Behav Res (2012) 44:608–621
DOI 10.3758/s13428-012-0211-3

Computer-based assessment of student-constructed responses

Joseph P. Magliano · Arthur C. Graesser

Published online: 12 May 2012

Abstract Student-constructed responses, such as essays, short-answer questions, and think-aloud protocols, provide a valuable opportunity to gauge student learning outcomes and comprehension strategies. However, given the challenges of grading student-constructed responses, instructors may be hesitant to use them. There have been major advances in the application of natural language processing of student-constructed responses. This literature review focuses on two dimensions that need to be considered when developing new systems. The first is type of response provided by the student—namely, meaning-making responses (e.g., think-aloud protocols, tutorial dialogue) and products of comprehension (e.g., essays, open-ended questions). The second corresponds to considerations of the type of natural language processing systems used and how they are applied to analyze the student responses. We argue that the appropriateness of the assessment protocols is, in part, constrained by the type of response and researchers should use hybrid systems that rely on multiple, convergent natural language algorithms.

Keywords Natural language processing · Student constructed responses · Computer-based assessment

J. P. Magliano (✉)
Northern Illinois University,
DeKalb, IL, USA
e-mail: jmagliano@niu.edu

A. C. Graesser
The University of Memphis,
Memphis, TN, USA

Computer-based assessment of student-constructed responses

Educators are often faced with a difficult choice when developing tasks to evaluate student learning. Specifically, they have the option of closed responses (e.g., multiple-choice tests, true/false tests) or student-constructed responses (e.g., short answers to questions or long essays). Although possible, it is not easy to develop multiple-choice tests that target important constructs that are fortified by cognitive theories of learning and comprehension (Graesser, Ozuru, & Sullins, 2009; Magliano, Millis, Ozuru, & McNamara, 2007; VanderVeen et al., 2007). Many would agree that open-ended responses provide a rich opportunity for students to engage in the active generation of knowledge, problem solving, explanation of difficult concepts, and reasoning that are not required or emphasized in closed-response test items. Open-ended items also provide opportunities for instructors to provide substantive feedback to students at a relatively fine-grained level, which is important for enhancing learning outcomes (Shute, 2008). Unfortunately, however, educators often do not to use open-ended items, because they do not have the time or, perhaps, expertise to evaluate them and provide critical feedback.

Thankfully, over the past 2 decades, there have been substantial advances in the application of natural language processing techniques to support the analyses of student-constructed responses (Graesser & McNamara, 2012; Landauer, McNamara, Dennis, & Kintsch, 2007; Shermis, Burstein, Higgins, & Zechner, 2010). We define student-constructed responses as those that require a student to produce an answer in natural language that may range from a couple of sentences to several paragraphs. These advances have been in the context of computer-based assessments of explanations and think-aloud

protocols during reading comprehension (Gilliam, Magliano, Millis, Levinstein, & Boonthum, 2007; Magliano, Millis, the RSAT Development Team, Levinstein, & Boonthum, 2011), the grading of essays and text summaries (Attali & Burstein, 2006; Burstein, Marcu, & Knight, 2003; Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Landauer, Laham, & Foltz, 2003), the grading of short-answer questions (Leacock & Chodorow, 2003), and intelligent tutoring systems and trainers that require students to produce constructed responses during interactive conversations (Graesser, Jeon, & Dufty, 2008; Litman et al., 2006; McNamara, Levinstein, & Boonthum, 2004; VanLehn et al., 2007). These can take the form of directed responses to specific questions or less directed think-aloud and self-explanation responses. These automated systems incorporate a variety of natural language processing tools and algorithms to assess the responses, make inferences about student learning, and make decisions to provide appropriate feedback. As such, these systems have the promise to provide assessment aides that can help ease the burden of including student-constructed responses.

Skeptics of automated constructed response systems have frequently raised worries about their use and accuracy (Calfee, 2000; Ericsson & Haswell, 2006). The skeptics point out important aspects of student responses that the automated systems fail to capture, the ethics of using computers rather than teachers to evaluate writing, and differences in the criteria that humans versus computers use in the automated analyses. Many of these concerns could also be raised about humans who score essays or give feedback on constructed responses. Indeed, humans often do not have the time or expertise to evaluate constructed verbal responses accurately, especially when there are many to evaluate. Just as human graders have limitations, so do the automated systems. Their assessments will always be probabilistic, rather than absolute. Automated systems cannot handle most forms of metaphor, literary devices, and content that is unrelated to the topics targeted by the automated systems, which limits their application across disciplines. Moreover, as noted by the critics, constructed responses may reveal rich processing that may not be amenable to computational analyses. To the extent that the automated systems fall prey to these limitations, they are best used as an aid to educators, rather than as a replacement for human graders.

Although we acknowledge these potential limitations of automated assessments, it is equally important to appreciate that the scoring and feedback can be useful, even when it is only moderately accurate. For example, intelligent tutoring and training systems have produced substantial improvements in subject matter learning even when the feedback to the student has been almost but not quite accurate (Graesser, Lu, et al., 2004; Millis et al., in press; VanLehn et al., 2007). Automated feedback on student writing over weeks can improve the writing even when the automated feedback is not quite perfect (Attali & Burstein, 2006; Elliott, 2003; McNamara et al., 2012). For these reasons, it is appropriate to take stock of the status of automated assessment of constructed responses with respect to use and accuracy and with the specific goal of making recommendations for the development of new systems that have the viability to be used in the classroom and other educational contexts.

The goal of this article is to provide a review of the literature that identifies dimensions that one needs to consider with developing new assessment protocols. Researchers developing new systems should consider at least two dimensions. First, one needs to consider the *use* of different types of constructed responses and the kinds of assessment that they can deliver. That is, the goals of the system constrain how instructors will use them, the length of the verbal responses, and the expected accuracy of the computer-based educational tool. There are important differences among (1) the assignment of a grade to an essay in a class or a high-stakes examination, (2) information provided to an instructor to evaluate the overall progress of a student, and (3) timely feedback to the student during tutorial dialogue or think-aloud protocols during reading. Second, one needs to consider the strengths and weakness of natural language processing tools with respect to the *accuracy* of the assessments and aptness of the resulting feedback. This is intimately related to the protocols that are developed to assess student responses and that utilize these algorithms. The natural language processing advances all compare the students' verbal responses with expected responses that are specified by some rubric, or what we call *expectations* (words, symbolic expressions, sentences, paragraphs, scripts, sets of essays graded in the past, criterial dimensions). The semantic matches between the student responses and the expectations are computed by a variety of modules developed in computational linguistics, cognitive science, and information sciences (Burgess, Livesay, & Lund, 1998; Jurafsky & Martin, 2008; Landauer et al., 2007). However, there are important methodological decisions that need to be made when constructing the expectations that are related to the type of response produced by a student. Additionally, as we will argue, given the strengths and weaknesses of different natural language algorithms, it is advisable to incorporate multiple natural language algorithms into the assessment protocols.

Types of student-constructed responses

Someone naïve to the advances in the computer-based assessment of natural language may assume that the types of

student responses that are readily handled by these systems are limited. However, this is not the case. The same wide variety of student-constructed responses extant in psychological and educational research and in educational practice are represented in computer-based assessment and tutoring systems. Moreover, the purposes for collecting and analyzing these responses mirror the same reasons for collecting them in research and educational contexts. We draw a broad distinction between systems developed to support meaning making (i.e., the pedagogical goal of students' active generation of information) and systems developed to provide insights into the quality and nature of one's level of understanding (i.e., the pedagogical goal of achieving reliable or valid information). These two classes of systems serve different functions in assessment and learning. However, this distinction is important to consider a priori to the development of assessment protocols, which will be discussed in the next section.

Meaning-making student responses

Meaning-making student responses occur when students engage in an activity that provides a window into what they do to comprehend an experience, such as reading a text or understanding a learning environment. Think-aloud and self-explanation responses (Bråten & Strømsø, 2003; Chi, de Leeuw, Chiu, & LaVancher, 1994; Coté, Goldman, & Saul, 1998; McNamara et al., 2004; Trabasso & Magliano, 1996) and dialogues between tutee and tutor (Graesser, Person, & Magliano, 1995) are examples of meaning-making student responses that have been investigated in the cognitive sciences and have been implemented in computer systems during the last 15 years. The developers have the goal of assessing and providing feedback on both the content and dynamic processes of learning and comprehension.

Researchers who target meaning-making responses may have participants provide them by writing, typing, or expressing them orally, the latter of which requires transcription (Magliano & Millis, 2003) or satisfactory speech recognition (D'Mello, Dowell, & Graesser, 2011).The nature and quality of oral versus typed responses are very similar but not equivalent (D'Mello et al., 2011; Muñoz, Magliano, Sheridan, & McNamara, 2006).

As an initial example, consider the information that is elicited in think-aloud data when students read text. Table 1 depicts example typed think-aloud protocols produced while texts on how cancerous tumors develop were read. The examples reveal some processes that contribute to constructing a coherent text representation (e.g., Magliano, 1999; Trabasso & Magliano, 1996) and illustrate individual differences in the extent to which readers enact them (see the Appendix for the text excerpt). For example, participant 1 produced a protocol that reflects a very close paraphrase of the sentence that was just read, which reflects a shallow approach to comprehension because the student is not contributing inferences and explanations (Cote et al., 1998; Magliano & Millis, 2003). In contrast, participants 2 and 3 produced statements that establish how the current sentence is related to the more global discourse structure, which is one of the primary bases for constructing a coherent representation of the text (Graesser, Singer, & Trabasso, 1994; W. Kintsch, 1998). The last example illustrates that some students do elaborate beyond the discourse contexts, but these types of responses support comprehension only to the extent that they are relevant to the text. As we will discuss below, detecting and classifying these kinds of elaborative responses is a major challenge because it is difficult to anticipate the variety of responses that students can produce (Magliano et al., 2011; Millis, Magliano, Todaro, & McNamara, 2007).

Both think-aloud and self-explanation protocols (which are very similar to think-aloud responses) are very time consuming to analyze by hand, which of course motivated the goal of developing computer-based algorithms (Magliano & Millis, 2003). Nonetheless, the development of automated assessments is warranted because these protocols expose individual differences in approaches to reading, particularly in the realm of prior domain knowledge and use of comprehension strategies (Chi, Bassok, Lewis, Reimann, & Glaser, 1989; Coté et al., 1998; Magliano & Millis, 2003, Magliano et al., 2011). For example, less skilled readers tend to paraphrase the sentence they just read, whereas skilled readers tend to bridge explicit statements in the text and elaborate the material with text-relevant inferences (Magliano & Millis, 2003; Magliano et al., 2011). Moreover, computer systems that train students to think aloud or self-explain effectively have been shown to promote comprehension (McNamara et al., 2004).

A central challenge of computer-based assessment or training systems is to identify the processes enacted by a reader that are readily apparent in the example protocols in Table 1. As is discussed below, one solution to this challenge is to infer the processes on the basis of the content in the protocols. If the readers' content has words that match or are close synonyms to the words in the current sentence, the researcher can readily infer that the readers are restating or paraphrasing the sentence. But if the words draw heavily upon content from sentences in the prior discourse, readers are likely to be engaging in the process of generating bridging inferences (Magliano & Millis, 2003; Magliano et al., 2011). Strategy detection systems, such as the Reading Strategy Assessment Tool (RSAT; Magliano et al., 2011), provide reasonable estimates of broad categories of process/strategy, such as paraphrasing and bridging, in the context of generating think-aloud or self-explanation protocols while text is read. This is accomplished by basing assessments of processes on the content of the protocols (e.g., information from current or prior discourse sentences).

Table 1 Example typed think-aloud protocols for the sentence "A message within each receptor cell becomes activated" (from the text "How Cancer Develops")

Participant	Clause	Protocol	Process
1	1.	A message is activated	paraphrase
2	1.	The growth factor attaches to the cell and	bridge
	2.	A message is activated	paraphrase
3	1.	Cells influence their surroundings.	bridge
	2.	They transmit signals: for example growth signals.	bridge
4	1.	I am wondering how they can use this information	elaboration
	2.	to find a cure for cancer.	elaboration

Another type of meaning-making responses occurs in tutorial dialogue when a human or computer tutor interacts with the student in natural language (Graesser, D'Mello, & Cade, in press; Graesser et al., 2008; VanLehn et al., 2007). That is, the student is given a difficult question to answer or a problem to solve. The student and tutor collaboratively interact in a multi-turn conversation in route to covering a good answer or solution. The student's response is distributed over many conversational turns, with tutor dialogue moves that elicit or elaborate on student contributions.

Table 2 presents an example dialogue between AutoTutor and student s who were collaboratively answering a main question: "The sun exerts a gravitational force on the earth as the earth moves in its orbit around the sun. Does the earth pull equally on the sun? Explain why" (see Graesser & McNamara, 2011). The students' contributions in these collaborations vary widely in volume and quality. AutoTutor classifies their contributions into different speech act categories, such as metacognitive statements (conversational turns 2 and 14), questions (turns 4 and 10), short answers (turn 6),

Table 2 Example AutoTutor–student dialogue with classified dialogue moves of AutoTutor, classified speech acts of the student, and comments (reprinted from Graesser & McNamara, 2010)

AutoTutor-1: *The sun exerts a gravitational force on the earth as the earth moves in its orbit around the sun. Does the earth pull equally on the sun? Explain why.* (MAIN QUESTION)

Student-2: *I don't know.* (A metacognitive speech act that expresses what the student knows.)

AutoTutor-3: *Could you at least try to give me an answer? That way, we'll have something as a starting place to work from.* (PUMP. This dialogue move encourages the student to say more, but does not direct the student to express particular information.) *How does Newton's third law of motion apply to this situation?* (HINT. This dialogue move leads the student to articulate a particular sentence, clause, or proposition.)

Student-4: *Does Newton's law apply to opposite forces?* (This is a YES/NO Verification Question; the student wants confirmation that the idea is correct.)

AutoTutor-5: *Yes, right!* (ANSWER. AutoTutor answers the student's question.) *Hum.* (DISCOURSE MARKER. This connects different dialogue moves within a turn.)

There's still something missing from your explanation. (DISCOURSE MARKER. AutoTutor lets the student know more is expected.) *Newton's Third Law refers to the forces exerted by one body on another _____?* (PROMPT. This dialogue move encourages the student to fill in a particular word or phrase. Prompts are needed for students who have lower knowledge and verbal ability).

Student-6: *body* (Short Answer)

AutoTutor-7: *Bravo!* (POSITIVE FEEDBACK)

Let's move on. (DISCOURSE MARKER) *What can you say about the magnitude of the force that the earth exerts on the sun?* (HINT)

Student-10: *What is the difference between a force and a vector?* (Student question.)

AutoTutor-11: *Force is any influence that tends to accelerate an object; a push or pull; measured in Newtons. Force is a vector quantity and vector is an arrow whose length represents the magnitude of a quantity and whose direction represents the direction of the quantity.* (ANSWER. This answer is fetched from the textbook)

Okay, I bet you can get this! (DISCOURSE MARKER) *What is going on in this situation between the earth and sun?* (HINT)

Student-12: *The sun is pulling on the earth because it is bigger.* (Student misconception).

AutoTutor-13: *No.* (NEGATIVE FEEDBACK) *the forces are equal in magnitude. Um, well, for every force there is a reaction force, such that the two forces constitute an action reaction force pair. The two forces have opposite direction and equal magnitude. No matter how different the masses of the two objects are, they will exert reaction forces on each other in the same magnitude*. (CORRECTION AutoTutor corrects the student's misconception.)

Do you understand? (COMPREHENSION GAUGING QUESTION)

Student-14: Yeah. I see. (Metacomprehension judgement)

misconceptions (turn 12), and lengthier assertions that are not represented in the protocol. The length of a student contribution in one conversational turn is quite short, only one word to two sentences over 90 % of the time in both human tutoring and AutoTutor (Graesser et al., 1995; Graesser, Jeon, Yang, & Cai, 2007). The tutor needs to formulate dialogue moves to get the student to express more information and to be responsive to what the student is saying.

Both human tutors and AutoTutor have a number of dialogue moves that guide the tutorial interaction. The initial dialogue move is frequently *feedback* on the quality of what the student just expressed in the previous turn, as exemplified in turns 5, 7, and 13 in Table 2. The short feedback can vary from extremely negative to neutral to very positive. The tutor has a number of dialogue moves that try to get the student to express a good piece of information (i.e., a word- or sentence-length expectation), such as *pumps* (turn 3), *hints* (turns 3, 7, and 11), and *prompts* for specific words (turn 5). Instead of trying to extract information from the student, the tutor also delivers information to the student in the form of *assertions*, *answers* to student questions (turn 11), *corrections* (turn 13), and *summaries*. However, a good tutor and AutoTutor try to get the student to do the talking and doing, rather than lecturing to the student. Tutors periodically ask *comprehension-gauging questions* (turn 13) on whether the student is understanding the recent collaborative changes. Students often give incorrect answers to the comprehension-gauging questions because their metacognitive knowledge is modest (Graesser, D'Mello, & Person, 2009). Indeed, it is the high-knowledge students who tend to answer that they don't quite understand (Graesser et al., 1995).

The conversations managed by AutoTutor are not always perfect but are sufficient for students to get through the sessions with minimal difficulties. Person, Graesser, and the Tutoring Research Group (TRG) (2002) reported a study in which half of the turns were generated by AutoTutor and half were substituted by a human expert tutor on the basis of the dialogue history. Participants who did not undergo tutoring (called third-person bystanders) were presented these tutoring moves in a written transcript and were asked to decide whether each was generated by a computer or a human. Signal detection analyses revealed that the bystanders had zero d' scores in making these discriminations. AutoTutor therefore passed a *bystander Turing test* for individual tutoring turns. Of course, a bystander would presumably be able to discriminate whether a sequence of turns was part of a dialogue with AutoTutor versus a human tutor. AutoTutor is close enough to human tutorial dialogue to keep the conversation going and also to promote active learning.

A more critical analysis of automated tutors like AutoTutor has unveiled four major challenges (Graesser et al., 2008).

1. *Errors in interpreting the content of student turns*. AutoTutor's evaluation of whether an expectation (or misconception) is expressed by a student is significantly correlated with the evaluation of experts ($r = .50$) and almost as high as the correlation between two experts ($r = .63$), but the correlation is far from perfect. Sometimes when such errors occur, the students get frustrated and conclude that the tutor is not listening.
2. *Misclassification of the speech acts in student turns*. The student turns are segmented into speech acts, and each speech act is assigned to 1 of approximately 20 speech act categories. The accuracy of classifying the student speech acts into categories varies from .87 to .96, which is almost but not quite perfect. The dialogue coherence breaks down when some misclassification errors occur, which ends up confusing students.
3. *Ignoring student contributions that fail to match any expectation or misconception*. The student may conclude that AutoTutor is unresponsive to the extent that this occurs. It should be noted that human tutors also fail to meaningfully respond to student contributions that are not on their content radar (Chi et al., 2004; Graesser et al., 1995).
4. *Failure to answer student questions*. Human and automated tutors can handle only a subset of student questions, so the student eventually stops asking them. Student questions are not prevalent in human tutoring (Graesser et al., 1995), so this is not a serious obstacle in automated tutors. However, active inquiry and self-regulated learning is encouraged in most pedagogical theories.

In sum, meaning-making responses provide opportunities to assess how a student is comprehending a text or co-constructing meanings with a tutor. In the context of reading or engaging in simulated dialogue, they can potentially provide insights into the dynamic processes that support learning. The successful detection of the target processes and outcomes rests on developing a set expectations or benchmarks that are representative of these processes and having the appropriate computational linguistic tools to compare the protocols with these expectations. In the section on natural language approaches for analyzing responses, we will discuss some approaches and guidelines for accomplishing this task.

Responses reflecting the products of comprehension

A second type of system for handling student-constructed responses is developed for the purpose of assessing the products of comprehension. These types of responses are generally used to assess student proficiencies in some task, rather than focused tracing of the process of accomplishing that task. These systems are beginning to be used for high-stakes assessments in addition to the formative assessments that facilitate the learning process. Students are asked to produce responses to *prompts* (i.e., questions, essay specifications) that are intended to evaluate their level of

understanding of the material that they have learned. These can take the form of short-answer questions (Cai et al., 2011; Leacock & Chodorow, 2003; Magliano et al., 2011) that require about one to three sentences to answer. Although one-word responses are possible, these typically do not require natural language processing algorithms to analyze. The more common span of texts in these assessments is longer essays or summaries (Attali & Burstein, 2006; Britt, Wiemer-Hastings, Larson, & Perfetti, 2004; Rudner, Garcia, & Welsch, 2006).

Although developing systems to code these types of responses is a challenge, quite surprisingly, many of these systems are as reliable as human raters (Elliott, 2003; Landauer et al., 2003; Shermis et al., 2010). These systems have had exact agreements with humans on a 5-point scale as high as the mid-80s, adjacent agreements in the high mid-90s, and correlations as high as the mid-80s. These performance measures are, surprisingly, a bit higher than agreement between trained human raters. Although these automated essay graders have impressive accuracy for some subject matters according to virtually any criteria, the performance is not sufficient for some subject matters. Defenders of high-stakes tests require extremely high reliability scores with human graders. Insufficient accuracy in grading is a frequent criticism of the skeptics, even though the criteria and accuracy of human essay graders is theoretically suspect and frequently limited (McNamara, Crossley, & McCarthy, 2010). There is no ideal gold standard for essay grading, but the performance of the automated systems is quite impressive, as compared with human grading, for many tasks and subject matters.

These short and long responses present different challenges for automatic scoring. Consider the short-answer questions that occur in the RSAT (Magliano et al., 2011). These questions are embedded in the text so that the test takers answer the questions while they are reading. Table 3 shows a text excerpt, question prompt, an expert answer, and sample student answers for one text used in RSAT. The excerpt contains the first six sentences from a text describing the first Battle of Bull Run from the U.S. Civil War. The question provides an assessment of how well the test takers comprehend the story and are able to draw upon world knowledge. Specifically, the text takers need to access knowledge from the text that the Confederates had made a decision to move their capital further north and access geographical knowledge that this would place their capital closer to Washington, D.C. The example student answers reflect the range of responses, and as can be seen, they vary in the extent to which the student answers contain semantic overlap with the expert answer (once again, the expectations). Scoring these protocols is a matter of assessing semantic overlap with the expert answer (Magliano et al., 2011) or a family of answers that range in quality (Attali & Burstein, 2006; Landauer et al., 2003; Shermis et al., 2010) and the extent to which they contain common misconceptions (VanLehn et al., 2007). Regardless, the assessment task for the computer is relatively straightforward because there is a relatively closed set of appropriate responses to the question.

Table 3 Sample text excerpt, short answer question prompt, and expert answer from RSAT

The Battle of Bull Run
1. The First Battle of Bull Run was the first real battle of the Civil War.
2. Union officials felt it would be an easy victory and would lead to a quick conclusion of the war.
3. History would prove otherwise.
4. At the end, the Confederacy was the winner, routing inexperienced Union forces and sending them fleeing all the way back to Washington.
NEW PARAGRAPH
5. Confederate leaders were eager to prove their mettle against the North and announced intentions to move the capital to Richmond, Virginia.
6. The Union government was angered by this prospect.
QUESTION: Why were they angry?
EXPERT ANSWER: The Confederate South decided to move its capital further north to Richmond Virginia near Washington D.C. closer to the Union capital.
SAMPLE STUDENT ANSWERS
1. Because of the Confederacy.
2. Because they were trying to make a move on the ground. Who wouldn't be angry?
3. Because the Confederates wanted to move the Capital to Richmond Virginia, which is pretty close to Washington.

Longer essays enjoy both advantages and disadvantages, as compared with shorter responses. A major advantage of length is that the reliability is higher by virtue of the fact that there is more information in the student discourse. However, a major liability is that there is a greater range of acceptable responses and more variability in the quality and composition of the essays. There also are more potential aspects of the essays that one could score. For example, one may want to assess the internal coherence of the essays that students write (Burstein et al. 2003; Foltz, Kintsch, & Landuaer, 1998; E. Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007). If readers are basing their essays on a text or set of texts, one may want to evaluate how effectively the essays cover content from the essays and whether the readers are appropriately sourcing where their ideas come from or whether they are plagiarizing the texts (Britt et al., 2004). Another challenge of these essays is that they are typically written in a context where students do not have the opportunity to reflect and revise their essays, so the quality of the student writing is considerably less than optimal (McNamara et al., 2012). It is up to the developer to decide whether the mechanics of

writing under these situations should be taken into consideration when evaluating the essays.

One very important issue is whether the student responses are based on source texts or reflect extemporaneous writing. When source text(s) exist, the sources impose constraints on the semantic content that should be in the essays. The developer can anticipate what kinds of information will be in the essays, which simplifies the development of assessment protocols. On the other hand, extemporaneous writing on a general topic creates a situation in which there will be dramatic variability in what students choose to write about or incorporate into their essays. This limits the assessments to an evaluation of the quality of the writing style or rhetorical format, as opposed to the richness of the semantic aspects of the essays. For example, one may want to detect the presence of a thesis statement, main ideas, supporting ideas, and conclusion in an argumentative essay on a specific subject matter (Burstein et al., 2003). This constrained writing assessment would be very different from the structure and wide-open content of a narrative of life's experiences.

In summary, there are a variety of different uses and types of student-constructed responses that are evaluated and constrained by the goals of the automated system. The systems vary in their accuracy of assessment when compared with human scoring of the open-ended responses. The goals and constraints of the system determine what type of response will be elicited from the user and what natural language processing approaches are needed for optimal analyses of these responses. We argue that the type of response has implications for the approach one uses to develop expectations to analyze those responses.

Computational approaches for analyzing student-constructed response

At the outset of this article, the claim was made that there have been dramatic successes in the automatic assessment of student-constructed responses (Graesser & McNamara, 2012). These successes can be attributed to landmark advances in computational linguistics (Jurafsky & Martin, 2008; Shermis & Burstein, 2003), discourse processes (W. Kintsch, 1998; McNamara & Magliano, 2009), statistical representations of world knowledge (Landauer et al., 2007), corpus analyses (Biber, Conrad, & Reppen, 1998), word dictionaries with psychological attributes (Pennebaker, Booth, & Francis, 2007), and automated analyses of discourse cohesion (Graesser & McNamara, 2011; Graesser, McNamara, Louwerse, & Cai, 2004). These advances enable one to evaluate semantic overlap between words, sentences, paragraphs, and entire texts.

One explanation of the successes has been the advent of statistical representations of world knowledge that are based on large corpora, instead of relying entirely on highly structured representations that attempt to capture precise and accurate meanings. Twenty years ago, the typical models in computational linguistics had highly structured lexicons, syntactic parsers, semantic analyzers, and representations of world knowledge (Allen, 1995; Lehmann, 1985). However, these traditional structured representations were brittle and could not account for the human language contributions that were frequently ungrammatical, vague, and imprecise. The statistical, corpus-based representations of meaning are somewhat simple but, nevertheless, powerful estimates of the extent to which the student responses reflect key constructs and expectations (Britt et al., 2012; Graesser & McNamara, 2012; Landauer et al., 2007). The ideal systems are a hybrid between symbolic and statistical representations of meaning (Cai et al., 2011; Jurafsky & Martin, 2008; Rus, McCarthy, McNamara, & Graesser, 2008).

This section identifies a variety of computational modules that have been used in analyses of student-constructed responses. The appropriateness of a module depends on the semantic complexity of the responses and the relationships to the expectations that comprise the assessment targets. By way of preview, we advocate the use of hybrid algorithms that make use of multiple computational mechanisms whenever the goal is to optimize the accuracy of the assessments. This section also describes general approaches used by many computational systems that tap different semantic and conceptual levels of meaning.

Natural language algorithms

A variety of natural language algorithms have been implemented to make the semantic comparisons between the student verbal responses and the assessment targets. For example, we might designate the student responses in an essay or a conversational turn as a set of n sentences (S_1, S_2, . . . S_n) and the assessment target as a set of m expectations (E_1, E_2, . . . E_m). Computations would be needed to assess the semantic similarity between a student sentence S_i and an expectation expression E_j. The output of the computation would a semantic similarity score that varies between 0 and 1. When the expectation expression is also a sentence, the computation would assess the semantic overlap between two sentences.

Perhaps the simplest algorithm in computing semantic similarity is word matching. This algorithm computes the specific words that overlap between the student response S_i and the assessment target E_j. The similarity scores is computed as [(2 \ast C)/(A + B)], given there are C common words, A words in the student response and B words in the expectation. However, this simple word overlap algorithm is limited by the fact that some words are more important than others (e.g., content > function words), words with the same

semantic stems have different spellings (e.g., *run, runner, running, runs*, etc.), and some words are misspelled. Fortunately, there are tools that compute semantic overlap in a manner that handles these variations (Jurafsky & Martin, 2008; McCarthy & Boonthum-Denecke, 2012). Part-of-speech taggers identify different syntactic classes of words on the basis of a word lexicon and the syntactic context of a word in a sentence. Algorithms such as Soundex (Christian, 1998) are able to identify key words even with misspelling or word transformations. Lemmatizers are available to extract core morphemes from words (e.g., the lemma "run" from *run, ran, runner, runners, running*, etc.). Words or lemmas are sometimes weighted in these overlap measures so that a higher weight goes to content words, rare words, or words directly relevant to the subject matter. Researchers can compute semantic similarity with each of these variants of operationally defining relevant linguistic units and weighting schemes.

There are algorithms that consider sequences of words (*n*-grams) and structured configurations of words or lemmas in regular expression templates (Cai et al., 2011; Jurafsky & Martin, 2008). *N*-grams are simply sequences of particular words, such as *b*-grams (word pairs), trigrams (word triplets), or more generally, sequences of *N* words. Algorithms can compute overlap of *n*-grams in its assessment of similarity. A very powerful technique for assessing semantic similarity is to represent an expectation E_j as a regular expression. Regular expressions represent the meaning of an expectation as a structured sequence of lemma-like units. For example, suppose that expectation E_j is the sentence "The dependent variable needs to be accurate, sensitive, and precise," with key words or phrases being "dependent variable," "accurate," "sensitive," and "precise." Different forms of words are recognized through regular expressions by creating abbreviated expressions such as "sensitiv," which allows for "sensitivity" and "sensitive." The regular expression ("/bdependent variable") for "dependent variable" is constrained, thereby keeping students from getting appropriate credit by saying "independent variable." Synonyms or functionally equivalent meanings of a word are also exemplified in the following type of expression: "b/dependent variable|DV|dv|outcome|criteri." Complex logical structured expressions can similarly be set up with regular expressions. Semantic matches between student sentences S_i and expectations E_j allow greater flexibility in student articulations and yield much higher accuracy than does any word overlap measure (Cai et al., 2011; Jurafsky & Martin, 2008).

High-dimensional semantic spaces, such as latent semantic analysis (LSA; Landauer & Dumais, 1997; Landauer et al., 2007), HAL (Burgess et al., 1998), and holographic models (Jones, Kintsch, & Mewhort, 2006), are also used in many systems to perform semantic matches. LSA is the most widely used statistical technique of this class of approaches to natural language processing, so it is the primary focus in this discussion. LSA is an important method of computing the conceptual similarity between text documents (e.g., words, sentences, paragraphs, or essays) because it considers implicit knowledge in addition to the explicit words. 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 word W in naturalistic documents; two words are similar in meaning to the extent that they share similar surrounding words. For example, the word *car* is highly associated with words of the same functional context, such as *engine, race, wheels, parking*, and *transportation*. These words are not synonyms or antonyms that would occur in a dictionary or thesaurus, but more like the co-occurrence of words in an encyclopedia article. LSA uses a statistical technique called *singular value decomposition* to condense a very large corpus of texts to 100–500 statistical dimensions (Landauer et al., 2007). The conceptual similarity between any two text excerpts (e.g., word, clause, sentence, entire essay) is computed as the geometric cosine between the values and weighted dimensions of the two text excerpts. The value of the cosine typically varies from approximately 0 to 1. For example, the cosines between *doctor* and *physician*, *nurse*, *office*, and *dog* are .61 .52, .29, and .03, respectively.

LSA is useful in the analysis of student-constructed responses because it is sensitive to inferences and to both proximal and distal semantic relationships. Consider situations in which the constructed responses are compared with the expectations in a source document. Students may use words in that document, words associated with inferences, synonyms, antonyms, and completely unrelated words. LSA is sensitive to the semantic distance of the words used by a student and the words in the source document. LSA can also evaluate how sentences and lengthier student responses can compare with the source document in meaning.

Researchers have performed systematic analyses to examine which of the natural language algorithms provide the best semantic matches to expectations (Cai et al., 2011; Graesser, Penumatsa, Ventura, Cai, & Hu, 2007; McNamara, Boonthum, Levinstein, & Millis, 2007; Rus et al., 2008). These analyses have confirmed that the best semantic similarity modules use hybrid models that take advantage of weighted keyword overlap, regular expressions, and high-dimensional semantic spaces. However, there are a few rules of thumb when deciding which system to explore and whether to consider developing

hybrid systems. When the responses are based on specific word- or sentence-length expectations and are likely to be constrained by specific source texts, keyword matching and regular expressions are often sufficient. This is because there is a narrow band of variation in the answers that semantically map onto the different assessment targets. Consequently, the higher dimensional spaces typically yield a modest improvement in the accuracy of the system. However, when responses are ill formed, are not strongly based on a source text, or are long (several sentences or even several paragraphs), using high-dimensional semantic spaces become critical for dealing with the variability in student responses.

Two approaches for developing assessment targets

All of the automated systems involve semantic comparisons between the student response and a set of assessment targets. There are two general approaches for developing the assessment targets, each requiring the development of expectations that reflect assessment targets. The *construct-based* approach involves developing expectations that reflect theoretical constructs, such as correct answers to a question or particular classes of inferences that a reader might generate. The *normative-based* approach involves developing expectations on the basis of a set of responses that are representative of different levels of student outcomes, such as essays that are reflective of receiving a grade of A, B, C, D, or F or responses that are representative of different types of misconceptions.

The construct-based approach requires the development of *models* of the students' cognitions, tasks, and a range of student products that should underlie the key constructs delineated by theory (Britt et al., 2012; Mislevy, 2007; Pellegrino & Chudowsky, 2003). That is, the assessment components of a computational system would ideally follow the tenets of an evidence-centered design approach to assessment development. This approach incorporates (1) a student model that identifies relevant cognitive states and processes, (2) a task model that specifies the task requirements and how to the map onto the constructs specified by the student cognitive model, and (3) guidelines for how to interpret a student's task performance with respect to the key constructs. In the first section, we mentioned that the type of response has implications for the protocols that one develops to analyze those responses. As such, we strongly advise that this approach be adopted whenever one is analyzing meaning-making responses, because it is typically the case that the developer wants to assess specific types of "events" that are reflected in the responses. Theory determines the importance of these events when developing coding protocols developed by human judges (e.g., Trabasso & Magliano, 1996), and the same should be true for automated assessments.

As an example, an evidence-centered approach was used for the development of RSAT for the analysis of some of the types of responses produced by students (Magliano et al., 2011). RSAT is a computer-administered test that is designed to assess a student's level of comprehension and the processes that support it while the student is reading (Gilliam et al., 2007; Magliano et al., 2011). Students read texts one sentence at a time and are prompted to answer *indirect questions* ("what are you thinking now") that require responses that are akin to thinking aloud (Trabassso & Magliano, 1996).

The evidence-based approach guided the development of semantic benchmarks that were associated with four types of processing theoretically important for comprehension (see McNamara & Magliano, 2009, for an extensive review): paraphrasing the current sentences, local bridging inferences, distal bridging inferences, and elaborative inferences. As is illustrated in Table 1, many of these processes can be induced by the informational content of the protocols. One version of RSAT relies solely on keyword matching for content words (Magliano et al., 2011) to theoretically derived expectations reflecting the processes described above (although we are currently exploring the development of hybrid systems that combine keyword spotting and LSA). The expectation for detecting paraphrasing consists of content words (nouns, verbs adjectives, adverbs) in the current sentences, whereas the expectation associated with bridging inferences consists of content words that are in the local sentence or distal sentences. RSAT detects elaboration by counting content words in the protocols that are not explicitly in the current sentence or prior discourse context. This is not a perfect measure, because synonyms for content words in the discourse context are counted as mere elaboration. Clearly, student elaborations need to be scored on a continuum from being relevant to the content to moving away from the discourse content (Magliano et al., 2011). Despite the simplicity of this approach, RSAT processing scores are highly correlated with human judgments of the verbal protocols (*r*s ranging from .48 to .78; Magliano et al., 2011). Elaboration scores are the lowest correlations with human judgments, perhaps in part because of the synonym problem, but another reason is the need to differentiate relevant versus irrelevant elaborations. Elaborations generally present a challenge to automatic detection, so future research is needed to understand these open-ended constructions (McNamara et al., 2007a; Millis et al. 2007; Rus et al., 2008).

ISTART (Interactive Strategy Trainer for Automated Reading and Thinking; McNamara, O'Reilly, Rowe, Boonthum, & Levinstein, 2007) is another example of the evidence-centered approach. ISTART is a reading strategy

training system that is intended to promote the comprehension of challenging texts by teaching students to self-explain during reading. For the purpose of this article, we are particularly interested in how iSTART analyzes and uses the practice self-explanations produced by the student user. Importantly, it implements a hybrid system involving word matching and LSA. ISTART algorithms are designed to provide a global assessment of the self-explanation based on a theoretical analysis of knowledge building when self-explaining (Coté et al., 1998). The algorithms attempt to provide an assessment of the protocols that range from vague and uninformative responses to responses that reflect building a rich mental model for a text. Word counts on the length and LSA cosine match scores between the student response and the text (i.e., the current sentence) provide the basis for determining the extent to which the students produce a sufficiently rich and relevant response, as opposed to simply repeating the current sentence or producing a vague and uninformative response (e.g., "Ok," "I get this," "sounds interesting"). If, for example, the word counts fall below a threshold, the student is prompted to say more. If the LSA cosine match scores are too similar to the current sentence, the student is prompted to produce self-explanations that go beyond the sentence that they just read. Self-explanations that are deemed rich enough are subjected to an assessment algorithm that takes advantage of word counts and LSA cosines between the self-explanation and assessment targets, including the current sentence, prior discourse contexts (local and distal sentences), and text title. The agreement between the algorithms and human judgments of the different levels of processing has been impressive, with 62 %–64 % agreement between the two (McNamara et al., 2007b). Moreover, algorithms generated from one text significantly generalize to other texts (Jackson, Guss, & McNamara, 2010).

An alternative construct-based approach is to use assessment targets with very specific semantic content that varies in quality and correctness. For example, the AutoTutor system developed by Graesser and colleagues has adopted such an approach (Graesser et al., 2008; Graesser, Lu, et al., 2004; VanLehn et al., 2007), as does also an AutoTutor derivative called Operation ARIES (Cai et al., 2011; Millis et al., in press). Students are asked to answer challenging questions (posed by an animated tutor agent) in a multi-turn conversation between the student and computer tutor. The students' responses are semantically compared with expectations that cover the set of expected sentence-length expressions in a good answer and also common misconceptions derived from a normative sample. The computational algorithms in the semantic matches include frequency-weighted content-word overlap (i.e., less frequent words get higher weight), LSAs (Landauer et al., 2007), and regular expressions (Jurafsky & Martin, 2008). The moves that the AutoTutor and Operation ARIES systems make when providing feedback are based on these semantic assessments. Again, consistent with the claim that an evidence-based approach is optimal, these expectations are developed on the basis of theories of tutoring and a cognitive model of student learning of the subject matter.

As was discussed earlier, there is an alternative to the construct-based approach—namely, a normative-based approach. In normative-based approaches, assessment targets are based on exemplar responses that vary in quality (Attali & Burstein, 2006; Foltz, Gilliam, & Kendall, 2000; Landauer et al., 2003). For example, there might be five grade levels for essays, and each grade would have a large set of example essays with the particular grade. When a new essay is to be graded, it is matched to the exemplars in the various grades; the grade received reflects higher matches to the exemplars in the grade category than to exemplars in other grade categories. This approach does not have as strong an alliance with an evidence-centered design unless there is a theoretically based method of classifying exemplar essays.

When is it appropriate to use a construct-based versus a normative-based approach? We argue that one should always rely upon a construct-based approach when analyzing meaning-making responses. This recommendation stems from the fact that most of these approaches are based on a rich empirical and theoretical history, which is certainly the case for systems that are designed to promote reading comprehension. Theories and research can and should provide a basis for detecting processes that support comprehension and deep learning. Normative approaches are appropriate when analyzing responses that reflect products of comprehension and understanding (e.g., Foltz et al., 2000), particularly when ideal answers and constructs are not available and the goal is to give global feedback. In this case, it is useful to identify a family of exemplar responses that reflect qualitative differences in student responses. That said, these approaches are by no means mutually exclusive and can be drawn upon to support different kinds of assessment. When one wants to evaluate and provide feedback regarding specific types of processing or products in a response, the construct-based approach should be used, whereas the normative-based approach can be useful for gauging overall quality that could support computer-aided grading.

In summary, as we have discussed, there are a variety of natural language algorithms that compute the extent to which the student response matches expectations in the theoretical rubric. These include word overlap, lemma overlap, regular expressions, and higher dimensional semantic spaces. Whenever possible, it is advisable to develop hybrid systems that take advantage of the strengths of the different approaches, and semantic spaces should always be considered with complex, variable, and longer responses. Regardless of the natural language algorithms one uses, one will

usually have to develop expectations that are compared with the protocols and that provide the basis for assessments of the quality and nature of the protocols. We have discussed two general approaches, and their appropriateness is, in part, contingent on the type of student response.

Conclusions

This review can be used to generate a few simple yet important recommendations for the development of future systems. First, it is important to explicitly identify whether the responses reflect meaning making or products of comprehension. If the latter, it is always advisable to identify expectations via a construct-based approach. Although a normative approach is very useful for providing global assessments of the quality of response that reflect products of comprehension, a construct-based approach should be used when there is a theoretical basis for developing an assessment protocol. Although simple word- and pattern-matching algorithms may be sufficient to analyze responses that are short and are likely to have a relatively narrow semantic range, it is advisable to utilize high-dimensional semantic spaces for longer responses. Moreover, we advise the use of hybrid systems that make use of both pattern-matching and high-dimensional semantic spaces.

As is evident in this review, computer-based assessments of student-constructed responses handle the same variety of responses that are used in research and educational contexts. Although sometimes imperfect, these systems are remarkably successful in categorizing critical dimensions of the student responses. The successes of many of the systems discussed in this article can be attributed to the adoption of a statistical approach for estimating the content of the student responses, in addition to some modicum of structured representations. It is remarkable how well these systems have performed when classifying and scoring open-ended responses (Graesser & McNamara, 2012), despite the fact that they do not attempt to engage in a precise symbolic computation of meaning. This is not to say that some aspects of student responses may be outside of current technological advances to detect (Calfee, 2000; Ericsson & Haswell, 2006). To be sure, there are imperfections in the systems from the standpoint of use and accuracy, but the case can be made that the technologies are ready for applications to diverse learning environments. However, we would hesitate to use them for high-stakes assessments in the league of SAT or ACT.

What are the most important future directions for this field? While we have advocated the use of hybrid systems, this review did not provide specific recommendations regarding how to develop improved systems. There are relatively few systems that use hybrid approaches to date (Britt et al., 2004; Cai et al., 2011; Graesser & McNamara 2012; McNamara et al., 2004), and each focuses on different types of protocols. We do know that the shorter responses require well- engineered word spotting in the expectations, with large gains from structured regular expressions and *n*-grams. We also know that LSA and other high-dimensional spaces are important for capturing the inferential meaning in longer essays. However, more research is needed to explore a broader landscape of hybrid systems. And once again, the solutions will depend on the type of response and assessment context.

A second line of important research should be directed at making these systems more accessible to practitioners and teachers. Many of these systems require extensive engineering to develop and, in particular, in the development of the expectations and algorithms used to analyze the protocols. Teachers and school districts will want to choose materials that map onto mandated curricula or reflect topics covered on high-stakes tests. Some systems, such as iSTART and Summary Street (Franzke et al., 2005) have been developed with this constraint in mind, and the algorithms for analyzing student responses were created such that they did not need extensive engineering to implement with new texts. This can be achieved by automatically identifying expectations based on features of the text (e.g., the current sentence being read is a semantic benchmark for detecting paraphrasing of the current sentence, whereas prior sentences are a benchmark for detecting bridging inferences), but this still requires developing algorithms that can generalize to new texts (Jackson et al., 2010). Additionally, although all the systems discussed in this article involve developing expectations that provide the assessment targets, this may not always be necessary. For example, Coh-Metrix (Graesser & McNamara, 2011; Graesser, McNamara, & Kulikowich, 2011; McNamara, Louwerse, McCarthy, & Graesser, 2010) automatically analyzes texts on a wide range of linguistic and discourse features that correspond to genre, referential cohesion, situation model cohesion, syntax, and word concreteness. Although there are challenges to developing systems that can be flexibly integrated into existing curricula, it is our hope that putting the control in the hands of the teacher will lead to a wider use of student-constructed responses in educational contexts.

Author Note The research was supported by the National Science Foundation (ALT-0834847, DRK12-0918409), the Institute of Education Sciences (R305B070349, R305F100007), the Center for the Interdisciplinary Studies of Language and Literacy, and the Institute for Intelligent Systems. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding sources. Inquiries can be directed to Joe Magliano, Department of Psychology, Northern Illinois University, DeKalb, IL 60115, jmagliano@gmail.com

Appendix

Table 4 Text excerpt from "How Cancer Develops" (the boldface sentence is where the example protocols in Table 1 were elicited)

How Cancer Develops
Cancer begins in genes, bits of biochemical instructions composed of individual segments of the long, coiled molecule deoxyribonucleic acid (DNA).
Genes contain the instructions to make proteins, molecular laborers that serve as building blocks of cells, control chemical reactions, or transport materials to and from cells.
In a cancerous cell, permanent gene alterations, or mutations, cause the cell to malfunction.
For a cell to become cancerous, usually three to seven different mutations must occur in a single cell.
Cancer may take many years to accumulate.
Understanding how cells communicate with one another is an important part of the story.
While each human cell performs its own specialized function, it also exerts influence on the cells around it.
Cells communicate with one another via receptors, protein molecules on the cell surface.
A cell may instruct other cells in its neighborhood to divide, for example, by releasing a growth-promoting signal, or growth factor.
The growth factor binds to receptors on adjacent cells.
A message within each receptor cell becomes activated.

References

Allen, J. (1995). *Natural language understanding* (2nd ed.). Amsterdam: Benjamin Cummngs.

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

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bråten, I., & Strømsø, H. I. (2003). A longitudinal think-aloud study of spontaneous strategic processing during the reading of multiple expository texts. *Reading and Writing, 16,* 195–218.

Britt, M. A., Wiemer, K., Millis, K. K., Magliano, J. P., Wallace, P., & Hastings, P. (2012). Understanding and reasoning with text. In P. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 133–154). Hershey, PA: IGI Global.

Britt, M. A., Wiemer-Hastings, P., Larson, A. A., & Perfetti, C. A. (2004). Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education, 14,* 359–374.

Burgess, C., Livesay, K., & Lund, K. (1998). Explorations in context space: Words, sentences, and discourse. *Discourse Processes, 25,* 211–257.

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

Cai, Z., Graesser, A. C., Forsyth, C., Burkett, C., Millis, K. K., Wallace, P., . . . Butler, H. (2011). Trialog in ARIES: User input assessment in an intelligent tutoring system. In W. Chen & S. Li (Eds.), *Proceedings of the 3rd IEEE International Conference on Intelligent Computing and Intelligent Systems* (pp. 429–433). Guangzhou: IEEE Press.

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

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, R., & Glaser, R. (1989). Self-explanation: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145–182.

Chi, M. T. H., de Leeuw, N., Chiu, M., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439–477.

Chi, M. T. H., Siler, S. A., & Jeong, H. (2004). Can tutors monitor students' understanding accurately? *Cognition and Instruction, 22*, 363–387.

Christian, P. (1998). Soundex—can it be improved? *Computers in Genealogy, 6,* 215–221.

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes, 25,* 1–53.

D'Mello, S., Dowell, N., & Graesser, A. C. (2011). Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language? *Journal of Experimental Psychology. Applied, 17,* 1–17.

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). Hillsdale, NJ: Erlbaum.

Ericsson, P. F., & Haswell, R. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments, 8,* 111–127.

Foltz, P. W., Kintsch, W., & Landuaer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25,* 285–307.

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,* 53–80.

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, Instruments, & Computers, 39,* 199–204.

Graesser, A. C., D'Mello, S. K., & Cade, W. (in press). Instruction based on tutoring. In R. E. Mayer & P. A. Alexander (Eds.), *Handbook of research on learning and instruction* (pp. 408–426). New York: Routledge.

Graesser, A. C., D'Mello, S., & Person, N. K. (2009). Metaknowledge in tutoring. In D. Hacker, J. Donlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education* (pp. 361–382). New York: Taylor & Francis.

Graesser, A. C., Jeon, M., & Dufty, D. (2008). Agent technologies designed to facilitate interactive knowledge construction. *Discourse Processes, 45,* 298–322.

Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journal, 15,* 199–213.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, & Computers, 36,* 180–193.

Graesser, A. C., & McNamara, D. S. (2010). Self-regulated learning in learning environments with pedagogical agents that interact in natural language. *Educational Psychologist, 45,* 234–244.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Sciences, 3,* 371–398.

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper & A. T. Panter (Eds.), *APA handbook of research methods in psychology*. Washington, DC: American Psychological Association.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40,* 223–234.

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, & Computers, 36,* 193–202.

Graesser, A. C., Ozuru, Y., & Sullins, J. (2009). What is a good question? In M. G. McKeown & L. Kucan (Eds.), *Threads of coherence in research on the development of reading ability* (pp. 112–141). New York: Guilford.

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). Mahwah, NJ: Erlbaum.

Graesser, A. C., Person, N. K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9,* 495–522.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101,* 371–395.

Jackson, G. T., Guss, R. H., & McNamara, D. S. (2010). Assessing cognitively complex strategy use in an untrained domain. *Topics in Cognitive Science, 2,* 127–137.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55,* 534–552.

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

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary Street ®: Computer-guided summary writing. In T. K. Landauer, D. M. McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis* (pp. 263–277). Mahwah, NJ: Erlbaum.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240.

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

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

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37,* 389–405.

Lehmann, W. P. (Ed.). (1985) Natural language processing. *Special issue of Computers and the Humanities*, *19*.

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,* 145–170.

Magliano, J. P. (1999). Revealing inference processes during text comprehension. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 55–75). Mahwah, NJ: ErIbaum.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure. *Cognition and Instruction, 21,* 251–283.

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.

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate assessment tools. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107–136). Mahwah, NJ: Erlbaum.

McCarthy, P., & Boonthum-Denecke, C. (Eds.). (2012). *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

McNamara, D. S., Boonthum, C., Levinstein, I., & Millis, K. K. (2007a). Evaluating self-explanations with word-based and LSA algorithms. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), *The handbook of latent semantic analysis* (pp. 227–241). Mahwah, NJ: Erlbaum.

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

McNamara, D. S., Levinstein, I. B., & Boonthum, C. (2004). iSTART: Interactive strategy training for active reading and thinking. *Behavioral Research Methods, Instruments, & Computers, 36,* 222–233.

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

McNamara, D. S., & Magliano, J. P. (2009). Towards a comprehensive model of comprehension. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 51, pp. 297–384). New York: Elsevier Science.

McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007b). 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). Mahwah, NJ: Erlbaum.

McNamara, D. S., Raine, R., Roscoe, R., Crossley, S., Jackson, G. T., Dai, J., . . . Graesser, A. C. (2012). The Writing-Pal: Natural language algorithms to support intelligent tutoring on writing strategies. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 298–311). Hershey, PA: IGI Global.

Millis, K. K., Forsyth, C., Butler, H., Wallace, P., Graesser, A., & Halpern, D. (in press) Operation ARIES! A serious game for teaching scientific inquiry. In M. Ma, A. Oikonomou, & J.

Lakhmi (Eds.), *Serious games and edutainment applications*. London: Springer.

Millis, K. K., Magliano, J. P., Todaro, S., & McNamara, D. S. (2007). Assessing and improving comprehension with latent semantic analysis. In T. Landauer, D. S., McNamara, S. Dennis, & W. Kintsch (Eds.), *The Handbook of Latent Semantic Analysis* (pp. 207-225). Mahwah, NJ: Erlbaum

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36,* 463–469.

Muñoz, B., Magliano, J. P., Sheridan, R., & McNamara, D. S. (2006). Typing versus thinking aloud when reading: Implications for computer-based assessment and training tools. *Behavior Research Methods, Instruments, & Computers, 38,* 211–217.

Pellegrino, J. W., & Chudowsky, N. (2003). The foundations of assessment. *Interdisciplinary Research and Perspectives, 1,* 103–148.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: LIWC 2007 [Computer program]*. Austin, TX: LIWC.net. Retrieved from www.liwc.net.

Person, N. K., Graesser, A. C., & the Tutoring Research Group (TRG). (2002). Human or computer? AutoTutor in a bystander Turing test. In S. A. Cerri, G. Gouarderes, & F. Paraguaçu (Eds.), *Intelligent tutoring systems 2002* (pp. 821–830). Berlin: Springer.

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

Rus, V., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2008). A study of textual entailment. *International Journal on Artificial Intelligence Tools, 17,* 659–685.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78,* 153–189.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross disciplinary perspective*. Mahwah, NJ: 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., pp. 20–26). Oxford: Elsevier.

Trabasso, T., & Magliano, J. P. (1996). Conscious understanding during text comprehension. *Discourse Processes, 21,* 255–288.

VanderVeen, A., Huff, K., Gierl, M., McNamara, D. S., Louwerse, M., & Graesser, A. C. (2007). Developing and validating instructionally relevant reading competency profiles measured by the critical reading sections of the SAT. In D. S. McNamara (Ed.), *Theories of text comprehension: The importance of reading strategies to theoretical foundations of reading comprehension* (pp. 137–172). Mahwah, NJ: Erlbaum.

VanLehn, K., Graesser, A. C., Jackson, G. T., Jordan, P., Olney, A., & Rose, C. P. (2007). When are tutorial dialogues more effective than reading? *Cognitive Science, 31,* 3–62.