Using Scenario-based Assessments to Measure Deep Learning

Educational Testing Service

Tenaha O'Reilly John Sabatini, Zuowei Wang

This manuscript is an early draft of a paper published in Deep Comprehension and thus may differ slightly from the final published version. Please see below for the official paper:

O'Reilly, T., Sabatini, J., & Wang, Z., (2018). Using scenario-based assessments to measure deep learning. In K. Millis, D. Long, J. Magliano, & K. Weimer (Eds.), Deep Comprehension: Multi-disciplinary Approaches to Understanding, Enhancing, and Measuring Comprehension, 1st Edition (pp. 197–208). NY, NY: Routledge.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305F100005, R305A150176, R305A160129 to the Educational Testing Service. The opinions expressed are those of the authors and do not represent views of Educational Testing Service, the Institute, or the U.S. Department of Education.

Context and background

Information is at the heart of lifelong learning, effective citizenship and good decision making. Yet information can come from a variety of sources with different degrees of trustworthiness, accuracy and depth (Metzger, 2007). The information age has resulted in an explosion of knowledge that can be readily accessed by anyone who uses the Internet. However, not only has the volume and access of information increased, but also the ability of any one individual to publish anything they want at any time. While these freedoms have opened up the playing field, allowing multiple voices to be heard, such a system without any checks and balances can be misused. Although information should be used for the common good, misinformation (Ecker, Lewandowsky, Chang, & Pillai, 2014), propaganda, monetary gain, and the lack of evidenced-based reasoning may serve as threats to our citizens and democracy.

Given these and other issues, many researchers and educators have advocated for an updated construct of reading, one that recognizes the opportunities and challenges faced by increased technology and communication (Goldman, et al., 2016; Leu et al., 2013; Magliano, McCrudden, Rouet, & Sabatini, in press; Sabatini, O'Reilly & Albro, 2012). While it is beyond the scope of the this chapter to describe the full range of construct changes that educators and researchers have advocated in the domain of reading, in the section below, we touch on some of the more common themes that are relevant to the current chapter on deep understanding. In short, we conclude that the time is ripe to revisit the notion of deep understanding, while simultaneously leveraging the research in the learning sciences to improve assessment.

What is deep understanding and learning?

The answer to this question is complex, and will vary depending upon whom you ask. Deep understanding is a phrase that many people discuss and strive towards, but few people agree on its definition. To help provide some clarity to the issue, in this section we outline seven claims and principles that characterize various elements of depth at different degrees of specificity. While this list is not exhaustive, our aim is to drive a discussion around a set of claims and features that can be measured during an assessment. Although one may notice the degree of depth seems to increase as the number of the claim increases, this observation is somewhat premature. For instance, the notion of depth and difficulty are intertwined with the level of students' background knowledge, text complexity, and task types (McNamara, Kintsch, Songer, & Kintsch, 1996; O'Reilly & McNamara, 2007). That is, the degree of depth for any particular claim may vary depending upon how much a particular student knows, how complex the text is (in terms of both content and the language), or what tasks are used to measure understanding (e.g., verification vs. essay). Thus, the order of claims do not necessarily entail a strict hierarchy of depth.

Claim 1: *Deep understanding is, in part, defined by the purpose and goals for reading.* Before a discussion of depth can be considered, it is important to define the reader's goals or purpose for reading. While a person can read for general understanding, research has shown that people process and comprehend text differently depending upon their goals of reading (van den Broek, Lorch, Linderholm, & Gustafson, 2001). The goals of reading help students to adopt a standard of coherence (Oudega & van den Broek, this volume; van den Broek, Young, Tzeng, & Linderholm, 1999) that helps define what level of attention and understanding is required by the task at hand. In other words, depth is relative to the goals that define the task. For instance, if a reader's goal is to find the date for a

particular event, a search and scanning strategy to locate that information is warranted. In this case, the person does not necessarily need to even read the text, but rather scan until they find what they are looking for. However, if the goal of reading involves making a decision, a number of additional processes may come into play (e.g., identify the options, pros and cons of various approaches, cost/benefit analysis, etc.). In this case, the person may not only read an entire source document on the topic, but also read multiple documents on the topic that may involve corroborating conflicting recommendations (Goldman et al, 2016).

Absent any meaningful goal for reading, students may read just enough to form a basic understanding, and as a result, they may not draw many inferences (McKoon & Ratcliff, 1992). However, according to Constructionist theory (e.g., Graesser, Singer, & Trabasso, 1994), deep understanding is achieved when people actively search for meaning. Graesser et al. (1994) outline three assumptions that enable deep processing. The coherence assumption suggests that readers form coherent models of reading at both the local and global levels. This implies that readers will reconcile any contradictions in their understanding of the text. The explanation assumption assumes that readers will attempt to develop explanations that describe how events or processes occur. These explanations usually focus on causes, and deep understanding involves making sure the antecedents leading to causes are sound. Related to claim 1, the Constructionist theory also suggests that people read with a particular goal in mind, and that the goal for reading will help dictate the level of understanding a person will achieve when reading a text (reader goal assumption). For instance, a shallow goal will result in shallow model of the text (i.e., low standard of coherence). In short, any discussion of depth should be contextualized by the reader and task goals (Graesser et al., 1994). Providing more complex goals should lead to deeper processing.

Claim 2: *Deep understanding involves going beyond the literal interpretation of text, by forcing readers to infer unstated ideas*. Superficial understanding is often discussed in the context of literal meaning of the text (Minguela, Solé, & Pieschl, 2015). However, authors write for a particular audience and they assume that readers have at least a minimal amount of knowledge of the topic in question. Thus, authors leave out certain details in their writing and readers need to infer the relations among ideas that are not directly stated in text (Beck, McKeown, & Gromoll, 1989). As such, a literal understanding of the text is often not sufficient to extract even the basic meaning intended by the author; inferences are required to make the text coherent. Given this discussion, it is not surprising that many researchers have made the claim that the ability to draw inferences is one of the hallmarks of reading ability and it is a key metric for distinguishing skilled from less skilled readers (Hannon & Daneman, 1998; McNamara, de Vega, & O'Reilly, 2007). Inferences allow students to connect proximal or distal ideas in text (bridging inferences), or in other cases, allow students to predict events or consequences that are not described in text (elaborative inferences). The ability to draw inferences is critical to the Constructionist theory (Graesser et al., 1994) as people seek to achieve local and global coherence for texts that are often not explicit. In short, we advocate *at minimum*, depth of understanding should include students' ability to draw inferences that connect ideas in text (Minguela et al., 2015).

Claim 3: *Deep understanding involves knowing not just the key facts and main ideas, but how the key ideas are related to each other.* A collection of facts, while complex at some level, is not a defining feature of depth. Depth of understanding is, in part, a comprehensive knowledge of the *set of relations* among global ideas in text, including causal and temporal relations (Graesser et al., 1994). These may

include explanations of why or how an event or process occurs (McNamara, 2004). Some of the relational information might be explicitly stated in the text, while other connections have to be inferred by the reader (claim 2). Several tasks that may be appropriate for measuring global ideas and their relations including summary writing, concept maps, and graphic organizer tasks. Indeed, interventions that focus on these features and text structure are effective for improving reading comprehension (Franzke, et al., 2005; Robinson & Kiewra, 1995; Williams, 2007). Thus, deep comprehension is more about the relational and causal connections among key ideas (Pascual & Goikoetxea, 2014) than a set of isolated facts. To think deeply, students need to develop a coherent model of text that represents the organization of essential concepts and their associations.

Claim 4: *Deep understanding involves the ability to ask meaningful questions that clarify the student's understanding, or challenge and extend the author's stated claims.* Reading comprehension is not a linear process, but an iterative process that involves revisiting the text to clarify meaning, and to update one's understanding as new information becomes available (van den Broek et al., 1999). As people read text, they may have questions that guide their initial reading, but they may also develop new questions to direct behavior when comprehension problems occur, or when new information peaks their curiosity. While the ability to successfully *answer* comprehension questions is one feature that is used to gauge reading ability, the ability to *generate*, *identify or ask relevant questions* of one's own understanding is also important (Graesser & Person, 1994; King, 1995; Pashler et al., 2007). In order to generate or identify a relevant question, students need to have some understanding of the text. More importantly, question generation can guide the process of metacognition (García et al., 2014) by helping students identify what they don't know. In this way, questions can serve as goals to self-regulate behavior (García et al., 2014) in the form of correcting student misunderstandings, or to guide further information gathering through inquiry learning.

However, not all questions are equal. In a review by Graesser and Person (1994), the authors found that most of the questions that occur in the classroom originate from the teacher and many of these are shallow. The authors' review also uncovered that when students did ask questions, they occur infrequently (median of three per hour of instruction) and many of these were also shallow. In prior work, Graesser, Person, and Huber (1992) created a classification scheme to characterize questions in terms of depth. The classification system includes 18 categories that vary in terms of depth. For instance, lower level questions may ask students to verify information (e.g., yes, no) or provide a concept completion (e.g., who, what?). In contrast, deeper questions are those that may ask students to interpret or explain enabling factors (e.g., why, how?). In short, the ability to ask and identify relevant questions is a form of deep comprehension as these actions may help students clarify their understanding or serve as goals for further inquiry (i.e., metacognition and self-regulation). However, the quality of the question matters; Graesser and Person (1994) found that the quality of the question, but not the frequency of questions is related to achievement. Thus, the identification and production of high quality questions is a form of deep understanding.

Claim 5: *Deep understanding involves the ability to evaluate the authors and their respective claims and evidence.* As mentioned earlier, the construct of reading has evolved with the ever increasing developments in technology and access to information (Leu et al., 2013), resulting in an explosion of source content. The complexity of evaluating sources has been compounded by the fact that anyone can publish information, a lot of information is not vetted, and thus, it may not be accurate or trustworthy (Goldman, 2012). This situation necessitates that readers evaluate the credibility of the

sources, authors, and claims (Britt, Rouet, & Durik, 2017; Graesser et al., 2007; Metzger, 2007). As such, in today's world, it is not enough to just be able to understand *what* an author is trying to say, but rather students must also evaluate the author's credibility and the truth and soundness of their arguments. This added layer of processing may place extra demands on the reader and thus require deeper processing. This claim is supported by research that suggests students are not good at evaluating texts, authors, or evidence (Britt et al., 2017; Foy, LoCasto, Briner, & Dyar, 2017), and it is difficult to train students to become more critical as they engage in inquiry learning tasks (Graesser et al., 2007). In sum, deep understanding requires a critical evaluation of the source, author, claims, and evidence.

Claim 6: *While deep understanding has its origins in individual experience, knowledge is a socially constructed phenomena and should be vetted by disciplinary communities of practice.* In the past, reading was primarily an individual activity, and in many cases, students could read and trust a single source to provide quality information that met their needs. However, increased communication, coupled with changes in the workforce, has reshaped how people interact with one another. More and more, people are communicating and interacting with each other in virtual groups with a variety of media including, email, chats, blogs, forums and virtual meeting spaces (e.g., Face Time). In short, today's youth are now both consumers and producers of information and culture through the power of social media (Luschen & Bogad, 2010). In other words, the world is becoming more "social" as people interact with others across time and place. This calls for a new set of skills focused on shared understanding of content, intention, motive, emotions, and maybe even cross cultural competence (Trejo, Richard, van Driel, & McDonald, 2015). Thus, while we are all individuals and have our own understandings of the world, today's digital and work environment requires that we interact more with others as we develop and reconcile shared understandings about the world. This necessarily requires people to engage in perspective taking (LaRusso et al., 2016), a process that involves understanding people's motives, intent, viewpoints, and reasoning about their beliefs and actions.

While individual opinion and an understanding of perspective matters, there is a larger social community of organized practice that is also important in academic domains that may impact deep learning. The information created in academic areas is cumulative and errors about prior claims, are corrected over time by communities of scholars. More specifically, each discipline has its own shared values, expectations, epistemologies, and ways of communicating with members of their community. This collective understanding and shared social practice is often called disciplinary literacy (Goldman, 2012, Goldman et al. 2016). That is, what is valued and considered as valid evidence in one domain may differ greatly from how information is processed and evaluted in another domain. For example, in history, considerations of primary versus secondary sources may impact the interpretation of past events, while in science, evidence for specific hypotheses is important for evaluating how well theories explain phenomena.

In sum, changes in the way that people interact and communicate in today's digital society have increased the demands for developing shared understandings of both content and human psychological states. These issues are often contextualized by larger communities of practice that help shape the way information is processed, valued, evaluated, and communicated. Deep understanding in the 21st century requires that individuals are not only required to understand a message in isolation, but also to engage in perspective taking, as meaning is co-constructed in local social settings or larger communities of practice. These changes have also required people to consult multiple sources as they corroborate information that may contain conflicting claims (Britt et al., 2017). Thus, deep understanding involves

being able to navigate multiple sources of information, perspectives, and social and disciplinary practices.

Claim 7: *At higher levels, deep understanding involves applying what one has read to solve a problem, make a decision, or transfer concepts to a new situation.* The above discussion outlines some of the possible ways to characterize depth from "relatively simple" processes such as making inferences, to more complex processes that require critical thinking and the navigation of social perspective. Collectively, these forms of depth focus on understanding a text, but they do not go too far beyond it. As such, our prior work has advocated for another element of depth, called applied comprehension (Deane et al., 2015; O'Reilly & Sheehan, 2009, Sabatini et al., 2013). By applied comprehension, we mean the set of processes that enable students to use the information contained in text for some particular purpose that *may go beyond the author's original purpose for writing*. More specifically, we view applied comprehension as instrumental; concepts, principles, or ideas read in text are used as instruments to achieve a larger, more complex goal. By complex goal, we mean the reader is trying to solve a problem (e.g., How can we improve the energy efficiency in our school?), make a decision (e.g., Should we adopt green school program?) or transfer what they know to a new situation (Will our actions for reducing energy consumption be successful with a school in a different location?). Collectively, we view applied comprehension as one of the key sources of evidence that a student understands a text at a very deep level.

Scenario-based Assessment

Based on proposed changes to the construct, the authors have designed a new type of assessment called scenario-based assessment (SBA) that might be useful for measuring deep understanding (O'Reilly & Sabatini, 2013; O'Reilly & Sheehan, 2009; Sabatini et al., 2013). The idea of SBA came out of an ETS initiative called Cognitively Based Assessment of, for, and as Learning (CBAL) that was designed to integrate learning and assessment (Bennett & Gitomer, 2009). The current rendition was motivated by a review of the literature in the learning sciences (O'Reilly & Sabatini, 2013; Sabatini et al., 2013; Sabatini & O'Reilly, 2013). The review resulted in five elements or dimensions of the construct that are designed to capture reading for understanding across development. The Print level includes all the typographical information and associated skills that are necessary to "get the words off the page" (e.g., decoding, word recognition, interpreting punctuation or *typographical emphases*). The Verbal level is designed to measure students' knowledge and understanding of vocabulary, morphology, sentences, and syntax. The Discourse level is designed to measure global understanding and includes the relations among key ideas and concepts (Claims 2, 3). However, the next two levels are designed to capture the heart of deep learning. The Conceptual level involves the processes associated with evaluating, integrating, and synthesizing multiple texts for a particular purpose (claims 1, 4, 5 and 7). The Social level is designed to measure students' ability to engage in perspective taking, social reasoning, productive discourse, and interpreting character, author, and collaborator's motives and intent (claims 6, 7).

During an SBA, students are asked to complete several complex tasks that tap deep understanding as described in this chapter. For instance, they may be asked to produce a flier that provides reasons and evidence for what to do with an empty city lot. They may be asked to write an essay that provides evidence for which of two theories best explains how early people arrived in North America. Prior to completing these larger tasks, students may be asked to draw inferences (claim 2),

complete summaries, or graphic organizers (claim 3) or identify relevant questions (claim 4). Other tasks require students to evaluate websites, authors, and evidence (claim 5), as they integrate multiple sources and engage in perspective taking, social reasoning, error detection, and repair (claim 6). They may also be asked to apply definitions, concepts, or principles to new situations (claim 7). Work from outside our lab suggests that our SBAs are measuring aspects of deep understanding. For instance, LaRusso et al., (2016) found that SBAs correlate with independent measures of deep understanding (as they defined them) such as complex reasoning, perspective taking, and academic vocabulary.

While the SBAs are designed to capture these and other skills, measuring whether students can do complex tasks is not particularly useful for instruction. Although complexity and deep understanding are worthy targets for an assessment, not all students have mastered these skills. In other words, if you build a traditional summative assessment that measures deep understanding, the scores will likely indicate that many students are not proficient. The question then becomes: How do you both measure deep understanding, while simultaneously providing information that is useful for instruction? Traditionally, there has been a firewall between summative assessment and instruction; summative assessment is designed to measure student competency, not support it. In contrast, we argue that summative assessment can, and should, measure achievement, while at the same time, support it.

To achieve these and other aims, SBAs are designed to structure assessment items and tasks to promote deep learning. During a typical SBA, students are provided with a realistic purpose for reading a collection of thematically related sources (Claim 1) (e.g., Should your neighborhood build a community garden?). While the sources are on a similar topic, they are diverse in format (e.g., e-mail, blog, website, and traditional printed text) and type of author (e.g., scientist, government, friend, community, company). Some sources may support the claims made in other sources, while other claims may be contradictory across sources. On balance, test takers are asked to evaluate the sources as they integrate and synthesize information (claim 5) to make a decision, solve a problem, or transfer what they learned to a new situation (claim 7).

To promote learning, the tasks and activities are structured and sequenced to help reveal what parts of the more complex task students can or cannot do. For instance, before students read any texts, their background knowledge is measured to determine what they know about the topic prior to the assessment. Subsequently, students read texts that build up their knowledge of the topic before they are required to do more complex tasks. These steps are taken to ensure they have adequate knowledge to engage in deeper thinking. The initial knowledge building tasks measure students' basic understanding of single texts through literal or inferential questions (claim 2). Students are also asked questions that require them to assemble the pieces together at a global level in the form of writing summaries or creating graphic organizers (claim 3), and identifying relevant questions to clarify meaning (claim 4). As the scenario unfolds, new information is introduced and the student is required to integrate this information as they proceed. In tandem, these features are designed to promote the use of instructionally relevant skills. By asking these questions on an assessment, we expect subsequent instructional approaches would follow suit.

To facilitate the process of learning, a set of simulated peer students are included in all SBA designs to model social interaction (claim 6). These may include a simulated teacher, or a similar authority figure, as well as a set of simulated students. The simulated characters are designed to serve several functions. In some cases, they provide hints and guidelines for task goals. For instance, they

may provide guidelines on how to write a summary or how to evaluate a website. They may help test takers by summarizing what has been learned so far in the assessment, or provide encouragement and feedback about key issues the test taker might have missed up to that point in the assessment. The peers may also comment on their understanding of the texts and sources (e.g., in a threaded discussion, peers may present information that is off topic). In many cases, the simulated peers present incorrect interpretations of the sources and the test taker is asked to identify and correct the errors. In this manner, the assessment is designed to promote and model effective metacognitive and self-regulatory behaviors.

Task decomposition and knowledge building sequences help the test taker form an initial understanding, and allow assessment designers to determine what parts of the more complex task students can or cannot do. For instance, some tasks may ask students to write a summary while others may ask them to evaluate a given summary based on specific criteria. When tasks are sequenced and structured properly, we can determine whether a student did not understand a key idea, a collection of ideas about a single text, the integration of ideas across multiple texts, whether they could evaluate what they read, or whether they could understand and correct misconceptions of simulated peers, solve the larger problem, or make a key decision based on reasons and evidence. These features are not only diagnostic, they are also designed to support deep learning through strategy use and social support. Thus, the assessment can both measure and support deep understanding at the same time.

Summary and conclusion

The way that people communicate and interact has changed how we view reading in the 21st century. These changes have prompted researchers and educators to re-conceptualize what we mean by deep understanding and how we should measure it. In this paper, we offer seven claims that help define deep understanding. We also provided an overview of SBAs and how they can be used to both measure and support learning in the context of instruction. We were prompted to do so, because we were asked to address a deep question in this chapter: what is deep understanding? A question that encouraged us to follow in the footsteps of Art Graesser. A question that led us to a deeper learning and understanding of our discipline. A question worthy of a larger goal of building innovations that serve the human condition. We hope that SBAs can serve as one vehicle of change in the assessment world that could have a positive impact on learning and instruction; we can only hope that future generations of SBAs can foster and perhaps mentor students to think more deeply about their learning and the world around them.

References

Beck, I., McKeown, M., & Gromoll, E. (1989). Learning from social studies texts. *Cognition and Instruction, 6*, 99–158.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Assessment issues of the 21st Century* (pp. 43–61). New York, NY: Springer.

Black, P. & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Britt, A., J. F, Rouet, & Durik, A. M. (2017). *Literacy Beyond Text Comprehension: A Theory of Purposeful Reading*. New York, NY: Routledge.

Deane, P., Sabatini, J., Feng, G., Sparks, J., Song, Y., Fowles, M., O'Reilly, T., Jueds, K., Krovetz, R., & Foley, C. (2015). Key practices in the English Language Arts: Linking learning theory, assessment and instruction. ETS Research Report Series RR-15-17.

Foy, J. E., LoCasto, P. C., Briner, S. W., & Dyar, S. (2017). Would a madman have been so wise as this?' The effects of source credibility and message credibility on validation. *Memory & Cognition, 45*(2), 281-295. doi:10.3758/s13421-016-0656-1

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.

García, F. C., García, Á., Berbén, A. G., Pichardo, M. C., & Justicia, F. (2014). The effects of question-generation training on metacognitive knowledge, self regulation and learning approaches in Science. *Psicothema, 26*(3), 385-390.

Goldman, S. (2012). Adolescent literacy: Learning and understanding content. *Future of Children, 22*, 89–116.

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., & ... Shanahan, C. (2016). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist, 51*(2), 219-246. doi:10.1080/00461520.2016.1168741

Gordon Commission. (2013). *To assess, to teach, to learn: A vision for the future of assessment*. Princeton, NJ: Author. Retrieved from: http://www.gordoncommission.org/rsc/pdfs/gordon_commission_technical_report.pdf

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. American Educational Research Journal, 31, 104–137.

Graesser, A. C., Person N. K., & Huber, J. D. (1992). Mechanisms that generate In T. E. Lauer, E. Peacock, & A. C. Graesser (Eds.), *Questions and information systems* (pp. 167-187). Hillsdale, NJ:

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101, 371–395. doi:10.1037/0033-295X.101.3.371.

Graesser, A. C., Wiley, J., Goldman, S., O'Reilly, T., Jeon, M., & McDaniel, B. (2007). SEEK Web tutor: Fostering a critical stance while exploring the causes of volcanic eruption. *Metacognition and Learning, 2*, 89–105.

Hannon, B., & Daneman, M. (1998). Facilitating knowledge-based inferences in less-skilled readers. *Contemporary Educational Psychology, 23*(2), 149-172. doi:10.1006/ceps.1997.0968

King, A. (1995). Inquiring minds really do want to know: using questioning to teach critical thinking. *Teaching of Psychology, 22*, 13-17.

LaRusso, M., Kim, H. Y., Selman, R., Uccelli, P., Dawson, T., Jones, S., . . . Snow, C. (2016). Contributions of academic language, perspective taking, and complex reasoning to deep reading comprehension. *Journal of Research on Educational Effectiveness, 9*(2), 201-222.

Leu, D., Kinzer, C., Coiro, J., Castek, J., & Henry, L. (2013). *New literacies: A dual-level theory of the changing nature of literacy, instruction, and assessment*. In D. E. Alvermann, N. J. Unrau, & R. B. Ruddell (Eds.), Theoretical models and processes of reading (6th ed., pp. 1150–1181). Newark, DE: International Reading Association.

Luschen, K., & Bogad, L. (2010). Youth, new media and education: An introduction. Educational Studies: *Journal Of The American Educational Studies Association, 46*(5), 450-456. doi:10.1080/00131946.2010.510402

Magliano, J., McCrudden, M., Rouet, J. F., & Sabatini, J. (in press). The Modern Reader

Should Changes to How We Read Affect Research and Theory?

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review, 99*, 440–466.

McNamara, D. S. (2004). SERT: Self-explanation reading training. Discourse Processes, 38, 1–30.

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

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. Journal of the American Society for Information Science and Technology, 58(13), 2078–2091.

Minguela, M., Solé, I., & Pieschl, S. (2015). Flexible self-regulated reading as a cue for deep comprehension: Evidence from online and offline measures. *Reading And Writing, 28*(5), 721-744. doi:10.1007/s11145-015-9547-2

Mislevy, R. J. (2009). Validity from the perspective of model-based reasoning. In R. L. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 83–108). Charlotte, NC: Information Age Publishing.

Mislevy, R. J. (2008). How cognitive science challenges the educational measurement tradition. *Measurement: Interdisciplinary Research and Perspectives, 6*, 124.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. *Discourse Processes 43*(2), 121-152.

O'Reilly, T., & Sabatini, J. (2013). *Reading for Understanding: How Performance Moderators and Scenarios Impact Assessment Design* (Research Report No. RR-13-31). Princeton, NJ: Educational Testing Service.

O'Reilly, T., & Sheehan, K. M. (2009). *Cognitively based assessment of, for and as learning: a framework for assessing reading competency* (RM-09-26). Princeton, NJ: Educational Testing Service.

Pascual, G., & Goikoetxea, E. (2014). Comprehension of university texts: Effects of domain-knowledge and summary. *Reading Psychology, 35*(2), 101-126. doi:10.1080/02702711.2012.664612

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007-2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Sabatini, J., O'Reilly, T., & Albro, E. (2012). *Reaching an understanding: Innovations in how we view reading assessment.* Lanham, MD: Rowman & Littlefield.

Sabatini, J., & O'Reilly, T. (2013). *Rationale for a New Generation of Reading Comprehension Assessments.* In Miller, B., Cutting, L., & P. McCardle (Eds), Unraveling Reading Comprehension: Behavioral, Neurobiological, and Genetic Components, (pp. 100-111). Baltimore, MD: Brookes Publishing.

Sabatini, J., O'Reilly, T., & Deane, P. (2013). *Preliminary Reading Literacy Assessment Framework: Foundation and Rationale for Assessment and System Design*. (Research Report No. RR-13-30). Princeton, NJ: Educational Testing Service.

Robinson, D. H., & Kiewra, K. A. (1995). Visual argument: graphic organizers are superior to outlines in improving learning from text. *Journal of Educational Psychology, 87*, 455-467.

Trejo, B. C., Richard, E. M., van Driel, M., & McDonald, D. P. (2015). Cross-cultural competence: The role of emotion regulation ability and optimism. *Military Psychology, 27*(5), 276-286. doi:10.1037/mil0000081

van den Broek, P., Lorch, Jr., R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory & Cognition, 29*, 1081–1087.

van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). *The landscape model of reading*. In. H. van Oostendorp & S. R. Goldman (Eds.), The construction of mental representations during reading (pp. 71–98). Mahwah, NJ: Erlbaum.

Williams, J. (2007). Literacy in the curriculum: Integrating text structure and content area instruction. In D. S. McNamara (Ed.), *Reading Comprehension Strategies: Theory, Interventions, and Technologies* (pp 199-219). Mahwah, NJ: Erlbaum.