jackson & McNamara, 2013). In generative games, students earn points for producing high-quality responses, such as explanations or summaries. In identification games, students read example responses to a text and earn points by correctly identifying which comprehension strategies were used in the examples. Students use their points to purchase customization features for students' avatars or to unlock new games. These "metagame" elements were designed to further enhance student motivation (Jackson & McNamara, 2013).

The core of iSTART is its focus on self-explaining challenging text using five empirically validated comprehension strategies: comprehension monitoring, paraphrasing, prediction, bridging, and elaboration. Comprehension monitoring is the reader's ability to assess their understanding of the text while reading. Paraphrasing is a restatement of the text in the reader's own words. Prediction is when a reader anticipates forthcoming information in a text either by making educated guesses or taking note of information that, if present, will aid in comprehension of a previous concept. Bridging is the act of drawing a connection between the current sentence to previous information in the text. Elaboration is using prior knowledge, either general or domain-specific, or logic to expand on the concepts in the text.

Several versions of the iSTART evaluation algorithm have been developed and assessed (McNamara et al., 2007). The ultimate goal was to develop an algorithm that was completely automated and did not rely on any human or hand-coded computations. The resulting algorithm uses a combination of both word-based approaches and semantic algorithms such as LSA (Landauer et al., 2007). Word-based approaches include a length criterion in which the student's explanation must exceed a specified number of content words that are in the text. The LSA-based approach relies on a set of benchmarks from the target text including the title of the passage, the words in the target sentence, and the words in the previous two sentences. The word-based algorithms provide feedback on shallow explanations (i.e., ones that are irrelevant or that repeat the target sentence). LSA augments the word-based algorithms by providing a deeper, qualitative assessment. More positive feedback is given for longer, more relevant explanations, whereas increased interactions and support are provided for shorter, less relevant explanations.

Students' self-explanations are assessed using a series of NLP algorithms. First, the response is

screened for metacognitive and frozen expressions (e.g., "I don't understand what they are saying here," "I'm bored"). If the explanation is dominated by the frozen expressions and contains little other content, then the pedagogical agent responds directly to those statements using a pool of responses that are randomly chosen, "Please try to make a guess about what this means" or "Can you try to use one of the reading strategies? Maybe that will help your understanding." After the frozen statements are removed from the explanation, then the remainder of the explanation is analyzed using both word-based and LSA-based methods (McNamara et al., 2007). If the length of the explanation does not reach a particular threshold, T, relative to the length of the target text, then the student is asked to add more to the explanation. The agent might then say, "Could you add to your explanation? Try to explain how it relates to something you already know." If the explanation does not have sufficient overlap in words or semantically meaning to the target and surrounding text, then it is assessed as irrelevant.

The explanation is further assessed in terms of its similarity to the target text. If it is too close to the target text in terms of the total number of words and the number of overlapping content words, as in the example below, then it categorized as a repetition. A repetition might receive feedback such as, "Try adding some more information that explains what the sentence means." The goal is to induce the student to go beyond the sentence. Paraphrasing is an excellent and optimal way to start an explanation, but the goal is usually to induce the student to go beyond paraphrasing by bringing in prior text or outside knowledge to the explanation. In that case, the student would receive feedback such as, "It looks like you've reworded the sentence. Now can you explain it by thinking about what else you know?" Once the explanation passes the thresholds for length, relevance, and similarity, feedback is provided on its quality. Students are provided with qualitative feedback, such as, "That's pretty good" for a medium-quality explanation and "You're doing a great job!" for a higher quality explanation. Lower quality explanations are just at the threshold and have little content that goes beyond the target text. They are provided with prompts and hints to help them use comprehension strategies.

iSTART can also adapt the difficulty of the texts that students read based on their performance in iSTART. When students' self-explanation quality is high, subsequent texts are more challenging, and vice versa, when self-explanation quality is low, subsequent text are adapted to students' ability levels (A. Johnson et al., 2017). Adapting the learning materials to the students' ability levels in iSTART leads to increased sense of learning (Watanabe et al., 2019) and leads to positive learning outcomes, specifically for less-skilled readers (McCarthy et al., 2018, 2020).

The accuracy of the iSTART evaluation algorithms has been assessed by computing linear equations based on a discriminate analysis of one data set and calculating its ability to predict human ratings for a variety of data sets (Boonthum et al., 2007; Jackson et al., 2010; McNamara et al., 2007; Millis et al., 2004). Across a number of evaluations, the iSTART algorithms have corresponded well to human ratings. McNamara et al. (2007) reported that algorithms corresponded highly with human evaluations of the self-explanations on two texts in the initial iSTART practice module; there was a 62% to 64% agreement between the algorithm and the human judgments ($r = .64 – .71$; $d' = 1.54 – 1.79$). The algorithms also successfully transferred to texts that were on a variety of science topics used in a classroom study that included 549 high school students who engaged in extended practice using iSTART across an academic year (Jackson et al., 2010). This study showed an $r = .66$ correlation between the human evaluations and iSTART's algorithms. This is remarkable given the variety of texts self-explained by the students in this study. Although this performance appears to be higher than AutoTutor, consider that the two systems target quite different information. iSTART assesses the quality of the student's self-explanation strategies whereas AutoTutor assesses the quality, depth, and accuracy of expected substantive content.

The analyses in this section support the claim that automated analyses are moderately successful in evaluating the quality of short verbal responses. A variety of algorithms have been used to compute semantic matches between student verbal responses and sentence expectations. Most of these algorithms are based on the overlap of content words and inferential content through LSA, but a few consider the order in which words are expressed and even deep symbolic analyses of the natural language. The performance of these computational analyses is moderately successful but not as impressive as automatic scoring of essays. We anticipate that future efforts will perform deeper analyses of the content with more sophisticated NLP.

Writing Pal

ITSs such as AutoTutor and iSTART focus on short, constructed responses. The Writing Pal is an ITS that has been developed as an extension to AWE systems that provide students with feedback on their writing. Specifically, the Writing Pal was designed to improve high school and college students' writing proficiency through explicit strategy instruction, deliberate practice, and automated feedback (Roscoe et al., 2014; Roscoe & McNamara, 2013). Contrary to the majority of computer-based writing systems (see Allen et al., 2016b, for a review), the Writing Pal strongly focuses on providing instruction and practice to use writing strategies in addition to providing opportunities to write essays with personalized feedback.

Strategy instruction in the Writing Pal system covers the three primary phases of the writing process: prewriting, drafting, and revising. In the system, these strategies are taught in the context of individual instructional modules that include Freewriting and Planning; Introduction Building, Body Building, and Conclusion Building; and Paraphrasing, Cohesion Building, and Revising. Each of these instructional modules contains multiple lesson videos, which are each narrated by an animated pedagogical agent. In these videos, the agent describes and provides examples of specific writing strategies. Once students have viewed the lesson videos, they can unlock mini-games that provide them with opportunities to practice the writing strategies in isolation before applying them in the context of a complete essay. In the Writing Pal, students can practice the strategies with identification mini-games, where they are asked to select the best answer to a particular question, or generative mini-games, where they produce natural language (typed) responses related to the strategies they are practicing.

An important component of the Writing Pal system is the AWE component (i.e., the essay practice component). This aspect of the Writing Pal contains a word processor in which students can write essays in response to a set of Scholastic Aptitude Test (SAT)–style prompts. Additionally, teachers have the option of adding their own prompts to the system. Once a student has completed an essay, it is submitted to the Writing Pal system for evaluation. As with other AES and AWE tools, the Writing Pal combines NLP and machine learning techniques to drive its automated feedback system and its adaptivity (Allen et al., 2016; McNamara et al., 2015). NLP techniques are used to assess students' essays across a variety of linguistic dimensions, such as the lexical sophistication or the organization of the essay. Once extracted, this information is used to drive essay scoring algorithms, which provide summative scores on a 6-point scale from *poor* to *great* similar to those used on the SAT rubrics (Roscoe et al., 2014). The formative feedback, on the other hand, provides information about strategies students can use to improve the quality of their essays. Formative feedback is an important component of writing development, as it provides knowledge about components of high-quality writing, as well as actionable recommendations on how to improve. The formative feedback in Writing Pal was developed with this in mind and provides recommendations that relate to multiple writing strategies. After they have read the feedback, students can revise their essays.

Students who have used the Writing Pal show significant improvements in writing skills, overall essay scores, and writing strategy knowledge

(Allen et al., 2015; Roscoe & McNamara, 2013). Receiving explicit writing strategy instruction helps students monitor their own strategy use and accuracy of their writing, which is beneficial when students need personalized writing feedback and their instructors are unavailable (Allen et al., 2015). The Writing Pal's success emphasizes the importance of individualized feedback for improving holistic writing quality.

THE FUTURE OF AUTOMATED LANGUAGE ANALYSES

This chapter surveyed the abundance of work that has been conducted on the automation of language analyses within the psychological domain. We know that language is an important construct in our understanding of a wide range of psychological and behavioral constructs; however, it is also highly complex, multimodal, and multidimensional (Allen et al., 2022; McNamara, 2021). Thus, the future of automated language analyses lies in the development of models that examine the complexity of language, considering language using multiple scales that range from examinations of word characteristics (e.g., the degree to which it is familiar, emotional, or abstract) to the organization of the discourse itself.

Such multidimensional analyses have the capacity to provide more nuanced information about the relations between language and psychological processes. For example, examination of languages at the word, sentence, and discourse level can provide more nuanced information about how certain experimental manipulations or individual differences influence discourse production and comprehension. Moreover, understanding and predicting human behavior calls for the integration of multiple sources of information from different modalities, such as gestures, eye movement, keystroke behaviors, and emotional responses. Thus, future research should not only consider the language being produced during psychological tasks but also consider other data sources that may be complementary. For instance, recent research has considered models that combine keystroke data with linguistic data (Allen et al., 2016), eye movements (Chukharev-Hudilainen et al., 2019), and click-stream data (Crossley et al., 2020). Multimodal work is likely to provide much more nuanced and robust understanding of discourse processes, cognition, and human behavior more broadly.

Overall, substantial progress has been made in our ability to provide automated assessments of natural language and discourse. This progress has been fueled by advances in computational power, statistical techniques, NLP tools, and theoretical understanding of discourse processes. These developments have undergirded techniques for scoring essays, analyzing characteristics of different types of writing, assessing text difficulty, assessing the accuracy, quality, and type of student contributions in tutoring systems, inferring psychological characteristics of speakers and writers, and detecting affective dimensions in discourse.

We expect that automated analyses of text and discourse will continue to grow and expand in the future. In this chapter, we have only covered a small slice of research at the intersections of computational modeling and psychology. Some colleagues will continue to have healthy skepticisms of the automated analyses of language and discourse. Others, however, will continue to discover how diverse aspects of psychological mechanisms can be captured using automated analyses of text and discourse. Both of these mindsets are needed to converge on automated assessments that most effectively and appropriately advance the field of psychology.

References

Allen, L. K., Creer, S. D., & Öncel, P. (2022). Natural language processing as a tool for learning analytics: Towards a multi-dimensional view of the learning process. In C. Lang, G. Siemens, A. F. Wise, D. Gašević, & A. Merceron (Eds.), *Handbook of learning analytics* (2nd ed., pp. 46–53). Society for Learning Analytics Research.

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016a). Cohesive features of deep text comprehension processes. In J. Trueswell, A. Papafragou, D. Grodner, & D. Mirman (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society in Philadelphia, PA* (pp. 2681–2686). Cognitive Science Society.

Allen, L. K., Jacovina, M. E., & McNamara, D. S. (2016b). Computer-based writing instruction. In C. A. MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research* (2nd ed., pp. 316–329). Guilford Press.

Allen, L. K., Magliano, J. P., McCarthy, K. S., Sonia, A., Creer, S., & McNamara, D. S. (2021). In T. Fitch, C. Lamm, H. Leder, & K. Tessmar (Eds.), *Proceedings of the 43rd Annual Conference of the Cognitive Science Society* (pp. 931–937). Cognitive Science Society.

Allen, L. K., Mills, C., Jacovina, M. E., Crossley, S., D'Mello, S., & McNamara, D. S. (2016). Investigating boredom and engagement during writing using multiple sources of information: The essay, the writer, and keystrokes. *Proceedings of the 6th International Learning Analytics & Knowledge Conference* (pp. 114–123). ACM.

Allen, L. K., & Perret, C. A. (2016). Commercialized writing systems. In D. S. McNamara & S. A. Crossley (Eds.), *Adaptive educational technologies for literacy instruction* (pp. 145–162). Taylor & Francis, Routledge. https://doi.org/10.4324/9781315647500-11

Allen, L. K., Snow, E. L., & McNamara, D. S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with Natural Language Processing techniques. *Proceedings of the International Learning Analytics & Knowledge Conference* (pp. 246–254). ACM.

Andrade, H. G., & Boulay, B. A. (2003). Role of Rubric-Referenced Self-Assessment in Learning to Write. *The Journal of Educational Research*, *97*(1), 21–34. https://doi.org/10.1080/00220670309596625

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater R V.2. *The Journal of Technology, Learning, and Assessment*, *4*(3), 1–30.

Barrett, L. F., Mesquita, B., Ochsner, K. N., & Gross, J. J. (2007). The experience of emotion. *Annual Review of Psychology*, *58*(1), 373–403. https://doi.org/10.1146/annurev.psych.58.110405.085709

Boonthum, C., Levinstein, I., & McNamara, D. S. (2007). Evaluating self-explanations in iSTART: Word matching, latent semantic analysis, and topic models. In A. Kao & S. Poteet (Eds.), *Natural language processing and text mining* (pp. 91–106). Springer. https://doi.org/10.1007/978-1-84628-754-1_6

Botarleanu, R.-M., Dascalu, M., Allen, L. K., Crossley, S. A., & McNamara, D. S. (2021). Automated summary scoring with ReaderBench. In *International Conference on Intelligent Tutoring Systems* (ITS 2021). Springer.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, and discourse. *Discourse Processes*, *25*(2–3), 211–257. https://doi.org/10.1080/01638539809545027

Burstein, J. (2003). The E-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Erlbaum.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, *18*(1), 32–39. https://doi.org/10.1109/MIS.2003.1179191

Calfee, R. (2000). To grade or not to grade. *IEEE Intelligent Systems*, *15*, 35–37.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. https://doi.org/10.1037/0022-0663.98.4.891

Chrysafiadi, K., & Virvou, M. (2013). Student modeling approaches: A literature review for the last decade. *Expert Systems with Applications*, *40*(11), 4715–4729. https://doi.org/10.1016/j.eswa.2013.02.007

Chukharev-Hudilainen, E., Saricaoglu, A., Torrance, M., & Feng, H.-H. (2019). Combined deployable keystroke logging and eyetracking for investigating L2 writing fluency. *Studies in Second Language Acquisition*, *41*(3), 583–604. https://doi.org/10.1017/S027226311900007X

Chung, C., & Pennebaker, J. (2007). The psychological functions of function words. In K. Fielder (Ed.), *Social communication* (pp. 343–359). Psychology Press.

Clark, A., Fox, C., & Lappin, S. (Eds.). (2013). *The handbook of computational linguistics and natural language processing*. John Wiley & Sons.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*(10), 687–693. https://doi.org/10.1111/j.0956-7976.2004.00741.x

Coté, N., & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 169–193). Lawrence Erlbaum.

Crossley, S. A., Karumbaiah, S., Ocumpaugh, J., Labrum, M. J., & Baker, R. S. (2020). Predicting math identity through language and click-stream patterns in a blended learning mathematics program for elementary students. *Journal of Learning Analytics*, *7*(1), 19–37. https://doi.org/10.18608/jla.2020.71.3

Crossley, S. A., Kim, M., Allen, L. K., & McNamara, D. S. (2019). Automated summarization evaluation (ASE) using natural language processing tools. In *Proceedings of the 20th International Conference of Artificial Intelligence in Education* (pp. 84–95). Lecture Notes in Computer Science, Vol. 11625. Springer.

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior Research Methods*, *51*(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. *Behavior Research Methods*, *49*(3), 803–821. https://doi.org/10.3758/s13428-016-0743-z

Dascalu, M., McNamara, D. S., Trausan-Matu, S., & Allen, L. K. (2018). Cohesion network analysis of CSCL participation. *Behavior Research Methods*, *50*(2), 604–619. https://doi.org/10.3758/s13428-017-0888-4

Denton, C. A., Enos, M., York, M. J., Francis, D. J., Barnes, M. A., Kulesz, P. A., Fletcher, J. M., & Carter, S. (2015). Text-processing differences in adolescent adequate and poor comprehenders reading accessible and challenging narrative and informational text. *Reading Research Quarterly*, *50*(4), 393–416. https://doi.org/10.1002/rrq.105

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pretraining of deep bidirectional transformers for language understanding. *arXiv*, 1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1), 1–35.

D'Mello, S. K., Craig, S. D., Witherspoon, A. W., McDaniel, B. T., & Graesser, A. C. (2008). Automatic detection of learner's affect from conversational cues. *User Modeling and User-Adapted Interaction*, *18*(1–2), 45–80. https://doi.org/10.1007/s11257-007-9037-6

D'Mello, S. K., Graesser, A., & King, B. (2010). Toward spoken human-computer tutorial dialogues. *Human-Computer Interaction*, *25*(4), 289–323. https://doi.org/10.1080/07370024.2010.499850

Dowell, N. M., Lin, Y., Godfrey, A., & Brooks, C. (2020). Exploring the relationship between emergent sociocognitive roles, collaborative problem-solving skills, and outcomes: A group communication analysis. *Journal of Learning Analytics*, *7*(1), 38–57. https://doi.org/10.18608/jla.2020.71.4

Dowell, N. M. M., Nixon, T. M., & Graesser, A. C. (2019). Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions. *Behavior Research Methods*, *51*(3), 1007–1041. https://doi.org/10.3758/s13428-018-1102-z

Elliott, S. (2003). IntelliMetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71–86). Erlbaum.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. MIT Press.

Ericsson, P. F., & Haswell, R. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. Utah State University Press. https://doi.org/10.2307/j.ctt4cgq0p

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary street: Computer support for comprehension and writing. *Journal of Educational Computing Research*, *33*(1), 53–80. https://doi.org/10.2190/DH8F-QJWM-J457-FQVB

Gill, A., French, R., Gergle, D., & Oberlander, J. (2008). Identifying emotional characteristics from short blog texts. In B. C. Love, K. McRae & V. M. Sloutsky (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 2237–2242). Cognitive Science Society.

Gilliam, S., Magliano, J. P., Millis, K. K., Levinstein, I., & Boonthum, C. (2007). Assessing the format of the presentation of text in developing a Reading Strategy Assessment Tool (R-SAT). *Behavior Research Methods*, *39*(2), 199–204. https://doi.org/10.3758/BF03193148

Graesser, A. C. (2016). Conversations with AutoTutor help students learn. *International Journal of Artificial Intelligence in Education*, *26*, 124–132.

Graesser, A. C., Hu, X., Rus, V., & Cai, Z. (2020). Conversation-based learning and assessment environments. In D. Yan, A. Rupp, & P. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 383–402). CRC Press/Taylor and Francis. https://doi.org/10.1201/9781351264808-21

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes*, *45*(4–5), 298–322. https://doi.org/10.1080/01638530802145395

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 180–192. https://doi.org/10.3758/BF03195563

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., Penumatsa, P., Ventura, M., Cai, Z., & Hu, X. (2007). Using LSA in AutoTutor: Learning through mixed initiative dialogue in natural language. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 243–262). Erlbaum.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. https://doi.org/10.1037/0022-0663.99.3.445

Griffiths, T. L. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, *135*, 21–23. https://doi.org/10.1016/j.cognition.2014.11.026

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, *30*(3), 141–158. https://doi.org/10.1119/1.2343497

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, *349*(6245), 261–266. https://doi.org/10.1126/science.aaa8685

Jackson, G. T., & Graesser, A. C. (2006). Applications of human tutorial dialog in AutoTutor: An intelligent tutoring system. *Revista Signos*, *39*, 31–48.

Jackson, G. T., Guess, R. H., & McNamara, D. S. (2010). Assessing cognitively complex strategy use in an untrained domain. *Topics in Cognitive Science*, *2*(1), 127–137. https://doi.org/10.1111/j.1756-8765.2009.01068.x

Jackson, G. T., & McNamara, D. S. (2013). Motivation and performance in a game-based intelligent tutoring system. *Journal of Educational Psychology*, *105*(4), 1036–1049. https://doi.org/10.1037/a0032580

Johns, B. T., & Jamieson, R. K. (2018). A large-scale analysis of variance in written language. *Cognitive Science*, *42*(4), 1360–1374. https://doi.org/10.1111/cogs.12583

Johnson, A., McCarthy, K. S., Kopp, K., Perret, C. A., & McNamara, D. S. (2017). Adaptive reading and writing instruction in iSTART and W-Pal. In Z. Markov & V. Rus (Eds.), *Proceedings of the 30th Annual Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 561–566). AAAI Press.

Johnson, L. W., & Valente, A. (2008). Tactical language and culture training systems: Using artificial intelligence to teach foreign languages and cultures. *Proceedings of the Twentieth Conference on Innovative Applications of Artificial Intelligence*. AAAI Press.

Jones, M. N. (2017). *Big data in cognitive science*. Psychology Press: Taylor & Francis.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice-Hall.

Kyle, K. (2021). Natural language processing for learner corpus research. *International Journal of Learner Corpus Research*, *7*(1), 1–16. https://doi.org/10.1075/ijlcr.00019.int

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, *50*(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *Modern Language Journal*, *102*(2), 333–349. https://doi.org/10.1111/modl.12468

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Erlbaum.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice*, *10*(3), 295–308. https://doi.org/10.1080/0969594032000148154

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405. https://doi.org/10.1023/A:1025779619903

Levinstein, I. B., Boonthum, C., Pillarisetti, S. P., Bell, C., & McNamara, D. S. (2007). iSTART 2: Improvements for efficiency and effectiveness. *Behavior Research Methods*, *39*(2), 224–232. https://doi.org/10.3758/BF03193151

Litman, D. (2016). Natural language processing for enhancing teaching and learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1), 4170–4176. https://doi.org/10.1609/aaai.v30i1.9879

Litman, D. J., Rose, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., & Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, *16*(2), 145–170.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods*, *44*(3), 608–621. https://doi.org/10.3758/s13428-012-0211-3

Magliano, J. P., Millis, K. K., The RSAT Development Team, Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). *Metacognition and Learning*, *6*, 131–154. https://doi.org/10.1007/s11409-010-9064-2

Magliano, J. P., Todaro, S., Millis, K. K., Wiemer-Hastings, K., Kim, H. J., & McNamara, D. S. (2005). Changes in reading strategies as a function of reading training: A comparison of live and computerized training. *Journal of Educational*

Computing Research, *32*(2), 185–208. https://doi.org/10.2190/1LN8-7BQE-8TN0-M91L

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology*, *91*(4), 615–629. https://doi.org/10.1037/0022-0663.91.4.615

McCarthy, K. S., Likens, A. D., Johnson, A. M., Guerrero, T. A., & McNamara, D. S. (2018). Metacognitive overload! Positive and negative effects of metacognitive prompts in an intelligent tutoring system. *International Journal of Artificial Intelligence in Education*, *28*(3), 420–438. https://doi.org/10.1007/s40593-018-0164-5

McCarthy, K. S., Watanabe, M., Dai, J., & McNamara, D. S. (2020). Personalized learning in iSTART: Past modifications and future design. *Journal of Research on Technology in Education*, *52*(3), 301–321. https://doi.org/10.1080/15391523.2020.1716201

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*(1), 1–30. https://doi.org/10.1207/s15326950dp3801_1

McNamara, D. S. (2017). Self-Explanation and Reading Strategy Training (SERT) Improves low-knowledge students' science course performance. *Discourse Processes*, *54*(7), 479–492. https://doi.org/10.1080/0163853X.2015.1101328

McNamara, D. S. (2021). Chasing theory with technology: A quest to understand understanding. *Discourse Processes*, *58*(5–6), 422–448. https://doi.org/10.1080/0163853X.2021.1917914

McNamara, D. S., Allen, L. K., Crossley, S. A., Dascalu, M., & Perret, C. A. (2017). Natural language processing and learning analytics. In G. Siemens & C. Lang (Eds.), *Handbook of learning analytics and educational data mining* (pp. 93–104). SOLAR. https://doi.org/10.18608/hla17.008

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 227–241). Erlbaum.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). Hierarchical classification approach to automated essay scoring. *Assessing Writing*, *23*, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavior Research Methods, Instruments & Computers*, *36*(2), 222–233. https://doi.org/10.3758/BF03195567

McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 397–421). Erlbaum.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionally. In *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates, Inc.

Millis, K., Kim, H. J., Todaro, S., Magliano, J. P., Wiemer-Hastings, K., & McNamara, D. S. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 213–221. https://doi.org/10.3758/BF03195566

Millis, K., & Magliano, J. (2012). Assessing comprehension processes during reading. In J. Sabatini, T. O'Reilly, & E. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment* (pp. 35–53). Rowman & Littlefield Education.

Müller-Frommeyer, L. C., Kauffeld, S., & Paxton, A. (2020). Beyond consistency: Contextual dependency of language style in monolog and conversation. *Cognitive Science*, *44*(4), e12834. https://doi.org/10.1111/cogs.12834

Nye, B. D., Graesser, A. C., & Hu, X. (2014). AutoTutor and family: A review of 17 years of natural language tutoring. *International Journal of Artificial Intelligence in Education*, *24*(4), 427–469. https://doi.org/10.1007/s40593-014-0029-5

Osgood, C. E., May, W. H., & Miron, M. (1975). *Cross-cultural universals of affective meaning*. University of Illinois Press.

Ozuru, Y., Best, R., & McNamara, D. S. (2004). Contribution of reading skill to learning from expository texts. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Cognitive Science Society* (pp. 1071–1076). Erlbaum.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Pennebaker, J. W. (2011). *The secret life of pronouns: What our words say about us*. Bloomsbury Press/Bloomsbury Publishing. https://doi.org/10.1016/S0262-4079(11)62167-2

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX: LIWC.net (https://www.liwc.net)

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language. use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pon-Barry, H., Clark, B., Schultz, K., Bratt, E. O., & Peters, S. (2004). Advantages of spoken language interaction in tutorial dialogue systems. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems* (pp. 390–400). Springer-Verlag.

Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. *Computers and Composition*, *34*, 39–59. https://doi.org/10.1016/j.compcom.2014.09.002

Roscoe, R. D., & McNamara, D. S. (2013). Writing Pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, *105*(4), 1010–1025. https://doi.org/10.1037/a0032340

Roscoe, R. D., Snow, E. L., Allen, L. K., & McNamara, D. S. (2015). Automated detection of essay revising patterns: Application for intelligent feedback in a writing tutor. *Technology, Instruction, Cognition, & Learning*, *10*(1), 59–79.

Ross, J. A., Rolheiser, C., & Hogaboam-Gray, A. (1999). Effects of self-evaluation training on narrative writing. *Assessing Writing*, *6*(1), 107–132. https://doi.org/10.1016/S1075-2935(99)00003-3

Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, *18*(8), 1121–1133. https://doi.org/10.1080/02699930441000030

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric essay scoring system. *The Journal of Technology, Learning, and Assessment*, *4*(4), 1–22.

Ruseti, S., Dascalu, M., Johnson, A. M., McNamara, D. S., Balyan, R., McCarthy, K. S., & Trausan-Matu, S. (2018). Scoring summaries using recurrent neural networks. In R. Nkambou, R. Azevedo, & J. Vassileva, (Eds.), *Proceedings of the 14th International Conference on Intelligent Tutoring Systems (ITS) in Montreal, Canada* (pp. 191–201). Springer.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*(1), 145–172. https://doi.org/10.1037/0033-295X.110.1.145

Samsonovich, A., & Ascoli, G. (2006). Cognitive map dimensions of the human value system extracted from natural language. In B. Goertzel & P. Wang (Eds.), *Advances in artificial general intelligence: Concepts, architectures and algorithms* (pp. 111–124). IOS Press.

Shermis, M. D., & Burstein, J. (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Erlbaum.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In E. Baker, B. McGaw, & N. S. Petersen (Eds.), *International encyclopedia of education* (3rd ed.). Elsevier. https://doi.org/10.1016/B978-0-08-044894-7.00233-5

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, *63*(4), 517–522. https://doi.org/10.1097/00006842-200107000-00001

Strapparava, C., & Valitutti, A. (2004). *WordNet Affect: An affective extension of WordNet*. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation* (pp. 1083–1086). European Language Resources Association.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rosé, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science*, *31*(1), 3–62. https://doi.org/10.1080/03640210709336984

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, *10*(2), 157–180. https://doi.org/10.1191/1362168806lr190oa

Watanabe, M., McCarthy, K., & McNamara, D. S. (2019). Examining the effects of adaptive task selection on students' motivation in an intelligent tutoring system. In *Proceedings of the 9th International Conference on Learning Analytics and Knowledge* (pp. 161–162). SOLAR.

Wiebe, J., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, *39*(2–3), 165–210. https://doi.org/10.1007/s10579-005-7880-9

Woolf, B. P. (2009). *Building intelligent interactive tutors*. Morgan Kaufmann Publishers.

Yan, D., Rupp, A. C., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. Chapman and Hall, CRC Press.