Scenario-based assessment of multiple source use

John Sabatini

Tenaha O'Reilly

Zuowei Wang

Kelsey Dreier

Educational Testing Service, USA

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305F100005 to the Educational Testing Service as part of the Reading for Understanding Research (RFU) Initiative. 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. We want to thank Kim Fryer, Jim Carlson, Paul Deane, and Jesse Sparks for helpful comments and editorial assistance.

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

Sabatini, J., O'Reilly, T., Wang, Z., & Dreier, K. (2018). Scenario-based assessment of multiple source use. In J. L. G. Braasch, I. Bråten, & M. T. McCrudden (Eds.), *The handbook of multiple source use* (pp. 447–465). New York, NY: Taylor & Francis.

Scenario-based assessment of multiple source use

Prologue

This book is a testament to the increasing role and importance of multiple source use in everyday and academic literacy activities in the 21st century. How should we introduce the topic of multiple sources here? For most readers, we need not, because this chapter is not their first literacy stop in the volume, so the topic has been adequately covered in other sources. As authors, we are keenly aware that our readers already may have developed mental models and critical stances that will influence the understanding and interpretation of the content we are about to present to them. No one is a *tabula rasa*. In the spirit of multiple source literature, we can only wonder whether the reader's aims correspond to the intended aims of this chapter, whether the text is relevant for their purpose for reading, and whether the mental models formed from reading prior chapters conflict or otherwise influence their interpretations of ours.

Overview of the Paper

Our plan is as follows. First, we present a recent and widely accepted conceptualization of 21st century constructs for understanding of single and multiple text sources, the MD-TRACE model (Rouet & Britt, 2011), as an analytic framework for thinking about assessment of multiple source use. In doing so, we acknowledge that the citations across this entire volume represent key sources that should inform a comprehensive assessment framework, but space limitations preclude integrating more of them into this chapter.

Next, we describe, explain, and justify the use of scenario-based assessments (SBA) as an approach to measuring multiple source use. The knowledge, skills, strategies, and dispositions necessary for proficiency in multiple source use pose a challenge to traditional assessment designs. On the one hand, multiple-choice paradigms that have privileged discrete, independent

items and tasks tend to make it difficult to elicit the complex, cross-source, cognitive activities that are core to multiple source processing (Sabatini, Petscher, O'Reilly, & Truckenmiller, 2015). On the other hand, traditional constructed response (mostly essay-based) and performance assessments have analogous challenges and limitations related to reliability, bias, and added value (Hift, 2014; Kafer, 2002; Lukhele, Thissen, & Wainer, 1994). Specifically, most exemplars of this approach in use today, such as the College and Work Readiness Assessment (or CWRA+) (Council for Aid to Education, 2017), combine multiple source reading and writing skills that culminate in a single, complex, writing task.

We view the cluster of techniques that comprise SBA as an alternative approach to assessment design that can be used to mediate or overcome many of the challenges and limitations associated with measuring multiple source use. We discuss different conceptualizations of SBAs in research by describing several of the most developed research programs and exemplars of SBAs. We also discuss desired consequences of using SBA approaches to impact learning and instruction, and how they are being applied in operational testing programs.

Multiple Source Use Constructs: What Is It and How Can We Assess It?

Multiple sources are not a new phenomenon to reading instruction or even reading assessment. For example, conducting a literature review, which requires accessing, evaluating, and synthesizing multiple sources, is a staple activity taught in schools. The *Advanced Placement* history exam test routinely includes a "document based" free response section that has students analyze and synthesize a set of multiple documents in order to explain a key historical event (College Board, 2017).

What has changed is the volume and diversity of sources available, as a result of the advent of the Internet and World Wide Web (Leu, Kinzer, Coiro, & Cammack, 2004). In turn, renewed interest in cognitive research on document use has informed more complex, multi-source models (Kirsch & Mosenthal, 1990; Mosenthal & Kirsch, 1991; Rouet, 2006). This has led to the need to expand the construct coverage expected in the assessment of reading comprehension expertise to take into account the wide repertoire of flexible and differentiated processes that are needed to achieve complex task goals that require examining multiple text sources (Rouet & Britt, 2011; Sabatini, O'Reilly, & Deane, 2013).

In this paper, we draw inspiration from the PISA reading framework in defining multiple source comprehension. Multiple sources are defined here as any collection of texts that have been written by a single author (or co-authors), but published across multiple time points, or any collection of texts written by different authors. These could be dynamic (e.g., hyperlinked) or static texts, printed texts, emails, blogs, webpages, or other digital sources. They can also include multimedia such as audio files, pictures, and videos. The diversity of text type and modality, coupled with an increased level of access to multiple sources, has placed additional demands on attention and resource allocation for the 21st century reader.

During single text comprehension tasks, some of the processes involved in multiple source comprehension are not called upon or are less complex to deploy, as a single source "should" be internally coherent in terms of its goals, arguments, and intended audience. While single texts may introduce controversies, they are usually written from one perspective and the related information is within close proximity. In contrast, when reading multiple sources, the reader usually has a specific goal directed at answering an overarching question, of which, only some sources are relevant or select elements of a source. Importantly, the reader goals may

differ from the author's intended goal for writing, requiring additional processes to identify, select, and interpret information relevant to the reader's aims. In multiple source processing, readers need to find information that is relevant to their goal, evaluate it for credibility, and corroborate sources to achieve their aims. Sources may support some points, while other sources may contradict each other. While some of these processes are required in single source reading, the source author often provides guidance regarding integrating, synthesizing, or representing conflicting information for the reader, presenting it from a single point of view, or explaining when different points of view are being represented. While the demand for multiple source skills has increased with wide access and use of the Internet, how are multiple source skills assessed? To address this, we need to understand the goals and purposes of assessments.

Assessments are used for many purposes, but chief among these are 1) to evaluate whether and how much individuals have learned or achieved in a domain; 2) to predict how well or whether individuals can apply what they know and can do in a context of use; or 3) to aid or guide instructional decisions and learning. For the first purpose, it may be sufficient to examine recall of taught knowledge and skills in a relatively decontextualized manner. In this case, the student is expected to act independently and no support for learning is provided. Conclusions drawn from such an assessment would indicate whether the student has learned factual or maybe even conceptual content. However, when one is assessing the application of skills when thinking or reasoning about content learned, then a different assessment strategy may be required. One could ask the learner to complete a complex and integrated task (e.g., write an essay that evaluates and integrates multiple sources), such that skills are called upon to engage and solve a novel problem.

Further, for inferences about depth of understanding in a context of use (the second purpose of assessment), ecological validity in the assessment strengthens the validity of claims that individuals can apply what they have learned and that the construct has been adequately measured. That is, we might ask whether the scores produced by a decontextualized assessment transfer to more realistic settings? In addition, when we want to inform future learning (the third purpose), we may want to know whether a low score on a complex task, such as an essay, essentially mean the student did not have any of the sub skills that feed into the more complex task.

In this chapter, we describe a new type of assessment called Scenario Based Assessments (SBAs) that provides both a framework for assessing multiple source use and takes a step towards ecological validity in assessments, in that it establishes credible literacy purposes or goals for the individual, and a structured, sequenced set of tasks or subgoals towards achieving those purposes. This also aligns such assessments better with contemporary models of discourse processing and reading comprehension (e.g., Kintsch, 1998; Rouet, 2006; Rouet & Britt, 2011). In this way SBAs can sample both complex, integrated performances (Can a student evaluate and integrate information independently?) and some of the key sub skills that support those performances (Can an assessment support and measure partial skill development, e.g., evaluate the credibility of a website?). The SBAs reviewed in this chapter take a further step, attempting to reflect what Bennett (2010) calls 'assessments as learning', that is, assessments that serve as models of or aids to learning and instruction.

MD-TRACE Model: Deconstructing the Construct of Multiple Source Use

As stated earlier, one of the goals of assessment is to measure a student's independent performance on a task. Another goal of assessment is to support learning that may also involve

obtaining information on whether a student has mastered any of the subcomponents of the more complex skill. To achieve this second aim, it would be useful to deconstruct and identify the essential elements of multiple source use so that assessments can be designed to measure the key sub skills.

To illustrate how multiple sources fit into assessment contexts, we employ a well-known, cognitive framework of multiple source processing, specifically, the Multiple Documents – Text-based Relevance Assessment and Content Extraction (MD-TRACE) model (Rouet & Britt, 2011). To summarize, the MD-TRACE model is comprised of internal and external resources, and cognitive activities described as a set of steps or processes. The external resources consist of the external task requirements; search devices, source material, and text organizers; document contents; and reader generated products (such as notes, summaries, or essays). The internal resources consist of the task model (the internal representation of the external task requirements) and the documents model (the representation that is the product of the cognitive operations). These internal resources are moderated by prior knowledge, reading/search skills, and self-regulation skills.

The MD-TRACE cognitive activities are decomposed into five interacting processing steps: 1) create and update a task model; 2) assess information needs; 3a) assess item relevance; 3b) process text contents; 3c) create/update a documents model; 4) create/update a task product; 5) assess whether the product meets the task goals. The authors note that the steps may occur out of order or in parallel in actual task performance. Embedded in each of these steps are multiple cognitive activities, plans, evaluations, and decisions.

From Cognitive Model to Assessment Design

It is a good moment to step back and remind ourselves of the relevant goals of assessment in contrast to a cognitive model. A cognitive model is a description or explanation of a process, in this case, multiple source processing. An assessment is an information gathering tool. So, while the MD-TRACE model is a detailed description of the reading sub skills, processes, and strategies necessary to use multiple sources to achieve a goal, it is precisely the question of whether, or to what extent, a student possesses each sub skill that the assessments we design seek to uncover.

In outcome tests, one is first and foremost interested in whether the individual has the relevant cognitive knowledge, skills, strategies, and disposition (hereafter simplified to cognitive skills). Typically, one derives a score that represents a point on a continuum of proficiency. Ideally, participants are required to apply their cognitive skills in task and text sets that are similar to or at least predictive of performance in applied, real world settings. Steps to minimize the influence of construct irrelevant variance or bias are taken, for example, trying to reduce the influence of background knowledge by using "familiar topics", using standard administration or scoring procedures, and avoiding sensitive topics that could disadvantage some groups or individuals. Constraints are put in place to ensure reliable, consistent scoring including the use of multiple-choice items, restricting the search and use of outside documents, creating effective scoring rubrics, and conducting training sessions for raters to ensure high inter-rater reliability of constructed responses (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

Collectively, the focus of these measurement standards is to ensure that the *product* of comprehension, defined here as the responses and the scores derived from them, yield reliable and valid inferences of the test taker's relative proficiency on the construct of interest. Of far

less interest in traditional testing paradigms, is the *process* by which the test taker responds, or which steps, activities, or processes (or sub skills) yielded the correct versus incorrect responses. In other words, traditional assessment is concerned more about whether a student is proficient or not, and less about *how* or *why* a particular student received a score, or whether a student was able to *do parts* of a more complex task.

The Assessment Paradox: How To Measure and Support Multiple Source Use

The above discussion illustrates the discrepancy between what is valued in cognitive models of reading and what is measured by traditional reading assessments. Theoretical models such as the MD-TRACE place emphasis on the growing importance of multiple source use in today's digital world. The model outlines the key skills and processing steps that proficient readers need to follow in order to successfully undertake the metacognitive, evaluative, and integrative mindset for 21st century multiple source reading environments. In short, the model presents a stance on what is important to measure, and underscores the importance of the component processes that are applied in achieving proficient performance.

In contrast, traditional assessment paradigms prioritize measuring reading ability in an efficient and cost effective way. In most current tests, this has meant measuring student ability to comprehend single passages in isolation. Passages are chosen to not demand much topical, background knowledge and the passages are not intended to be related to one another. There is no overarching goal for reading other than to answer the questions accurately (Rupp, Ferne, & Choi, 2006), and the questions are also assumed to be independent of each other. This is not to say that traditional assessments of reading ability are not valid for some inferences of proficiency; the assessments typically have strong psychometric properties and the scores are

predictive of success on a number of metrics. However, this paradigmatic design poses significant constraints in adapting to a changing, multiple source literacy environment.

The key challenge for assessment designers then, is how to assess modern constructs (and sub constructs) of reading including constructs such as multiple source use, while simultaneously providing information that is psychometrically sound. It would also be of value to gather evidence of the process that leads to the final goal (e.g., understanding the goal, assess document relevance), preferably without the need to parse each sub-skill into discrete, decontextualized items.

In SBA designs, both process and product can be considered (O'Reilly & Sabatini, 2013; Sabatini et al., 2013). One can apply cognitive and learning science insights in the assessment design with an aim of enhancing the assessment's instructional relevance and construct coverage. SBAs have the potential to enhance such construct and instructional utility and are especially well suited for multiple source assessment, over traditional testing paradigms. In the remainder of the chapter, we describe three well-developed research programs that have pioneered the development of SBAs and how they address the construct of multiple source use.

The Global, Integrated Scenario-based Assessment (GISA) approach

The scenario-based, reading comprehension assessments which we call global, integrated scenario-based assessments (GISA) can be useful for achieving a number of such construct and process goals, while maintaining psychometric integrity. While the approach was not designed to explicitly measure multiple source use, many of the design features can be used to both support and measure many of the key elements of multiple source comprehension. O'Reilly and Sabatini (2013) defined SBAs as a collection of techniques that allow test designers to *structure* tasks and items to enhance construct coverage using valid task designs, with a goal of enhancing

instructional value. Unlike traditional assessments that present texts and tasks in a discrete manner, the GISA approach to SBAs *organize the texts and tasks* into units of integrated activities, rather than a collection of stand-alone items. This approach to structuring and sequencing makes the SBA inherently amenable to assessing the steps of multiple source use.

Key Features of GISA

The key features of the GISA SBA approach include: 1) an initial goal and context for reading; 2) a collection of thematically related sources; 3) a set of simulated social agents (i.e., simulated peers, teacher); and 4) a set of techniques to model good reading habits, while simultaneously providing opportunities for students to demonstrate their partial skill development. In addition, because reading is an integrated activity, what we call performance moderators are also included in the design. The performance moderators of background knowledge, motivation, metacognition/self-regulation, and reading strategies are not directly considered a part of the construct, but may impact reading performance (O'Reilly & Sabatini, 2013). In many GISA assessments, background knowledge is directly measured and we look for indicators of student motivation. Reading strategies are often incorporated into GISA forms as specific reading tasks (e.g., summary, graphic organizer).

How the Features are Assembled in an Assessment Context

How are these features implemented? In a typical GISA form, students are given a specific goal for reading a collection of thematically related materials (e.g., should your school adopt a clean energy program). The sources are chosen to be diverse in terms of *format* (e.g., blog, email, website, textbook passage), *depth* (e.g., comprehensive, selective), trustworthiness (e.g., reliable or unreliable sources) and *accuracy* (e.g., contains errors or misconceptions). This diversity is not only used to set the stage for multiple source use, but also to engage students in

critical thinking, encourage perspective taking, and broaden students' awareness and appreciation of genre similarities and differences. Learning new ideas across different text formats and contexts is also designed to promote transfer by not restricting the conditions of learning to a single source. While the sources are diverse on a number of dimensions, they are all related to each other at some level in connection to the reading goal.

Tasks and items in an SBA require the students to engage in both traditional "single text" forms of reading (e.g., identify key ideas, draw local inferences), and more demanding multiple source tasks. Multiple source tasks may require integrating and synthesizing cross-textual information, evaluating source utility and trustworthiness, identifying and potentially resolving conflicting claims and evidence, or making decisions about how to apply text content in new situations or contexts (transfer). In GISA forms, all of these activities are thematically related to achieving the larger, scenario goal.

These features of GISA SBAs enable designs that extend beyond the traditional, discrete item paradigms of reading assessment, better aligning the assessment with modern models of goal driven, multiple source processing accounts of reading (e.g., Magliano, McCrudden, Rouet & Sabatini, in press; Rouet & Britt, 2011). In addition, GISA forms are designed to model good reading habits, as well as provide opportunities for test takers to demonstrate partial skill development (e.g., the stages in the MD-TRACE model).

Illustrating Multiple Source Use Through a GISA Example: Connecting the MD-TRACE Model to Assessment

To illustrate how some of these features work together, we briefly describe the structure and sections (tasks) of a GISA form on the topic of community gardens, and how these GISA sections roughly correspond to stages of the MD-TRACE model (see Table 1). While the GISA

form was designed to measure constructs that go beyond multiple source use, it does cover many aspects of the key sub skills involved in multiple source use. Table 1 includes the section number, name, and intended key function. The key function describes what the section was designed to do, which includes, and may go beyond, multiple source use constructs. The table also includes a column that identifies the elements of the MD-TRACE that roughly correspond to the GISA sections of the community gardens assessment.

According to the MD-TRACE model, students undertake multiple activities in the course of reading and understanding multiple source texts. Some of these are strategic in nature, and *may* involve interactions outside of text comprehension in a more restricted sense (e.g., defining a task model, defining information needs, assessing item relevance, and assessing whether a task product meets task requirements). Other activities may involve skills characteristic of a traditional, single source reading construct (processing text content), while others require critical thinking and written communication (e.g., building a documents model and creating a task product). How does the GISA SBA align with this model?

Section 1: Setting Up the Task Model: What Are Students Supposed to Do and Produce?

The community gardens assessment is a 45-minute, computer-delivered assessment designed for use with 5th to 6th-grade students. In the introductory section of the assessment, students are given the goal of helping decide whether or not to build a community garden on a vacant lot. Sub-goals include: find out more about community gardens; decide if they support the construction of the community garden; and prepare a flyer to inform community members of their recommendation. During this process, simulated peers are introduced who help define task goals, support the test taker by providing hints, or provide stimuli that need to be evaluated by the test taker. In the introductory section, the test taker is also introduced to a simulated teacher

who will provide guidance as the test taker and simulated peers "work" on the community gardens project.

In the MD-TRACE model, this first section of the GISA corresponds to the "create a task model" step (Table 1, row 1). Note that the task model goals and sub-goals are structured and organized for the students in this scenario. We do not directly assess students' ability to formulate a complex task model, though future SBA forms could target those multiple source sub-skills in the design.

Section 2: Measuring Background Knowledge: What Do Students Know About the Topic?

The thematic nature of a scenario-based assessment could be considered a limitation of the design. This is because students enter a scenario with different levels of prior knowledge about the topic, and with different levels of skill in performing each of the component tasks. This variability can cause students with very different skills profiles to perform similarly. For instance, a student with high knowledge of the topic, but weak ability to integrate information from multiple sources might produce a final product of similar quality as a student who had low initial knowledge on the topic, but strong text integration skills (e.g. a quick learner of a new domain). The GISA approach is designed to capture information about different phases of this complex process, making it easier to develop skills profiles that suggest hypotheses about why students did well or poorly on a specific GISA form.

For example, background knowledge is not in and of itself part of multiple-text reading comprehension, but it is a performance moderator (in the GISA framework and in MD-TRACE), since some students will enter a task with high or low knowledge, and may learn the content presented more or less quickly and completely as the task unfolds. To address the potential limitation of the thematic nature of GISA, the form is designed to take variability in levels of

knowledge into account. Therefore, at the beginning of the assessment, students' background knowledge is measured. In this case, the background knowledge concerns the topics of community gardens and farming in general. This step provides the test user with some evidence to determine whether the students had sufficient knowledge to understand the topic, or whether they had so much knowledge that the assessment is essentially a test of knowledge, rather than a test of comprehension. In addition, some of the background questions will be answered in the text of subsequent passages. This feature also allows the assessment to measure whether or not the students *learn* passage content from reading versus recalling it from prior knowledge (see Table 1, row 2).

Section 3: Building up Students' Understanding: Single Source Comprehension

The GISA form is also sequenced to *build* background knowledge up over the course of the assessment (Table 1, row 3). For instance, an initial text introduces the topic of community gardens. To support (and evaluate) test takers understanding of key ideas in this introductory text, we present a sequence of tasks that probes their global understanding. In addition, to support low knowledge readers, the peers engage in a dialogue that explains what a community garden is. Students are also introduced to the controversy that drives the scenario goal: some groups want to build a playground on the vacant lot, others want to build a community garden on the lot.

Students may have weak, single-text reading skills, or may be inclined to put as little effort into reading as is feasible given the task they are assigned. The GISA SBA addresses motivation by providing a more meaningful, and scaffolded, task sequence, while measuring single-text reading comprehension with tasks (such as summarization) that also encourage students to build the deeper knowledge they will need during task integration as the assessment

unfolds. For example, in a four part sequence, we gather information on whether test takers can write a summary independently, whether they can detect if a fellow student's summary violates one of the guidelines, whether they can locate where in the summary the violation occurred, and whether they can fix the error. Such sequencing is useful for identifying what parts of the more complex task a student could or could not do. Collectively, this section is related to the "process text contents" part of the MD-TRACE model, but also provides clues as to whether test takers have developed some accuracy in their mental model of the single text, when later they are required to integrate this knowledge into a multiple source use task where it is applied (Gil, Bråten, Vidal-Abarca & Strømsø, 2010).

The above sequence is intended to measure basic single text understanding (Table 1, row 3). Other tasks in the section are designed to help test takers formulate their argument, and can be thought of as fostering/assessing multiple source processing. For instance, the test taker is asked to complete a graphic organizer (a reading strategy) that outlines who supports a particular position and whether the text provides information to make the position judgment. Requiring that students recognize who supports what position aligns with assessing multiple source use.

Section 4: Evaluating Web Links: Assess Item Relevance

After the introductory text and tasks are presented, the test taker is given a list of simulated websites that contain the URL and a short description similar to what one would find in a typical search engine output (Table 1, row 4). Some of the sites are relevant to the task goal and others are not. The test taker is asked to identify the relevant sources, as well as engage in some perspective taking tasks. A subsequent task asks the test taker to identify the parts of an actual website that are useful towards their goal.

Section 5: Gaining a Deeper Understanding of the Topic: Update Task Model and Create a Documents Model

Next, test takers are given a second, detailed article that explains more about community gardens, what they are, why people create them, and their relation to real-world problems, such as food deserts. This second text is key for building an argument supporting the creation of a community garden (Table 1, row 5). Items in this section measure understanding through the use of graphic organizers, identifying causes and effects, and identifying correct and incorrect paraphrases of key information. Also part of this section are items tapping knowledge of key vocabulary that was formerly presented in the background knowledge section. Here, we measure if students learned formerly unknown terms (or whether they knew terms prior to taking the assessment). To support argumentation, another task asks students to classify statements that support a position.

Section 6: Opposing Viewpoints and Counterargument: Update Documents Model

The subsequent task presents a third text that opposes the building of the community garden and offers reasons for building a playground in its place (Table 1, row 6). This section requires the students to "process text contents" and "update their documents model" (in the MD-TRACE process) by presenting information that is not consistent with prior texts. Here, the test taker can compare the conflicting arguments across multiple sources.

Section 7: Produce a Flier: Create a Task Product

The culminating task requires the test takers to complete parts of a flyer (Table 1, row 7). In particular, they are asked to take a position and provide reasons that justify the position. In theory, test takers' decision to take a side should be based on their ability to weigh the evidence on both sides of the argument as they consider the information across multiple sources. The first

task requires the test taker to provide this information in a constructed response format (open ended), after which the test taker is given a second attempt, this time with selected responses. This sequence allows less skilled readers to demonstrate their partial skill development. The task is related to the "create/update a task product" part of the MD-TRACE model.

In sum, this example from the community gardens GISA form was used to illustrate many of the features of SBAs and how they can be applied to measuring aspects of multiple source use. While multiple source measurement was not the sole construct targeted in the assessment, it does include tasks that call upon most of the processing steps of the MD-TRACE model. At the same time, it probes and monitors student performance on single source texts, to help identify and distinguish single versus multiple source skill strengths versus weaknesses. We believe that the set of GISA SBA features, coupled with performance moderators, enables richer construct coverage of goal-driven, single and multiple source comprehension than discrete item, multiple choice or single, culminating writing task test designs. While SBA's do take a lot of thought to design and implement, the extra effort may pay off and they are easier to build once initial designs are developed.

Properties of GISA

While there is much more work to be done, we have created over 20 SBAs for students in grades pre-K through 12th grade. Most of these SBAs have multiple source use tasks similar to those in the community garden form, though the difficulty and support for students vary across developmental levels. We have piloted them in several states and locales with large numbers of students. Our analyses reveal that the assessments demonstrate adequate psychometric properties including reliability, variability of scores, and appropriateness for the intended population (O'Reilly Weeks, Sabatini, Halderman, & Steinberg, 2014; Sabatini, Halderman,

O'Reilly, & Weeks, 2016; Sabatini, O'Reilly, Halderman, & Bruce, 2014a, 2014b). We have also created a vertical scale among the GISA forms, which allows comparisons across grade levels and forms. In other words, even if students take different GISA forms, their scores can still be compared to each other, thanks to the vertical scale. This is useful in pre/post intervention designs and to explore changes in development over time. In short, we believe the scenario-based assessment is a feasible way to measure reading comprehension inclusive of multiple source use.

Applications of GISA

Elements of the GISA approach, including measurement of multiple source use, have been operationally implemented in the PISA reading literacy assessments. The triennial Programme for International Assessment (PISA) surveys 15-year-old students around the world, assessing the extent to which students near the end of compulsory education have acquired key knowledge and skills that are essential for full participation in modern societies. The PISA Reading Literacy Framework discusses and cites the GISA framework and designs of SBA tasks have been developed for use in the 2018 implementation of PISA, where reading literacy is the main cognitive focus.

Other Scenario-based Assessment Research Programs

To our knowledge, there are only a limited number of other research programs currently investigating the use of scenario-based assessment of reading and writing that could also be considered measures of multiple source processing. We summarize the research of each program, with strong emphasis on how they define and operationalize SBAs to address elements of multiple source constructs. We do not review the work of Goldman and colleagues in this

chapter as the research is covered elsewhere in this book (Goldman, Blair, & Burkett, 2018/this volume).

The ORCA (Online Research and Comprehension Assessment) Project

As a response to the proliferation of new literacies, the ORCA, Online Research and Comprehension Assessment, was created. ORCA was developed under a new literacies perspective and defines online reading comprehension as a "web-based problem-solving inquiry process involving skills and strategies for locating, critically evaluating, synthesizing, and communicating information on the Internet" (Coiro, 2011 p. 352). The strategies involved in the forms are summarized by the acronym LESC, which stands for reading to *locate*, reading to *evaluate*, reading to *synthesize*, and reading and writing to *communicate*; several of the processes described by "LESC" align to the cognitive activities of the MD-TRACE model.

Three types of ORCA forms were originally created, and each form was designed with a problem-solving scenario related to human biology, a subject familiar to most 7th grade students, who provided the majority of the assessment's sample population. The initial forms, ORCA-Open and ORCA-MC (multiple choice), were piloted first in 7th grade classrooms (Coiro & Dobler, 2007). The ORCA-closed form was created and piloted later, based on the iSkills assessment (Katz, 2007) and the digital literacy assessment in PISA (Organisation for Economic Co-operation and Development, 2011). ORCA-Open lets students navigate the open internet while ORCA-Closed, similar to a scenario based assessment, requires students to write a report within a simulated online environment that includes several internet capabilities like instant messaging, search engines, emails, and several web pages. ORCA-MC was created similarly and used the same content, except that all student response types were multiple choice items. After extensive pilot research, the researchers decided that the ORCA-Closed and multiple

choice had adequate feasibility and psychometric properties for continued development (Leu et al., 2014).

The ORCA-Closed forms had students utilize multiple source skills in a scenario based environment. Students were given a clear purpose— a problem solving task related to issues in human biology, such as asthma, decorative contact lenses, or safe music volume levels. The overall task was presented in a Facebook like interface, which involved multiple sources including a feed, instant messaging, and emailing. Students were given 45 minutes to use a simulated search engine, "Gloogle", to locate, evaluate, synthesize, and communicate (LESC) information from multiple sources in order to complete the overall task. Students were given smaller scenario related tasks throughout the test that helped lead to the final response. In the ORCA-MC forms, students were given a similar scenario, except instead of having free-range across sources, they were guided through several key stopping points that were aimed to also utilize students' LESC skills. The scenarios presented in both ORCA-Closed and ORCA-MC required students to examine multiple sources from the "Gloogle" pseudo search engine. The scenarios called upon several multiple source skills such as purposeful reading, searching sources, evaluating sources, and perspective taking.

General Properties of ORCA

The ORCA-Closed and ORCA-MC demonstrate adequate reliability and validity, although ORCA-Closed had slightly higher reliability (Leu et al., 2014). The results from the forms indicate that offline and online reading comprehension skills each contributed to performance on reading tasks, which, following the theory of new literacies, suggests online reading comprehension involves skills beyond offline reading comprehension (Coiro, 2011).

However, it was also found that the majority of students are not skilled in online reading comprehension, especially the ability to critically evaluate the information, which was the most difficult for students compared to the three other online reading skills assessed (Forzani & Maykel, 2013). One might infer that the multiple source evaluation and integration skills required in the ORCA tasks contributed to the challenges students had with the online reading skills, though that was not the authors' research focus.

Applications of ORCA

The primary design elements of the ORCA designs have been incorporated into the ePIRLS design (Mullis & Martin, 2015). The Progress in International Reading Literacy Study (PIRLS) is conducted on a regular five-year cycle on a population of children in their fourth year of formal schooling. The ePIRLS was designed as an extension to the traditional paper-booklet PIRLS reading measures to assess reading in an online environment. The ePIRLS assessment consists of four school-based online reading tasks, each involving two to three different websites, with any student completing two of the four tasks in a 40 minute session. Thus, an SBA approach to multiple sources measurement as represented in an online computer environment construct is being operationally implemented in the ePIRLS program.

The Educational Testing Service (ETS) CBAL Initiative

Another SBA approach that has an extensive history of development is called, CBAL™, or Cognitively Based Assessment of, for, and as Learning. CBAL is a research initiative funded by ETS, aimed at creating "a model for an innovative K-12 assessment system that documents what students have achieved, facilitates instructional planning, and is considered by students and teachers to be a worthwhile educational experience in and of itself " (Bennett, 2010, p. 70).

The aforementioned GISA SBAs owe their origins to research commenced as part of the CBAL initiative. Despite similarities between GISA and CBAL assessments, CBAL assessments place more emphasis on using assessments to facilitate learning. CBAL assessments not only serve as a documentation of what students have learned (of learning), but also help teachers with instructional planning (for learning), and provides a model for students to follow when learning a content area (as learning). While the ability to use multiple sources is one assessment target in GISA, multiple sources of information serve as a learning opportunity for students to learn the content area. This is because in reality, no content area can be learned with a single source of information. In other words, multiple source is an inherent feature of learning.

In CBAL assessments, materials from multiple sources are organized by cognitive models, including competency models and associated learning progressions, which are developed from cognitive and learning sciences research in English Language Arts (Deane, Sabatini, & O'Reilly, 2012), mathematics, and science domains. During CBAL assessments, students participate in extended scenario-based tasks that are created by modeling high quality teaching practices that have been shown to improve classroom learning. These assessment scenarios not only help students learn the content while they go through the assessment, but also set up good examples for teachers to make their own instructional plans.

By using materials from multiple sources, CBAL assessments also have advantages over traditional assessments in terms of the intended consequences of testing. Under the pressure of traditional high-stakes testing, teachers may focus their instructions on the test content. Although this may improve student performance on the test itself, it does not generalize to the content domain. This problem is partially mitigated in CBAL assessments because the

assessments are based on domain-specific competency models (e.g., Liu, Rogat, & Bertling, 2013; O'Reilly & Sheehan, 2009). The competency models include key practices, strategies, and habits of mind, and CBAL assessments are developed to represent these processes with scenarios that students may experience in real life. By using scenarios that involves multiple sources of information, the CBAL assessments promote learning gains that can potentially generalize to the broader content domain.

Multiple source materials are prepared for CBAL assessments through key practices of the related content area. A key practice is a class of activities in which students use their skills to carry out complex tasks within a purposeful social context (Deane et al., 2015). The CBAL ELA key practices includes 1) basic literacy skills such as reading, writing, listening, and speaking, 2) model building skills such as building and sharing knowledge (O'Reilly, Deane, & Sabatini, 2015), and 3) applied practices such as conducting inquiry and research (Sparks & Deane, 2015) and discuss and debate ideas (Deane & Song, 2015). Participating in key practices allows students to gain competence in a domain (Deane et al., 2015). The concept of a key practice originates from the social constructivist view of learning (Vygotsky, 1978), which emphasizes the importance of social interaction in cognitive development. Thus, key practices often involve interaction with peers and teachers. Following this reasoning, multiple source materials are developed by considering the activities of key practices.

In short, the ETS CBAL initiative is aimed at creating the next generation assessments that not only measure student learning, but also facilitate it (Bennett, 2010). A typical CBAL SBA 1) provides a realistic purpose, 2) sequences tasks to follow hypothesized learning progressions of a domain and thus provide support for student performance, and 3) includes texts and information coming from multiple sources. CBAL assessments use scenarios to organize

multiple assessment materials that reflect key practices in related content areas. Below we provide an example to illustrate these features.

CBAL Example: SBA of Argumentation

In CBAL SBAs that target argumentation skills (Bennett, Deane, & van Rijn, 2016), students are asked to write on controversial issues such as whether advertising to children under age 12 should be banned in the U.S. Following the learning progression of argumentation proposed by Deane and Song (2015), the assessment target four elements: 1) understand the issue, 2) consider positions, 3) create and evaluate arguments, 4) organize and present arguments. First, students are presented with several source materials related to the topic and are asked to summarize the materials in preparation for using these documents in writing an argumentative essay (Gil et al., 2010). Following the summarization task, students work on an argument classification task, which requires them to classify people's positions based on the reasons/evidence they provide. The third task is an evidence classification task, which asks students to determine whether a piece of evidence supports or weakens a claim. This task also evaluates students' ability to create and evaluate arguments. The fourth task requires students to write an argument essay. The task provides students access to the source documents reviewed in earlier tasks, along with a writer's checklist to help them write the essay. This task addresses students' ability to understand the issue, to create and evaluate arguments, and to organize and present arguments. Finally, a fifth task requires students to write a few sentences to identify logical flaws in example arguments on the issue.

Across the portfolio of CBAL ELA SBA forms, the use of multiple sources is a staple design element. Students are typically provided with several document sources on the scenario topic. Initially, each source is accompanied by tasks probing specific learning progressions that

target understanding of the individual text. This may be in the form of writing a summary, answering questions about key content, or evaluating claims and evidence. As the SBA progresses (typically in a second, 40-minute session), the multiple sources become foundational and are made available in completing an extended constructed response essay task such as writing an argument, framing a proposal, or creating a research synthesis.

Applications of CBAL

Several of the primary elements of the CBAL ELA designs have been incorporated into NAEP reading blocks. The National Assessment of Educational Progress (NAEP), sometimes referred to as the nation's report card, is conducted on a regular cycle with a United States nationally representative sample of 4th, 8th, and 12th grade students. Scenario-based tasks derived from principles and prototypes of CBAL ELA forms have been adapted for subsequent implementation in NAEP, which uses tablet delivery of assessments for the first time in 2018. Thus, an SBA approach to multiple sources measurement as represented in CBAL style tasks, consistent with the NAEP reading framework, is being operationally implemented in the NAEP assessment program.

Implications for Research and Practice

While still in its infancy as a research program, we see several key lines of research and application with respect to multiple source assessment, scenario-based assessment approaches, and their intersection. First, continued theoretical research is warranted to clarify what is common versus unique in processing of multiple sources in comparison to single source analogs, with careful attention to how individual differences in ability and other characteristics may interact with performance. Assessments are best when targeted (with respect to construct) and efficient (with respect to time, cost, and effort). Multiple source assessment is likely to impact

efficiency, so it is helpful to know when or whether inferences can be made from single to multiple source tasks, and vice versa. Also, within the multiple source construct, which elements are most valuable as targets of assessment and are there contingent relations among elements? Ideally, learning and instructional approaches to enhancing multiple source processing will be documented, as these serve as models for assessment scenarios.

Future research is, of course, needed to clarify which features or techniques of SBAs are necessary or effective in achieving their intended goals of enhancing construct relevant processing, versus those that are ineffective or sources of construct irrelevant variance. SBA techniques and research is in its infancy – and agreed upon definitions or descriptions of SBA elements are still emerging. In order for SBAs to be used at scale, research must demonstrate practicality, utility, and efficiency, along with psychometric reliability and efficiency (Haertel, 1999).

Future applications of SBAs that are designed to model and inform instructional programs (e.g., GISA and CBAL) need to be evaluated to see whether they foster intended consequences, that is, facilitating effective instructional approaches to multiple source use. One promise of SBAs is that they provide models of instruction and application of reading skills. But a history of teaching to the test has mostly yielded unintended, negative consequences to instruction and learning (Au, 2007; Jones & Egley, 2004), so there is much research to be done to reverse this trend.

Summary

In this chapter, we have reviewed the application of multiple source constructs in scenario-based assessment approaches. We referenced the MD-TRACE model as a basis for analyzing how scenario based assessment designs encourage multiple sources processing. We

also reviewed three prominent research programs that are developing and evaluating SBAs – GISA, ORCA, and CBAL. We noted how each of these research programs has been influential in impacting the test approaches of national and international testing programs. We hope that this chapter encourages other research and test development programs to experiment with SBA design, theory, and research.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *NCME: Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Au, W. (2007). High-stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36,* 258-267.

Bennett, R. E. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement, 8,* 70–91.

Bennett, R., Deane, P., & van Rijn, P. (2016). From cognitive-domain theory to assessment practice. *Educational Psychologist, 51*, 82–107.

College Board (2017). The AP United States History Exam. Retrieved from: http://apcentral.collegeboard.com/apc/members/exam/exam_information/2089.html

Coiro, J. (2011). Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research, 43*, 352–392.

Coiro, J., & Dobler, E. (2007). Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the Internet. *Reading Research Quarterly, 42*, 214–257.

Council for Aid to Education (2017). K-12 Faculty or Administrator CWRA+ Sample Instrument. Retrieved from: http://cae.org/education-professionals/k12-faculty-or-administrator/cwra-sample-instrument/

Deane, P., Sabatini, J., Feng, G., Sparks, J.R., Song, Y., Fowles, M., . . . Foley, C. (2015). *Key practices in the English language arts (ELA): Linking learning theory, assessment, and instruction* (Research Report No. RR-15-17). Princeton, NJ: Educational Testing Service.

Deane, P., Sabatini, J., & O'Reilly, T. (2012). English language arts literacy framework. Princeton, NJ: Educational Testing Service. Retrieved from http://elalp.cbalwiki.ets.org/Table+of+Contents

Deane, P., & Song, Y. (2015). *The key practice, discuss and debate ideas: conceptual framework, literature review, and provisional learning progressions for argumentation* (Research Report No. RR-15-33). Princeton, NJ: Educational Testing Service.

Forzani, E., & Maykel, C. (2013). Evaluating a representative state sample of Connecticut seventh-grade students' ability to critically evaluate online information. *CARR Reader, 10*, 23–37.

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H. (2010). Summary versus argument tasks when working with multiple documents: Which is better for whom? *Contemporary Educational Psychology*, *35*, 157-173.

Haertel, E. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*, 5–9.

Hift, R. (2014). Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? *BMC medical education, 14*, 249.

Jones, B. D., & Egley, R. J. (2004). Voices from the frontlines:Teachers' perceptions of high-stakes testing, *Education Policy Analysis Archives, 12*(39), 1-34.

Kafer, K. (2002, December 1). High-poverty students excel with direct instruction. *Heartlander Magazine.* Retrieved from https://www.heartland.org/news-opinion/news/high-poverty-students-excel-with-direct-instruction

Katz, I. (2007). Testing information literacy in digital environments: ETS's iSkills assessment. *Information Technology and Libraries, 26,* 3–12.

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

Kirsch, I., & Mosenthal, P. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25*(1), 5–30.

Leu, D., Kinzer, C., Coiro, J., & Cammack, D. (2004). Toward a theory of new literacies emerging from the Internet and other information and communication technologies. In Ruddell, R. B., & Unrau, N. (Eds.), *Theoretical models and processes of reading* (pp. 1570–1613). Newark, DE : International Reading Association.

Leu, D., Kulikowich, J., Sedransk, N., Coiro, J., Liu, C., Cui, W., . . . Maykel, C. (2014, April). *The ORCA project: Designing technology-based assessments for online research, comprehension, and communication.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Liu, L., Rogat, A., & Bertling, M. (2013). *A CBAL™ Science Model of Cognition: Developing a Competency Model and Learning Progressions to Support Assessment Development*. (Research Report No. RR-13-54). Princeton, NJ: Educational Testing Services.

Lukhele, R., Thissen, D., & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement, 31,* 234–250.

Magliano, J. P., McCrudden, M. T., Rouet, J. F., & Sabatini, J. (in press). The modern reader: Should changes to how we read affect research and theory? In M. F. Schober, M. A. Britt & D. N. Rapp (Eds.), *Handbook of Discourse Processes* (2nd ed.). New York: Routledge.

Mosenthal, P., & Kirsch, I. (1991). Toward an explanatory model of document literacy. *Discourse Processes, 14*, 147–180.

Mullis, I., & Martin, M. (Eds.). (2015). *TIMSS & PIRLS International Study Center*. Chestnut Hill, MA: Boston College.

Organisation for Economic Co-operation and Development (2011). *PISA 2009 results: Students online: Digital technologies and performance*. Retrieved from https://www.oecd.org/pisa/pisaproducts/48270093.pdf

O'Reilly, T., Deane, P., & Sabatini, J. (2015). *Building and sharing knowledge key practice: What do you know, what don't you know, what did you learn?* (Research Report No. RR-15-24). Princeton, NJ: Educational Testing Service.

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*. (Research Report No. RR-09-43). Princeton, NJ: Educational Testing Service.

O'Reilly, T., Weeks, J., Sabatini, J., Halderman, L., & Steinberg, J. (2014). Designing reading comprehension assessments for reading interventions: how a theoretically motivated assessment can serve as an outcome measure. *Educational Psychology Review, 26*, 403–424.

Rouet, J. F. (2006). *The skills of document use: From text comprehension to web-based learning*. Mahweh, NJ: Erlbaum.

Rouet, J. F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. J. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Charlotte, NC: Information Age Pub.

Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*, 441–474.

Sabatini, J., Halderman, L., O'Reilly, T., & Weeks, J. (2016). Assessing comprehension in kindergarten through third grade. *Topics in Language Disorders, 36*, 334–355.

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.

Sabatini, J., O'Reilly, T., Halderman, L., & Bruce, K. (2014a). Broadening the scope of reading comprehension using scenario-based assessments: Preliminary findings and challenges. *L'Année psychologique, 114*, 693–723.

Sabatini, J., O'Reilly, T., Halderman, L., & Bruce, K. (2014b). Integrating scenario-based and component reading skill measures to understand the reading behavior of struggling readers. *Learning Disabilities Research & Practice, 29*, 36–43.

Sabatini, J., Petscher, Y., O'Reilly, T., & Truckenmiller, A. (2015). Improving comprehension assessment for middle and high school students: Challenges and opportunities. In K. Santi & D. Reed (Eds.), *Improving reading comprehension for middle and high school students* (pp. 119–152). Baltimore, MD: Springer Literacy Series.

Sparks, J. R., & Deane, P. (2015). *Cognitively based assessment of research and inquiry skills: Defining a key practice in the English language arts* (Research Report No. RR-15-35). Princeton, NJ: Educational Testing Service.

Vygotsky, L. (1978). *Mind in society: The development of higher mental process.* Cambridge, MA: Harvard University Press.

Table 1. Overview of the structure of the GISA Community Gardens assessment including the key stages of the assessment, their basic function, and the relationship to the MD-TRACE model.

GISA Section number	GISA Section name	Key function of the GISA Section	Rough correspondence to Relevant MD-TRACE stage
1	Introduction	To provide the goal and sub goals of the assessment; provide an overview of the task product; and to introduce the simulated peers and teacher	Create and update a task model
2	Background knowledge assessment	To measure students' knowledge on the topic of the sources.	Internal resources are moderated by prior knowledge, reading/search skills, and self-regulation
3	Text 1 The controversy	To build up students' knowledge of the topic and to assess basic understanding; to model self-regulation through peer dialogue.	Process single text contents; build prior knowledge.
4	Web links	To assess students ability to evaluate the relevance of source material	Assess information needs Assess item relevance
5	Text 2 Community gardens	To deepen students understanding of the issues and to assess learning; Multiple source comprehension	Process text contents Update task model Create/update a documents model
6	Text 3 Opposing view	To present a counterargument; Multiple source comprehension	Process text contents Update documents model
7	Produce a flyer	Culminating task that communicates a position and supports it with evidence.	Create/update a task product