Summer 7-2014

Developing a Measure of Scientific Literacy for Middle School Students

Helenrose Fives

Mark Nicolich

Amanda Birnbaum

Wendy Huber

Developing a Measure of Scientific Literacy for Middle School Students

Article *in* Science Education · July 2014

Impact Factor: 2.92 · DOI: 10.1002/sce.21115

4 authors, including:

Helenrose Fives
Montclair State University
31 PUBLICATIONS **420** CITATIONS
SEE PROFILE

Mark Nicolich
Cogimet
82 PUBLICATIONS **1,504** CITATIONS
SEE PROFILE

Science
Education

Developing a Measure of Scientific Literacy for Middle School Students

HELENROSE FIVES,[1] WENDY HUEBNER,[2] AMANDA S. BIRNBAUM,[2]
MARK NICOLICH[3]
[1]*Department of Educational Foundations and* [3]*Department of Health and Nutrition Sciences, Montclair State University, Montclair, NJ 07043, USA;* [2]*Cogimet, Lambertville, NJ 08530, USA*

Received 15 July 2013; accepted 10 March 2014
DOI 10.1002/sce.21115
Published online 18 June 2014 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: Scientific literacy reflects "a broad and functional understanding of science for general education purposes" (DeBoer, 2000, p. 594). Herein, we present the ongoing development of the Scientific Literacy Assessment (SLA), a work-in-progress measure to assess middle school students' (ages 11–14) scientific literacy. The SLA includes a selected response measure of students' *demonstrated* scientific literacy (SLA-D) and a motivation and beliefs scale based on existing measures of self-efficacy, subjective task value, and personal epistemology for science (SLA-MB). Our theoretical conceptualization of scientific literacy guided the development of our measure. We provide details from three studies: Pilot Study 1 ($n = 124$) and Pilot Study 2 ($n = 220$) describe the development of the SLA-D by conducting iterative item analyses of the student responses, think-aloud interviews with six students, and external expert feedback on the items in the SLA-D. Study 3 describes the testing of our prototype measure ($n = 264$). We present a validity argument including reliability evidence that supports the use of the current version of the SLA to provide evaluation of middle school students' scientific literacy. Our resulting SLA includes the SLA-D in two versions, each with 26 items and the SLA-MB with 25 items across three scales: value of science, scientific literacy self-efficacy, and personal epistemology. *Sci Ed* **98:**549–580, 2014

Correspondence to: Helenrose Fives; e-mail: fivesh@mail.montclair.edu
Contract grant sponsor: Science Education Partnership Award (SEPA), supported by the National Center for Research Resources.
Contract grant sponsor: Division of Program Coordination, Planning, and Strategic Initiatives of the National Institutes of Health.
Contract grant number: 8R25 OD011117-05.
Supporting Information is available in the online issue at wileyonlinelibrary.com.

DEVELOPING A MEASURE OF SCIENTIFIC LITERACY FOR MIDDLE SCHOOL STUDENTS

> At its core, scientific inquiry is the same in all fields. Scientific research, whether in education, physics, anthropology, molecular biology, or economics, is a continual process of rigorous reasoning supported by a dynamic interplay among methods, theories, and findings. It builds understanding in the form of models or theories that can be tested. (Scientific Research in Education; National Research Council [NRC], 2002, p. 2)

Scientific literacy is the ability to understand scientific processes and to engage meaningfully with scientific information available in daily life. Meaningful learning is understood as the connection of new information with prior knowledge in personally relevant ways (e.g., Aikenhead, 2011; Ausubel, 1977; Berry, Loughran, & Mulhall, 2007). Thus, we see scientific literacy as "a broad and functional understanding of science for general education purposes and not preparation for specific scientific and technical careers"; this functionality refers to the ability to use science to "live more effectively with respect to the natural world" (DeBoer, 2000, p. 594). This definition draws on perspectives from multiple sources including research and policy documents (NRC, 1996, 2012; Organisation for Economic Co-operation and Development [OECD], 2007) and science education researchers (e.g., Bybee, 2008; DeBoer, 2000; Laugksch, 2000; Roberts, 2007). There is no single accepted definition of scientific literacy; rather, the many characterizations of scientific literacy discussed in the literature include varying elements of competencies in science inquiry, content knowledge, and attitudes toward science (e.g., DeBoer, 2000; Roberts, 2007). Trends in science education policy have emphasized the importance of scientific literacy as a transferable outcome of science education. Several measures of scientific literacy currently exist (e.g., Bybee, 2008; OECD, 2006; Wenning, 2006, 2007). However, none of these target middle school students (aged 11–14 years, Grades 6–8 in the United States). Furthermore, most current measures draw on some degree of complex knowledge of one or more specific science field/disciplines and most measures do not include assessment of attitudes toward science. The purpose of our investigation was to develop a measure to assess the scientific literacy of middle school students that included assessments of the ability to think scientifically and students' motivation and beliefs toward science while being as field/discipline general as possible.

In designing this measure of scientific literacy, we sought to achieve a degree of science field (e.g., life, physical)/discipline (e.g., biology, astronomy) generality. That is, we attempted to measure the aspects of scientific literacy identified in our framework (described below) in ways that did not rely on field/discipline scientific knowledge (e.g., photosynthesis, simple machines, atomic structure); rather, we focused on the processes of science that span specific fields/disciplines. In part, this decision is based on the recognition that middle school students may not have a common depth of field/discipline knowledge from which to draw; thus, our goal was an instrument that can be used broadly to make valid inferences and evaluations by educators and educational researchers.

SCIENTIFIC LITERACY: THE CONCERN FOR SCIENCE EDUCATION

Educators and policy makers have made repeated calls for improved K-12 science education and defined performance expectations to reinforce the need for science as inquiry to improve scientific literacy (American Association for Advancement of Science [AAAS], 1993; National Assessment Governing Board [NAGB], 2010; NRC, 1996, 2012; OECD, 2007). These expectations are based on the premise that science is a recursive, dynamic process of asking questions, investigating, and then asking more questions, and that

these approaches can better engage children, who are naturally curious and learn through experience.

Most recently, in the United States, the National Research Council's report entitled *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* stated that hands-on investigative science is crucial in that it "gives [students] an appreciation of the wide range of approaches that are used to investigate, model, and explain the world" (NRC, 2012, p. 42). Moreover, authentic experiences in science can begin at a young age; the NRC (1996) report claimed that "[s]cience literacy begins with attitudes and values established in the earliest years . . ." (p. 18) and ". . . attitudes and values established toward science in the early years will shape a person's development of scientific literacy as an adult" (p. 22). This requires that education provide learners with a science curriculum that will facilitate the development of scientific literacy. We posit that given the goal of improved scientific literacy among students, we need to fill a gap in our ability to measure such scientific literacy.

Our first step was to develop an up-to-date conceptualization of the nature of scientific literacy reflective of extant theory in the field (described in the next section). This conceptualization framed the development of our measure. Findings from our two rounds of iterative testing of the measure informed our final conceptualization of scientific literacy. The pilot testing and final prototype testing are described later in the manuscript followed by a presentation of our findings and recommendations for use.

THEORETICAL REVIEW: CONCEPTUALIZATION OF SCIENTIFIC LITERACY

We engaged in a systematic review of the literature on scientific literacy that focused on how scientific literacy is defined, the various components that have been identified, and previous measures used to assess it. Reviews by Laugksch (2000), DeBoer (2000), Dillon (2009), Holbrook and Rannikmae (2009), and Roberts (2007) provided essential historical context for understanding the development of the concept of scientific literacy over the last 50 years. We also reviewed policy documents from leading science education agencies (e.g., AAAS, 1993; National Science Teachers Association [NSTA], 1991; NRC, 1996, 2012; OECD, 2007) to identify core capabilities that are considered essential to scientific literacy. Each definition or capability list was broken down into specific capabilities and compared across documents to identify components for assessment. We initially generated 12 components of scientific literacy to assess. We independently engaged in a theoretical analysis of these components to synthesize and better reflect the field. We then compared and discussed the individual syntheses and grouped similar components together, resulting in a total of six components comprising our initial framework.

While our perspective on scientific literacy is informed by the extant literature (described below), we focus more on the processes of science that span specific fields of study and disciplines. While some scholars contend that an information-rich knowledge of science is necessary for true scientific literacy (e.g., Shamos, 1995), others emphasize scientific literacy as active participation in the sociocultural potential and consequences of science (e.g., Cross & Price, 1992) that could lead to social activism (Hodson, 1999), and still others define scientific literacy as the "ability to deal with science in the news" (Hazen & Trefil, 1991, p. xii). The perspective we take attempts to find a middle ground that recognizes scientific literacy as knowledge of the nature of the field and its processes so that one can engage (in whatever form that takes for the individual) with science pragmatically and meaningfully in daily life. Our framework for conceptualizing scientific literacy is presented in Table 1 with a summary of supporting references. This initial framework

TABLE 1
Summary of Initial Scientific Literacy Construct and Components

Components	Supporting Literature												
	Showalter (1974)	Shen (1975)	Arons (1983)	Miller (1983)	AAAS (1993)	Hazen and Trefil (1991)	NSTA (1991)	NRC (1996)	DeBoer (2000)	Duit and Treagust (2003)	OECD (2007)	Holbrook and Rannikmae (2009)	NAGB (2010)
Role of science: Identify questions that can be answered through scientific investigation; understand the nature of scientific endeavors; understand generic science concepts	✓	✓	✓	✓				✓		✓	✓		
Scientific thinking and doing: Describe natural phenomena; recognize patterns; identify study variables; ask critical questions about study design; reach/evaluate conclusions based on evidence	✓		✓		✓		✓	✓	✓	✓	✓		✓
Science and society: Apply scientific conclusions to daily life; identify scientific issues underlying policy decisions; understand the role of science in decision making	✓	✓		✓	✓	✓	✓	✓	✓		✓		
Science media literacy: Develop questions to assess the validity of scientific reports; question the sources of science reporting[a]								✓	✓				
Mathematics in science: Use mathematics in science; understand the application of mathematics in science					✓								
Science motivation and beliefs: Value of science; self-efficacy for scientific literacy; personal epistemology of science	✓	✓	✓				✓		✓	✓	✓	✓	

[a]After our pilot studies, we subsumed science media literacy into the science and society component.

included six components that together reflect our perspective on the nature of scientific literacy: role of science, scientific thinking and doing, science and society, science media literacy, mathematics in science, and science motivation and beliefs. In the sections that follow, we briefly describe each of these components and how each is reflective of scientific literacy.

Role of Science

The first component in our framework, role of science, reflects the way that science can function in terms of understanding (a) the kinds of questions that can be answered through science, (b) the nature of scientific activities, and (c) generic scientific concepts present across field/discipline areas (e.g., variables, experiment, observation, etc.). As indicated in Table 1, seven of the 13 resource frameworks of scientific literacy included the ability to identify scientific questions as part of their conceptualization (i.e., Arons, 1983; Duit & Treagust, 2003; Miller, 1983; NRC, 1996; OECD, 2007; Shen, 1975; Showalter, 1974). The ability to recognize scientifically investigable questions (Duit & Treagust, 2003) provides some access to individuals' understanding of the nature of science, scientific methods, and what counts as evidence in science. A scientifically literate person, at the very least, must be able to determine whether and how science can be used to address questions in daily life. For instance, when shopping for a car, the scientifically literate person can determine what details provided by the salesperson are scientifically verifiable (e.g., fuel consumption, safety ratings) and which are issues of personal preference (e.g., interior color, prestige).

Scientific Thinking and Doing

While the first component emphasized the ability to recognize when science can be used to answer questions, this second component refers to actually *doing* the science needed to answer those questions. Thus, the scientifically literate are able to engage observational and analytical processes that are required for scientific thought. Scientific thinking includes the abilities to "describe, explain, and predict natural phenomenon" (NRC, 1996, p. 22), generate and evaluate scientific evidence (NRC, 1996), understand the difference between inference and observation (Arons, 1983), and identify patterns in data (NAGB, 2010). Thus, this component refers to the ability to design and conduct studies to address questions that can be answered by science.

In addition, this component includes the ability to question scientific methods, use evidence to support or refute arguments, and apply evidence-based conclusions (Duit & Treagust, 2003; NAGB, 2010; NRC, 1996; NSTA, 1991). The individual's ability to understand and apply scientific methods well enough to question and critique those methods when presented, to evaluate types of evidence offered in light of a research design, and to make conclusions about the findings presented are also included in this component. Furthermore, the abilities to apply what one knows about science (methods, theories, etc.) to new scientific endeavors and to evaluate those new endeavors through scientific reflection are also conceived to be part of scientific thinking.

Science and Society

The abilities to identify (a) scientific issues underlying local, national, and international policy and (b) science in decision-making processes are reflected in this component. We argue for a broad perspective on policy to include decisions made in the individual's home, workplace, or school, to larger contexts of town, state, national, and international arenas.

This component refers to the individual's ability to both recognize the role of science in decision making as well as the reasons for why science may not always be the deciding factor on policy decisions. The interaction between science and society was also noted by several other frameworks for scientific literacy (i.e., AAAS, 1993; DeBoer, 2000; Hazen & Trefil, 1991; Liu, 2009; Miller, 1983; NRC, 1996; NSTA, 1991; OECD, 2007; Shen, 1975; Shonwalter, 1974). Similar to our conceptualization, these other researchers described the need for people to develop a balanced perspective that integrates scientific thinking with social norms and ethical values (e.g., Miller, 1983; NSTA, 1991).

Science Media Literacy

Scientific media literacy refers to the individual's ability to critique scientific findings described or portrayed in the popular media and is closely related to the previous component in practice (Jarman & McClune, 2007). This includes the ability to develop questions to assess the validity of scientific reporting found in news reports or other media outlets and to question the sources of evidence provided for alternative goals or priorities. There is some precedence for science media literacy from other frameworks of scientific literacy that recognized this component as a unique aspect of scientific literacy (i.e., DeBoer, 2000; NRC, 1996). Furthermore, akin to the NRC (1996) *National Science Education Standards*, and the *Beyond 2000: Science Education for the Future* report from the United Kingdom (Millar & Osborne, 1998), we see a distinction between the ability to read and understand scientific reports and the ability to recognize the need to engage that scientific thinking when exposed to popular media. Similarly, DeBoer (2000) argued for the development of citizens who are able to "critically follow reports and discussion about science that appear in the media . . . " and recognize the direct role that science has in daily life (p. 592). This component recognizes the need for the individual to be able to activate scientific thinking when necessary and apply that thinking to information presented in the "normal" world of the person. Finally, as individuals move into adult life, the source of ongoing scientific literacy is often the news media and with this source comes embedded biases and persuasive tactics. Science media literacy seeks to facilitate the ongoing learning of science throughout adulthood with the ability to be a critical consumer of that information (Jarman & McClune, 2007; Zimmerman, Bisanz, Bisanz, Klein, & Klein, 2001). The current communications environment exposes individuals to constant information and arguments; being an informed citizen requires the ability to critically, quickly, and accurately ascertain the basis for arguments from among obviously scientific arguments about climate change to the more subversive use of science seen in advertisements for weight loss supplements and devices. Note that in our final conceptualization of scientific literacy we collapse science and society with science media literacy into one component.

Mathematics in Science

Our framework includes mathematics in science as a distinct component within scientific literacy. The inclusion of this component is supported by the AAAS (1989, 1993) and others in the field (e.g., Hamm, 1992; Yore, Pim, & Tuan, 2007). Domain-specific literacy may be described as *fundamental*, able to engage in domain-specific discourse, and *derived*, having an understanding of the content in the domain (Norris & Phillips, 2003). Yore and colleagues (2007) used this distinction to draw a parallel between literacy in the domains of mathematics and science and argued that literacy within either of these fields would require an interaction of both fundamental and derived literacies. The importance of literacy in both mathematics and science is underscored by the assessment of these separate literacies

by the Programme for International Student Assessment (PISA), which considers literacy to include the ability to apply knowledge across disciplines (OECD, 2003).

While we agree with the importance of mathematical literacy for reasons similar to the need for scientific literacy, in our framework, akin to that offered by AAAS (1989), we attempt to identify the kinds of mathematical understandings that are inherent to evaluating scientific findings. A working knowledge of mathematics as used in science (e.g., graph reading and the understanding of proportions and percentages) is necessary to fully understand science in everyday life; we consider this different from basic computation. The use of statistics and visual representations of numerical data has become commonplace in U.S. media. Everything from the representation of the number of search results as "O's" in Google to the forecasting of tomorrow's weather is reported through mathematics and visual representations of that mathematics. As such, an understanding of the mathematics that is used to communicate scientific findings and results is required to recognize or critically examine and understand science in media or understand issues of science in society.

Science Motivation and Beliefs

More than knowledge is needed to be a scientifically literate person; one must also have the motivation and beliefs necessary to engage that knowledge when needed as part of one's daily life. Therefore, we chose to include components related to students' motivation for and beliefs about science. Attitudes, values, and beliefs have been identified by others as components of scientific literacy (e.g., AAAS, 1989; Arons, 1983; Holbrook & Rannikmae, 2007; NRC, 1996; NSTA, 1991; OECD, 2007; Ryder, 2001; Shen, 1975). Despite the recommendation to address values and beliefs in conceptions of scientific literacy, little effort has been made to articulate just what those values and beliefs should be. To engage in an analysis of that nature was beyond the scope of our current work. However, we felt that to ignore this aspect of scientific literacy entirely would be disingenuous to a modern understanding of this concept. We were guided by Gauld's (1982) description of the "scientific attitude" as the motivation needed to convert knowledge and skills into scientific procedures and engagement. From this perspective, we reviewed some of the work on students' motivation and beliefs in science and selected three constructs relevant to the successful engagement of scientific literacy: value (subjective task value), confidence (self-efficacy), and beliefs about knowledge and knowing (personal epistemology). Therefore, we perceive a scientifically literate person as one who values science (intrinsically and for utility purposes: Wigfield & Eccles, 2000), feels capable of engaging in scientific activities (self-efficacy: Ketelhut, 2010), and believes that knowledge in science is developed by humans and is changing (personal epistemology: Conely, Pintrich, Vekiri, & Harrison, 2004).

Motivation researchers have examined the relations between the value students attribute to content area or achievement tasks and their engagement and achievement in school (e.g., Bøe, 2012; Eccles & Wigfield, 2002). Subjective task value is used to describe the value that learners have for academic tasks and has been described in four ways: *intrinsic value* refers to learners' experiences of "fun" or enjoyment for the task itself, *attainment value* refers to how important success on a task is for the learner's sense of self, *utility value* occurs when the learner sees the task as useful for some other goal, and the last area is *cost* that refers to what a learner must give up to engage in the task (Eccles, Barber, & Jozefowicz, 1999; Wigfield & Eccles, 2000). Task value is salient to scientific literacy. If a learner is to engage in his/her scientific literacy as part of daily life, then they must see some value for doing so, either because they enjoy it, they see themselves as the kind of person to thinks scientifically, or they see it as useful.

In the field of achievement motivation, self-efficacy beliefs are identified as beliefs held by an individual about his/her ability to organize and execute acts to bring about the desired outcome (Bandura, 1997). In other words, this refers to their perceived confidence in completing tasks. One definition of scientific literacy explicitly addressed the importance of feelings of self-efficacy in terms of confidence. Scientific literacy was described as

> [t]he capability to function with understanding and *confidence*, and at appropriate levels, in ways that bring about empowerment in the made world and in the world of scientific and technological ideas. (emphasis added, UNESCO, 1993, p. 15)

This definition and others indicated that it is not enough for students to be able to know about science or how to engage in science but that they must actually do so and feel confident about that capability, that is, they must have self-efficacy for science. Self-efficacy beliefs influence learners' choices, effort, and persistence, and routinely predict academic achievement (e.g., Britner & Pajares, 2001; Bryan, Glynn, & Kittleson, 2011; Lent, Brown, & Gore, 1997; Pajares & Valiante, 1997; Shell, Colvin, & Bruning, 1995). Thus, we chose to include the construct of self-efficacy for engaging in activities associated with scientific literacy.

Personal epistemology refers to individuals' domain-specific beliefs about knowledge and knowing (Hofer & Pintrich, 1997). Hofer's (2000) epistemological theories perspective suggests that beliefs about knowledge and knowing serve as interconnected theories that learners use as they engage with content and the world. Specifically, there are beliefs about "the nature of knowledge (what one believes knowledge is)" and beliefs about "the nature or process of knowing (how one comes to know)" (Hofer, 2000, p. 361). Within each of these frames, two dimensions of beliefs have been identified. Beliefs about the nature of knowledge have been described along two continua: certainty (knowledge is certain vs. knowledge is fluid) and simplicity (knowledge is made up of discrete separate units vs. knowledge is integrated and complex). Beliefs about knowing are described as beliefs about the source of knowledge (from authority or outside the person vs. constructed by individuals) and the justification of knowledge. Students' epistemological beliefs influence learning outcomes (e.g., Perkins, Jay, & Tishman, 1993; Songer and Linn, 1991), strategic processing, and reading comprehension (e.g., Braten, Stromoso, & Samuelson, 2008).

Limitations in Our Conceptualization of Scientific Literacy

We constrained our conceptualization of scientific literacy to what we felt could be tested in middle school students through paper and pencil measures. We appreciated calls from DeBoer (2000) and Holbrook and Rannikmae (2009) to maintain an open-ended and situation/culturally specific conceptualization of scientific literacy and we agree that the construct of scientific literacy includes fluid situation-specific applications of science in daily life. For example, Duit and Treagust (2003) argued for the inclusion of *collaboration* in a conception of scientific literacy, referring to individuals' abilities to interact with each other as they engage in scientific inquiry. However, we omitted such conceptions of scientific literacy from our framework, not because they are not valued components of this construct, but because we felt that these conceptions required more nuanced performance-based assessments to accurately assess individuals' abilities in these areas. Thus, we recognize that our conceptualization of scientific literacy is limited to those components we felt could be appropriately assessed through the type of measure we wanted to design.

MEASURING SCIENTIFIC LITERACY

Considering the importance of science literacy outcomes, it is surprising to discover the paucity of available measures that attempt to assess it. For example, in the United States, many state and national standardized tests for students attempt to measure scientific literacy, yet the scope of such tests (and the classroom instruction that precede them) is so broad that teachers engage in surface level coverage of a wide range of topics at the cost of allowing students time to focus deeply and learn a few central scientific concepts (Lambert, 2006). In a similar vein, it is recognized that most U.S. state and national testing programs do not address abilities in scientific inquiry (Fuchs, 2008). Several measures of scientific literacy currently exist (Bybee, 2008; Laugksch & Spargo, 1996; Liu, 2009; OECD, 2007; Wenning, 2007). However, existing measures have three key limitations in that they (1) tend to be field/discipline specific, (2) are intended for students at the secondary or university levels, and (3) ignore the assessment of students' motivation for and beliefs about science.

Field/Discipline Specificity

Items on existing measures of scientific literacy are largely information-dependent. By that we mean that learners must have sufficient scientific information (Jenkins, 2003) to respond accurately to test items. In contrast, scientific literacy should emphasize those aspects of science that transcend specific fields/disciplines, focus on the processes of science, and reflect scientific training (Jenkins, 2003). Science field/discipline-specific measures of scientific literacy are evidenced in the PISA science literacy measure that utilizes the *environment and natural resources* as appropriate context for measuring scientific literacy among 15-year-olds in 57 countries (Bybee, 2008). Similarly, the test by Wenning (2007) emphasized understanding of *physics*, with some of the items focused on general scientific thinking.

We agree that there is an important place for assessment of student understanding of specific science topics or information; however, we were interested in a measure of "their understanding of science as an approach" (NRC, 2012, p. 263). The NRC (2012) proposed a framework for K-12 science education and standards that defined eight science *practices* (e.g., asking scientific questions, engaging in argument from evidence) and seven *cross-cutting concepts* (e.g., pattern recognition, identifying cause and effect relations) that span fields/disciplines and are reflective of scientific literacy from our perspective (NRC, 2012). We agree with the premise of this recent framework for science education that "[a]lthough the practices used to develop scientific theories ... differ from one science domain to another, all sciences share certain common features at the core of their inquiry-based and problem-solving approaches" (NRC, 2012, p. 26).

In this way, we, perhaps, sidestep one common tension in the discourse around scientific literacy, the tension between content-focused and issues-focused science (DeBoer, 2000; Roberts, 2007). Roberts (2007) framed this tension as revealed in two "visions" of scientific literacy (p. 730). Vision I reflected a focus on the knowledge of science from within the discipline, and emphasized the importance of knowledge of scientific findings, principles, and laws as a basis for engagement with the field. In contrast, Vision II garnered its focus from the issues and experiences of daily life that hold within them a scientific component. Taking each of these perspectives to extreme outcomes suggests that in Vision I only expert scientists can ever become *truly* scientifically literate; it is only with vast knowledge of the content and process of the domain that one can converse and understand fully the meaning of scientific discourses (Shamos, 1995). Similarly, a Vision II extreme perspective may lead to a conception of scientific literacy as merely functional, an ability to engage

superficially with science and to understand when it is applicable to daily life (Shamos, 1995). To use a metaphor, Vision I literacy would be the equivalent of knowing a foreign language well enough to write, produce, create, appreciate, and consume literature in that language, whereas Vision II literacy would be equivalent to conversational language that would facilitate navigating the local areas, communicating to purchase goods, and finding directions.

In his seminal review of the literature on scientific literacy, Roberts (2007) stated "[t]here is no consensus about the meaning, or even the constituent parts, of SL [scientific literacy] – with one exception: everyone agrees that students can't be scientifically literate if they don't know any science subject matter" (p. 735). This conclusion highlights the inherent challenges in assessing scientific literacy, to which our work responds in two key ways. First, given the lack of consensus, we offer a definition and framework for scientific literacy that is operationalized by the measure we constructed and tested. We understand that this definition is limited and may be contested on theoretical grounds and sociopolitical goals. Thus, we offer *one* way to assess scientific literacy. Second, the content assessed in this test is intended to be relevant across the fields and disciplines of science. Our test is designed to assess middle school students' ability to recognize the underlying science processes and concerns at issue across a range of fields/disciplines. Success on this test should rest on the learners' understanding of scientific processes rather than recall of information from different disciplines of science.

We believe it is imperative to develop measures to assess whether or not principles of science, critical thinking, and problem solving are being effectively taught and learned. In our development of field/discipline-general items, we recognized that any particular item would have science content topics in it, and if the person responding to the item has prior knowledge of that topic, he/she will most likely perform better on that item (e.g., Alexander, Kulikowich, & Schultze, 1994). Thus, in our efforts to address this, we sought to vary the science topics across the items to emphasize everyday examples.

Middle School Level

Despite the work that has been done in the field to develop tools to measure scientific literacy, this work has not addressed the specific needs of middle school students. This age group requires a tool to assess their specific abilities and needs for three reasons. First, the middle school period marks the preparation for secondary education in the United States and as such there are frequently changes in how science is taught and by whom. It is usually marked by a change from science being taught by a classroom teacher who offers multiple subjects to a dedicated science teacher with content area expertise. A tool for assessing scientific literacy at this juncture of a student's academic career can provide meaningful information for classroom teachers as a possible formative assessment and researchers targeting science education to promote scientific literacy. Second, and in conjunction with the previous reason, science education starting at Grade 7 is a typical and entrenched academic subject worldwide (Holbrook & Rannikemae, 2007). Thus, we can, with some certainty, argue that formal instruction in science taught by science experts is offered starting in sixth grade (around age 11). The amount of variability in instruction, content, and expertise prior to Grade 7 is potentially very high. Therefore, targeting this tool for Grades 6–8 (11–14 years old) provides a good baseline for possible future development. Finally, students tend to report less interest or value for school subjects in general (e.g., Wigfield & Eccles, 2000) and science in particular (e.g., Osborne, Simon, & Collins, 2003) as they transition from elementary to secondary school. Targeting this measure for middle school allows for the assessment of the relation among knowledge, motivation, and beliefs

at this known transitional time, as well as possible predictors or facilitators of scientific literacy.

Motivation and Beliefs in Scientific Literacy Assessment

Existing measures of scientific literacy do not assess motivation and beliefs in science despite the theoretical call from scholars and organizations for these perspectives to be included in the conception of a person who is scientifically literate (Arons, 1983; DeBoer, 2000; Holbrook & Rannikmae, 2007; NSTA, 1991; Shen, 1975; Showalter, 1974). We identified task value, self-efficacy, and personal epistemology as salient motivation and belief constructs for including in a measure of scientific literacy. Together these constructs tap into the value individuals hold for science, their confidence to engage in science, and their belief in the nature of science knowledge.

AIMS OF THE INVESTIGATION

Our overarching aim was to develop and test the Scientific Literacy Assessment (SLA) measure that would allow researchers and educators to make valid inferences about middle school students' scientific literacy. To achieve this goal, we developed two sets of measures to be administered together in a single instrument. The SLA-D assesses *demonstrated* scientific literacy through a series of multiple-choice items that use everyday situations and examples, rather than field/discipline-specific scientific knowledge, to test scientific literacy through the examination of understandings of the role of science, scientific thinking and doing, science and society, science media literacy, and mathematics in science. The other component of the SLA, the SLA-MB, assesses *motivation and beliefs* associated with scientific literacy. The SLA-MB includes three adaptations of three previously developed Likert-type scales to assess students' motivations and beliefs in relation to science.

OVERVIEW OF METHODOLOGY: A MULTISTAGE APPROACH TO MEASURE DEVELOPMENT

Our approach to the design and development of this measure was informed by the unitary construct of validity advocated by the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999). The unitary construct of validity recognizes a variety of types of evidence that may support a validity argument but that "validity involves an overall evaluation of the plausibility of the intended interpretations," that is validity is a property of the inferences made and not the measure itself (Kane, 1994, p. 136). The *Standards* put forth five types of evidence to use in supporting a validity argument for the use of a measurement tool; these include evidence based on test content, response process, internal structure, relations to other variables, and consequences of testing (AERA, APA, NCME, 1999). As Kane (2012) recently wrote "[t]he kinds of evidence required for validation are determined by the claims being made . . . " (p. 3). We argue that the SLA will provide users with data that can be used to make valid inferences about individuals' demonstrated scientific literacy (SLA-D) and scientific literacy motivation and beliefs (SLA-MB). With these goals in mind, we emphasized in our measure development evidence based on *test content*, *response process*, and *internal structure.* Table 2 overviews our three-study validation process and the types of validity evidence we offer to support the use of this tool.

TABLE 2
Validation Process and Sources of Validity Evidence

Overview of Validation Process and Sources of Validity					
Measure Details			Sources of Validity Evidence		
SLA Measure	Item Type	Intended Inferences	Test Content	Response Process	Internal Structure
SLA-D Study 1	57 multiple-choice (MC) items	Higher scores on this measure indicate stronger demonstrated scientific literacy	• Alignment of items to definition scientific literacy • Readability statistics • Well written items, follow item writing guidelines • Construction and review of items by interdisciplinary test construction team • Review of items by SEPA colleagues and four science teachers in Study 1	Think aloud four students	• Item analysis • Discrimination index • Item-total correlations
SLA-D Study 2	53 MC items			Think aloud two students	
SLA-D Prototype Study 3 (two versions)	26 MC items			No think aloud	*Same as above and* • Kuder—Richardson 20 • Factor analysis
SLA-MB Prototype Study 3	25 Likert scale items	Higher scores on these subscales indicate motivation and beliefs reflective of a scientifically literate person	Alignment of items to definition of scientific literacy motivation and beliefs		• Cronbach's alpha • Factor analysis

Evidence Based on Test Content

To construct the SLA-D, we engaged in an iterative process of selected response item (i.e., multiple choice) item generation, evidence gathering, and redesign that fostered the development of the measure. Throughout our process, we relied on *evidence based on test content* as recommended in the *Standards* (AERA, APA, NCME, 1999), using the following strategies: (1) developing items specifically to align with each of the components identified in our framework; (2) discussion among our interdisciplinary research team composed of two epidemiologists (one who has extensive experience implementing inquiry science in K-12 education), a statistician, and an educational psychologist (who was a middle school science teacher for 6 years); (3) subjecting our initial measure to evaluation by experts in science education; and (4) crafting the items following the item writing recommendations of Haladyna, Downing, and Rodriquez (2002).

To develop a pool of candidate items, we reviewed existing measures to identify items to include or adapt, and also developed original (de novo) items. For the *de novo* items, each member of the research team initially drafted multiple items aligned to specific components of our framework of demonstrated scientific literacy. We wrote all items at or below the sixth-grade level according to the Flesch–Kincaid index provided by Microsoft Word. We did this because we did not want reading ability to confound our assessment of scientific literacy for this measure. We do recognize, however, that differences in reading ability and language fluency will impact individual student's scores on this, or any, measure of scientific literacy. Furthermore, the Flesch–Kincaid index does not take into account what the reader brings to the task of reading such as prior knowledge of the subject matter or interest in the topic and factors that can influence reading comprehension (e.g., Alexander et al., 1994; Baldwin, Peleg-Burckner, & McClintock, 1985).

All items were passed on in a "merry-go-round" fashion so each team member reviewed all items constructed, making changes as needed, and including new or revised items. As described below, we ran several empirical trials of the SLA-D; after each trial, we reviewed all items, distractors, and responses as a team, and engaged in collaborative discussion and review to revise, delete, or generate new items as needed. At the end of our testing processes, most of the final SLA-D items had been developed by our team (see Appendices 1 and 2 in the Supporting Information), with a total of six items derived (and used with permission) from AAAS Project 2061 (2011) and Dillashaw and Okey (1980).

For the assessment of scientific literacy motivation and beliefs (SLA-MB), we followed the same strategy of aligning our description of scientific literacy to items for inclusion in our measure described above. We examined the existing measures in the field and compared them to the definition of this component in our theoretical framework and identified three existing measures for use and adaptation. The measures were selected based on the theoretical basis of their development and congruence with the motivation and beliefs identified as salient for scientific literacy. We selected Wigfield and Eccles' (2000) measure of achievement value to assess each of the three achievement values for science: intrinsic value (fun), attainment value (importance), and utility value (usefulness). Kettlehut's (2010) measure of self-efficacy for scientific inquiry was adapted to measure students' perceptions of capability for engaging in activities reflective of scientific literacy. Ketelhut developed this tool from a sound theoretical base and engaged in measure development that provided ample validity evidence to support the use of this scale for the intended purpose. To assess students' beliefs about the source and certainty of knowledge in science that seemed most connected to issues of scientific literacy, such that they underscore the nature of science as an evolving domain with multiple responses to questions in the field, we used the two subscales from Conley and colleagues (2004).

Evidence Based on Response Process

In addition to validity evidence based on test content, we also gathered evidence for the SLA-D based on response processes by performing think-aloud interviews (e.g., Presser et al., 2004) with six middle school students (Pilot Studies 1 and 2).

Evidence Based on Internal Structure

We collected data from middle school students (Pilot Studies 1 and 2; Prototype Study 3) for evidence based on internal structure by examining responses to the SLA-D using statistical analysis of each item and of the overall measure. The scales selected for the SLA-MB had previously demonstrated evidence of internal structure and reliability. Wigfield and Eccles' (2000) task value measure has demonstrated sound reliability. Kettlehut (2010) tested her self-efficacy scale with 2,000 middle school students and reported a Cronbach's alpha of .86. The personal epistemology scales were used with fifth-grade students to assess their beliefs about science at two time periods and demonstrated acceptable reliability (i.e., the source of scientific knowledge: $\alpha = .81, .82$; the certainty of knowledge in science: $\alpha = .78, .79$; Conley et al., 2004). In Prototype Study 3, we evaluated the factor structure of the motivation and beliefs measure using principal components factor analysis and examined the reliability evidence for these scales. A brief overview of the three studies, research questions, procedures, and findings can be found in Table 3.

PILOT TESTING SLA-D: STUDIES 1 AND 2

Pilots: Participants and Procedures

Pilot testing of the SLA-D iterations (i.e., multiple-choice items) was conducted in two phases: Pilot Study 1 and Pilot Study 2. For both phases of pilot testing, the participants were seventh- and eighth-grade students (12–13 years old) from urban middle schools in the mid-Atlantic region of the Unite States. Study 1 included 124 participants (75 in seventh grade and 49 in eighth grade) from a single school. Study 2 included 220 participants (170 in seventh grade and 50 in eighth grade) from four other schools.

The SLA-D for Studies 1 and 2 contained 57 and 53 multiple-choice items, respectively. Assessments were completed during the students' scheduled science class periods (blocks); these classes ranged from 50 to 80 minutes in duration. All participants submitted parental consent and student assent forms. Study personnel introduced the study procedure to students, administered the forms, and monitored the testing. All consent procedures and interactions with study subjects were conducted with approval from Montclair State University's Institutional Research Board.

Pilots: Data Analysis

The following analyses were employed to establish validity evidence based on test content, response process, and internal structure. Four key analyses were used to evaluate each item and inform internal structure evidence: (a) percent correct, (b) frequency of responses to each option, (c) discrimination index, and (d) item-total correlation coefficients. We examined each item for the overall percentage of correct responses, expecting that, at most, half of participants would select the correct option for an item. Our participants were not receiving any special preparation in scientific literacy beyond their current science courses in schools. Furthermore, with the goal of developing a measure for use by researchers and educators to assess effects of instruction on scientific literacy, the measure needed to be

TABLE 3
Overview of Research Design

Study →	Pilot Study 1	Pilot Study 2	Prototype Study 3
Research questions	1. How do middle school students respond to items on this? 2. Are indicators of test difficulty (discrimination index, correlations) reasonable for all items? 3. How should the test be revised to better measure scientific literacy?		1. How do middle school students respond to items on this? 2. Are indicators of test difficulty reasonable for all items? 3. Should the test be revised to better measure scientific literacy? 4. How do participants respond to the attitude and belief items?
Participants	124 seventh- and eighth-grade students four students: think-aloud	220 seventh- and eighth-grade students two students: think-aloud	264 seventh- and eighth-grade students
Procedures	• SLA-D = 57 MC items • Think-aloud with four students • Expert review	• SLA-D = 53 MC items • Two variations • Think-aloud with two students	• Two versions SLA-D each composed of 26 MC items (11 shared items and 15 unique items) and SLA-MB composed of 25 Likert-type items from 1 to 5 • Two ordering variations for each version
Analyses	(a) Test score descriptive statics (b) Item discrimination index (*D*) (c) Response choice by item (d) Item-total correlation coefficients (*r*) (e) Thematic analysis	(a) *t* test to compare test versions (b) Analyses a-e from Study 1 (c) Kuder–Richardson-20 (d) Review and revision of SLA framework	(a) ANOVAs to compare test versions and schools (b) Kuder–Richardson-20 (c) Analyses a-e from Study 1 (d) Principal components analysis SLA-MB (e) Review and revision of SLA framework
Findings	• *D* ranged from 0.06 to 0.83, with 19 items demonstrating *Ds* below 0.30 • *r* ranged from .02 to .59 • 31 items were revised • Seven items were deleted • Three new items were created • Revised test of 53 multiple-choice items	• *t*-test revaled no difference between versions • *Ds* ranged from 0.05 to 0.82; eight items demonstrated *Ds* below 0.30 • *r* ranged from .11 to .59 • 24 items identified for revision • 12 items were deleted • Revised test of 41 multiple-choice items	• ANOVAs revealed no difference between versions • Two 26-item versions of final SLA-D • Kuder–Richardson-20 = 0.83 for SLA-D1 and 0.82 for SLA-D2 • *Ds* ranged from 0.30 to 0.85 • *r*s ranged from .13 to .62 • SLA-MB included three clear factors: Value of science ($\alpha = .80$); science literacy self-efficacy ($\alpha = .72$), and personal epistemology ($\alpha = .88$)

sensitive enough to pick up changes in learning. Thus, items with more than a 50% correct response were either revised to be made more difficult (in Study 1) or were dropped from the measure (in Study 2).

We also examined response frequencies to each item as selected by the participants in each quartile of the distribution of total scores. This method enabled us to revise several item options to make the distractors more attractive, and therefore increase item difficulty. We calculated the discrimination index (*D*) for each item as the proportion of top total scores (top quartile) who chose the correct response minus the proportion of bottom total scores (bottom quartile) who choose the correct response (Johnson, 1951). The discrimination index provides information as to how well each item discriminates participants from the top and bottom percentiles. Using Hopkins (1998) guidelines for evaluating items based on the *D* values, we considered items with a *D* of 0.40 as very strong, 0.30–0.40 as good, and items below 0.30 as needing work (Reynolds, Livingston, & Wilson, 2006).

We calculated the relationship of performance on individual items with the total test score based on the point biserial correlation coefficient. Shaw and Young (2004) offer four recommendations for item retention or deletion based on the item-to-total score correlation coefficients for classroom tests. Specifically they recommend to (1) delete, replace, or revise items with a negative correlation coefficient; (2) replace or rewrite items with zero or (3) low correlation coefficient (i.e., coefficients from .00 to .20, p. 20); and (4) consider using .30 as the "cut-off point for identifying items that may merit retention" indicating that correlations falling in the range of .20–0.30 " . . . are fairly good to quite good items. They could stand as written" (p. 20). It is important to note that these recommendations focus on improving the overall reliability of a teacher's classroom test used as a summative assessment to evaluate learning that had occurred. Thus, their goal in offering these recommendations was different from our perspective of developing a measure sensitive enough to assess changes in scientific literacy in response to instructional interventions. That is, in this initial development stage we were testing the measure with students who we expected would be relatively naïve to the content, if we followed all of the recommendations we would risk making the test too easy for students with preparation in this content. For that reason, we adhered to the first two recommendations but were more flexible in accepting or retaining items with item-to-total score correlation coefficients lower than .20.

We were also guided by qualitative think-aloud interviews during Studies 1 and 2 from four and two participating students, respectively, to gather response process validity evidence. The classroom teacher was asked to identify students for the think-aloud activity who would be likely to develop rapport and communicate openly with the researcher but who were not necessarily the most accomplished students. These students then completed the test with one of the researchers who prompted each student to "think out loud" while completing the test. The researchers took field notes on a copy of the test, recording comments the students made (including both cognitive and affective responses).

Think-Aloud Interview Insights

The think-aloud interviews provided insights into adolescents' cognitive as well as affective responses, contributing unique information to the range of evidence the research team considered in deciding to keep, revise, or reject tested items. A specific result of think-aloud feedback involves an item that was developed to assess students' application of scientific findings to everyday life. In an effort to provide unambiguous directions, the original stem was: "A family decided to make all their decisions based on the results of scientific studies. The adults want the children to get better grades in school and so they set a new rule—all the children must be in bed by 9 PM. Upon which study result was this rule

based?" The think-aloud participants in Study 2 were perplexed by the implausible premise that a family would make all decisions based on the results of scientific studies, although they were able to understand the question. Based on their feedback, we revised the stem to a more relatable scenario: "Arturo's parents want him to get better grades in school. His mother read a research study on the topic. After reading the study, she decided that from now on Arturo needed to be in bed by 9 PM. Which of these studies did Arturo's mother read?"

A related insight, based on observations of students' affective responses during the think-aloud interviews, was that they were energized by questions that they found to be inherently worthwhile or "fun" based on the topic or context. Consistent with findings of Drennan (2003), the qualitative information generated by the think-aloud interviews helped inform decisions about which items to eliminate, and bolstered the research team's understanding of students' stamina and willingness to work through complex questions.

Pilot Study 1 Results

The overall mean number correct score for this test was 32 of 57 (56%) with minimum and maximum scores of 14 and 51 (24% and 89%), and the distribution resembled a normal curve. The discrimination index (*D*) ranged from 0.06 to 0.83 with 38 items demonstrating *D*s of 0.30 and greater; this indicated how well the items discriminated low scorers from top performers on the test overall. The bi-serial correlations between individual items and the total score ranged from .02 to .59, with a median of 0.32. As noted above, Shaw and Young (2004) recommend revision or deletion of items with correlations coefficients less than .20.

We gathered additional evidence of validity of test content by sharing a copy of the Pilot Study 1 test with a group of 25 other science education researchers with current Science Education Partnership Awards (SEPA) from the National Institutes of Health. They reviewed and commented on items during a session at a SEPA Principal Investigators' annual meeting. A member of our research team gathered their feedback and reported it to our team.

Items with correlation coefficients lower than .20 and *Ds* lower than 0.30 were closely scrutinized and then deleted or revised based on a number of criteria including the distribution of responses to each distractor, the overall number of items aligned with the five target components of scientific literacy, feedback on items from SEPA colleagues (particularly regarding comments about cultural appropriateness), and responses from the think-aloud interviews. Through detailed discussions of the above data for each item, we decided to revise 31 items, delete seven items, and create three new items for the next pilot test (Study 2).

Pilot Study 2 Results

We created two orderings of the 53-item SLA-D to address any possible order or fatigue effects on students' responses to the items. A series of analyses of variances (ANOVAs) indicated that the ordering did not lead to significant differences for the total score; thus, for the following analyses, responses from both versions were considered together.

Based on the 220 observations, the Kuder–Richardson equation 20 reliability for the 53 items was 0.90, suggesting that responses to the multiple-choice items were collectively reliable. Because responses were scored on a binary scale (i.e., correct or incorrect), we used the Kuder–Richardson reliability equation as equivalent to the Cronbach alpha. Although we considered the items on the SLA-D to reflect a single construct of demonstrated scientific

literacy, we were also interested in how items related to each of the five components of our framework (see below for an explanation of our framework reduction from six to five components). The reliability for the individual components was lower, ranging from 0.46 to 0.76, with the magnitude of the coefficient approximately proportional to the square root of the number of items in the component. It is important to note that because dividing up items across the components allowed fewer items per component (which decreases reliability) and because the subjects were expected to be naïve to the material, the relatively low reliability scores on the individual components were expected. However, we interpret the substantial decline in reliability when examining the five individual components compared with the full complement of items as providing initial evidence regarding the internal structure of the SLA-D, suggesting that it is most appropriately viewed as a single construct.

We followed the same procedures for item review as in Pilot Study 1 to identify items for revision or deletion. As before, our decisions concerning item revisions and deletions were guided by our goals of aligning with our framework of scientific literacy as well as the statistical results (detailed below), and two additional think-aloud interviews. The overall mean number correct score for this test was 26.6 of 53 (50%), slightly lower than in Study 1, with a wider range of scores from 4 to 52 (8–98%). Overall, 31 of the 53 items were correctly answered by 50% or fewer students. Items to which more than 50% of participants responded correctly were considered "too easy" and were marked for either deletion or revision to ensure an adequate sampling of the construct of interest (AERA, APA, NCME, 1999).

The *Ds* ranged from 0.05 to 0.82, with only eight of the 53 items demonstrating *Ds* of 0.29 or lower (an improvement over Study 1). Biserial correlations between the item score and the total score ranged from .11 to .59. Based on our discussion and analysis of each item, we deleted 12 items and identified 24 items for revisions, primarily for simplification of stems and clarification of language. The resulting multiple-choice measure comprised 41 items.

In addition to clarifying items, the team also used the Study 2 results to review and adjust the overall framework for demonstrated scientific literacy underpinning the measure development. We did so by mapping backwards from constructed items to the initial components and considering the reliability analysis for each component. Based on this, we felt that combining what was initially conceived as two separate components ("science media literacy" and "science and society") into a single component provided an adequate and more parsimonious framework. Although as stated earlier, we believe that the SLA-D is most appropriately viewed as a single construct. In Table 4, we illustrate the alignment of sample items with each component of the revised conceptual framework.

PROTOTYPE TESTING

The goal of Prototype Study 3 was to evaluate our prototype measure and identify any final items for revision or deletion (see Table 3 for the research questions). Based on our findings from the pilot studies, in Study 3 we tested four variations of the SLA that included both the SLA-D and SLA-MB subscales (the latter was not tested in Studies 1 and 2).

Measures

SLA-D. The SLA-D included two versions (i.e., SLA-D1 and SLA-D2) each composed of 26 items. Eleven items are shared between the two versions and each version includes 15 unique items. The two versions of the SLA-D represent 41 unique items in all. The

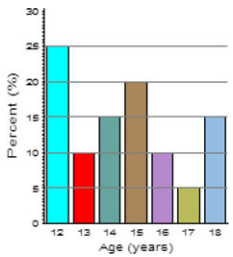
TABLE 4
Final Scientific Literacy Framework and Sample Items

Component	Sample Item
Role of science • Identify questions that can be answered through scientific investigation; • Understand the nature of scientific endeavors; • Understand generic science terms/concepts.	A country has a high number of decayed teeth (cavities) per person. Which question about tooth decay can only be answered with scientific experiments? (a) Do the men in this country have more tooth decay than women? (b) Would putting vitamin D in the water supply affect tooth decay? (c) Has the number of decayed teeth increased in the past 10 years? (d) Is tooth decay more common in some parts of the country than others?
Scientific thinking and doing *observational and analytical abilities* • Describe natural phenomena; • Recognize patterns; • Identify study variables; • Ask critical questions about study design; • Reach/evaluate conclusions based on evidence.	The principal of Riley middle school wants to remove candy and soda (pop) vending machines. In their place, she wants to put in healthy food machines. She wants to know what her students will think of these changes. What would be the best way to get an accurate answer to this question? (a) Give a survey to all students who play on sports teams. (b) Give a survey to all students who attend a health fair. (c) Give a survey to every 20th student on a list of all students. (d) Give a survey to all students who use the vending machines.
Science and society *Critique of scientific findings described in the popular media* • Apply scientific conclusions to daily life; • Understand the role of science in policy decision-making; • Develop questions to assess validity of scientific reports; • Question the sources of science reporting; • Identify scientific issues underlying policy decisions.	A student finds a website created by the "No Homework Committee." He wants to find out the reasons for and against assigning homework to students. Is this a trustworthy source of information? (a) Yes. This group is against homework and knows all of the arguments. (b) Yes. Information on Web sites is always balanced and correct. (c) No. This group might give more attention to arguments against homework. (d) No. This group is probably not very good at arguing for or against homework.

a

(Continued)

TABLE 4
Continued

Component	Sample Item
Mathematics in science • Use mathematics in science; • Understand the application of mathematics in science.	What percent of the sample of people shown in the graph is older than 15 years? (a) 20% (b) 30% (c) 40% (d) 50% 
Motivation and beliefs • Value of science (Wigfield & Eccles, 2000); • Self-efficacy for scientific literacy (adapted from Kettlehut, 2010); • Source and certainty of scientific knowledge (Conley et al., 2004).	Value: In general, I find working on science assignments (1: very boring to 5: very interesting) Self-efficacy: I can use science to make decisions about my daily life (1: strongly disagree to 5: strongly agree) Personal epistemology: All questions in science have one right answer (1: strongly disagree to 5: strongly agree)

SLA-D1 and SLA-D2 can be found in Appendices 1 and 2 in the Supporting Information, respectively.

SLA-MB. The SLA-MB is composed of three motivation and belief scales assessed on a 1–5 scale: value of science (six items), self-efficacy for scientific literacy (eight items), and personal epistemology of science (11 items—reverse coded; see Appendix 3 in the Supporting Information for these scales). Higher scores indicate stronger value, self-efficacy, and more sophisticated beliefs about science. We adapted Wigfield and Eccles (2000) measure of achievement value to assess students' value of science by replacing "math" with "science" in each of the six items. Scientific literacy self-efficacy was assessed with eight items. Four of these items came from Kettlehut's (2010) measure and we drafted four new items to better align with our conceptualization of scientific literacy. Specifically, we added items to assess student's confidence reflective of topics in the role of science, science and society, science media literacy, and mathematics in science. Beliefs about the source and certainty of knowledge in science were assessed with two scales from Conely et al.'s (2004) measure of personal epistemology in science. These items were worded such that low scores demonstrated more sophisticated epistemological beliefs, and as recommended by Conley et al. (2004), we reverse coded these items in our analyses to align the scores with the other two scales. The SLA-MB can be founded in Appendix 3 in the Supporting Information.

Variations. During the prototype testing in addition to testing two SLA-D versions, we also varied the *order* in which participants received the SLA-D version with the SLA-MB scales. This resulted in four variations of the SLA that were tested: (1) SLA-D1 → SLA-MB; (2) SLA-MB → SLA-D1; (3) SLA-D2 → SLA-MB; (4) SLA-MB → SLA-D2. We engaged in this variation to ameliorate any order effects in completing the assessments.

Participants and Procedures

Data collection was initiated with 321 middle school students in seventh or eighth grade from five schools in Northern New Jersey. Data collection took place in May and June 2012 during students' scheduled science classes. All participants submitted parental consent and student assent forms and all study activities were conducted with approval from Montclair State University's Institutional Research Board. The enrollment response rate was 26% overall, with a range by school of 17–45%, a likely reflection of the written parental consent required to participate. Female students were slightly overrepresented (56%) and the participants reported a range of ethnicities including Hispanic (34%), White (24%), Asian (21%), African American (14%), and other or multiple (7%).

Evaluation of the Prototype

Consistency Across Schools, Grade Levels, and Variations of SLA-D. We deleted responses from four students who had two or more missing responses on the SLA-D portion, leaving data from 317 participants for analysis. To determine if there were any significant differences related to the schools, grade levels, or variations of the SLA, we conducted a univariate ANOVA. School, grade, measure variation, and their two and three-factor interactions served as the independent variables, with the score on the SLA-D items as the dependent variable. Because of the higher order interactions, we were not able to test for the homogeneity of variance assumption, but the Shapiro—Wilk (1965) test for normality

TABLE 5
Description of Participants in Study 3

Characteristic	Number of Participants		
School by grade	Seventh	Eighth	All
1	0	38	38
2	0	133	133
3	41	0	41
4	20	32	52
Total	61	203	264
School by gender	Male	Female	All
1	13	25	38
2	65	68	133
3	17	24	41
4	25	27	52
Total	120	144	264
Ethnicity by gender	Male	Female	All
Asian	29	35	64
Black or African American	9	14	23
Hispanic or Latino	31	52	83
White	41	34	75
Other ethnicities (American Indian, Pacific Islander); or two or more ethnic backgrounds indicated; or missing	10	9	19
Total	120	144	264

of the residuals indicated the residuals were normally distributed and there were no outliers ($p < .23$), which gave us confidence that we had not violated either the homogeneity of variance or the normality of residuals assumptions. ANOVA results indicated a significant difference across schools [$F(4, 289) = 38.67$; $p < .0001$, $\eta^2 = 0.35$, where η^2 is the effect size] but no significant differences across test variations [$F(3, 289) = 2.01$; $p < .11$, $\eta^2 = 0.02$] or grade level [$F(1, 289) = 2.16$; $p < .14$, $\eta^2 = 0.007$], nor in the 2 two- and three-factor interactions. Follow-up analyses used Duncan's multiple range test (Duncan, 1975), a multiple comparison test that tests pairs of group means while considering the multiple comparisons problem associated with a test like the multiple *t*-test. These tests indicated that School 2 had statistically significantly higher scores than all others ($M = 17.1$, $SD = 4.9$), Schools 1 and 3 were similar to each other and different from all others (Ms = 14.8, 13.4, SDs = 5.0), School 4 ($M = 11.1$, $SD = 4.0$) and School 5 ($M = 7.7$, $SD = 2.5$) were different from all other schools and scored the lowest in the order presented. In addition, the students in School 5 were administered the measure on the last day of school. Based on anecdotal evidence from the researcher administering the SLA at this school and the combined teaching experience of other members of the research team, we determined that the extremely low scores from both grades in School 5 could well be due to this contextual variable. Therefore, we dropped all data from School 5 for all remaining analyses ($n = 54$, of which one had been dropped because of too many missing responses), leaving data from 264 participants for further analysis. Table 5 provides complete demographic information for the participants used in the remaining analyses.

SLA-D Evaluation. For the 264 responses from the four schools retained for the remainder of our investigation, the mean SLA-D was 15 correct of 26 (58%), with a range of 2–25 (8–96%). An ANOVA was performed in which school, grade, measure form, and their two- and three-factor interactions served as the independent variables and score on the multiple-choice items as the dependent variable. The Shapiro–Wilk (1965) test for normality of the residuals indicated that the residuals were normally distributed and there were no outliers ($p < .15$), which, again, gave us confidence that we had not violated either the homogeneity of variance or the normality of residuals assumptions. As in the previous analyses, the ANOVA results indicated a significant difference across schools [$F(3, 244) = 13.93$; $p < .0001$, $\eta^2 = 0.14$]. Follow-up univariate analyses indicated that School 4 ($M = 11.0$, $SD = 4.0$) had a significantly lower mean score than the other schools on the SLA-D. This was indicated in our initial comparison across the five schools. We elected to keep School 4 in our analyses because the overall population in School 4 was the most diverse with respect to ethnicity and socioeconomic status and we wanted to ensure that decisions we made about this measure would reflect a wide range of student groups. The goal of this project is to create a measure of scientific literacy that would be sensitive enough to pick up variability across groups and educational experiences. If we dropped school 4 from our analyses, we may have elected to delete more items and consequently develop a test that was too difficult or culturally limited to assess a wide variation in demonstrated scientific literacy among diverse students.

There were no statistically significant differences in the four test variations [$F(3, 244) = 2.64$; $p < .06$, $\eta^2 = 0.03$], or grade level [$F(1, 244) = 2.21$; $p < .14$, $\eta^2 = 0.01$]. Since some schools provided test results for only 1 grade, it was not possible to test for the interactions with school and grade, but none of the calculated higher order interactions were statistically significant. The lack of differences on these measures suggests that the order of the measure in terms of multiple-choice items or motivation and belief scales coming first or second had no bearing on participants' scores on the multiple-choice items, and that the two versions (SLA-D1 and SLA-D2) are not statistically significantly different (i.e., they are equivalent). The means and standard deviations of scores on the multiple-choice items by school, grade level, and test form can be found in Table 6.

The Kuder–Richardson equation 20 reliability (1937) for the two versions of the SLA-D was 0.83 and 0.82, respectively. The discrimination indices for the 41 items that make up both versions of the SLA-D ranged from 0.30 to 0.85 with a mean and median of 0.58. The highest percent correct for any one item was 89% and the lowest was 22%. Overall 12 of the 41 items (29%) were selected correctly by 50% or fewer students.

To assess the reasonableness of the scientific literacy components assessed by the SLA-D items, we conducted a principal components factor analysis. The purpose was to test whether the items aligned with each component could be used as independent measures, despite our previous finding of low reliability by component (see Study 2). The analysis indicated that the SLA-D for this participant pool also assessed a single factor of demonstrated scientific literacy. As such we recommend using the SLA-D as a single measure of demonstrated scientific literacy.

SLA-MB Evaluation. One of the 264 participants did not respond to the motivation and belief scale items, leaving 263 responses for analysis on SLA-MB scales. Exploratory principal components factor analysis on the 25 items of the SLA-MB indicated that they formed three well-defined and unique components with the items having high eigenvalues (minimum eigenvalues of 0.62, 0.55, and 0.36 on the first three components). This suggests that the SLA-MB assesses three distinct sets of beliefs associated with scientific literacy

TABLE 6
Study 3: Prototype SLA-D Mean Scores by Version, Variation, School, and Grade

		School					
		1	2	3	4		
Test Form	Grade →	Eighth	Eighth	Seventh	Seventh	Eighth	All
1: SLA-D1 → SLA-MB	*n*	9	35	13	4	9	70
	M (*SD*)	17.1 (4.2)	18.3 (4.7)	13.8 (4.9)	11.5 (2.1)	13.4 (3.5)	16.3 (4.9)
2: SLA-MB → SLA-D1	*n*	8	33	7	5	8	61
	M (*SD*)	13.6 (4.1)	18.1 (5.1)	14.1 (5.5)	10.6 (5.9)	13.0 (3.3)	15.8 (5.4)
3: SLA-D2 → SLA-MB	*n*	10	32	10	5	7	64
	M (*SD*)	13.1 (6.8)	15.6 (4.9)	12.9 (5.5)	8.0 (2.4)	10.6 (4.3)	13.6 (5.5)
4: SLA-MB → SLA-D2	*n*	11	33	11	6	8	69
	M (*SD*)	15.3 (3.9)	16.5 (4.7)	12.9 (4.9)	9.3 (3.9)	10.5 (4.0)	14.4 (5.1)
All	*n*	38	133	41	20	32	264
	M (*SD*)	14.8 (5.0)	17.1 (4.9)	13.4 (5.0)	9.8 (3.9)	12.0 (3.8)	15.0 (5.3)

TABLE 7
Study 3: Descriptive Statistics and Reliability Scores for the SLA-MB ($n = 263$)

Component of SLA-MB	*M*	*SD*	*A*
Value of science (Wigfield & Eccles, 2000)	3.9	0.7	0.80
Self-efficacy for scientific literacy (adapted from Kettlehut, 2010)	3.8	0.6	0.72
Source and certainty of scientific knowledge (Conley et al., 2004)	3.7[a]	0.8	0.88

Minimum score = 1, Maximum score = 5.
[a]Reverse coded.

TABLE 8
Study 3: Correlation Matrix for SLA-MB and SLA-D

Study 3: Correlation Matrix for SLA-MB and SLA-D ($N = 263$)				
Component of SLA-MB	Value	Self-Efficacy	Knowledge[a]	SLA-D
Value of science (Wigfield & Eccles, 2000)	1.000	.530 ($p < .0001$)	−.110 ($p = .060$)	.100 ($p = .120$)
Self-efficacy for scientific literacy (adapted from Kettlehut, 2010)		1.000	.050 ($p = .400$)	.400 ($p < .0001$)
Scientific knowledge is uncertain and constructed[a] (Conley et al., 2004)			1.000	.370 ($p < .0001$)
SLA-D				1.000

[a]Reverse coded.

and that these scales can be used independently of one another. Each scale consistent with previous findings demonstrated sound reliability (value of science: $\alpha = .80$; self-efficacy for science literacy: $\alpha = .72$; source and certainty of scientific knowledge: $\alpha = .88$). Table 7 provides the means, standard deviations, and Cronbach's alpha for each of these components.

SLA-MB Relation to SLA-D Scores

Correlations among the three SLA-MB components and the SLA-D are shown in Table 8. Among the three constructs, there was a strong positive correlation ($r = .53$) between the mean score on the value of science and self-efficacy. This is consistent with what is seen in the expectancy-value research, that students frequently report valuing tasks that they feel confident in achieving (or the reverse). There was no correlation between the students' total score on the SLA-D and the mean response on the value of science construct, and moderate correlation with self-efficacy ($r = .40$) and personal epistemology ($r = .37$). The relation between self-efficacy and achievement suggests that students may have a good sense for their ability to engage in scientific literacy. Furthermore, the moderate correlation with personal epistemology demonstrates a relation between understanding knowledge as tentative and constructed with students' ability to demonstrate scientific literacy. Visual inspection of scatter plots (not shown) confirmed these interpretations.

LIMITATIONS

Our validation argument is limited. First, we do not provide evidence based on relations to other variables. This first omission is a serious limitation of our current work. Future research with this measure needs to establish relations between scores on the SLA-D with other criteria that provide evidence of scientific literacy, such as science grades, performance evaluations of scientific literacy, or scores on a similar test. Unfortunately, the development process at the time of this investigation did not allow for such comparisons as data were gathered from students anonymously. Furthermore, while the measure was still in development we did not think that this would be an adequate use of our resources. This is a much needed next step to add to the validation argument for the SLA. The SLA should be tested as a pre–post assessment tool for lessons, units, or courses in which scientific literacy is systematically addressed. Gains in scores at posttest, especially relative to an independent assessment of student learning, would provide further evidence that the SLA assesses scientific literacy. Future work needs to establish this evidence to support widespread use.

Second, our participant pool for the prototype study was limited to four schools in one district, in one U.S. state. This limits the generalizability of the findings in this investigation. Further testing of this SLA needs to be conducted with participants in other states and countries to ensure the effectiveness of this tool in assessing scientific literacy in different cultural contexts. It should be noted that our participants represented a range of ethnicities and socioeconomics levels.

Third, additional content and response process evidence could be garnered from a sample of practicing middle school level teachers. While initial versions of the SLA-D were shared with four teachers (two middle level and two high school) and the development team included a former middle school science teacher and a science educator who develops programming for middle school students, a review of these tools by a larger group of teachers could prove informative in moving forward with revised versions of this measure.

Fourth, we do not provide any evidence based on test consequences. Evidence based on consequences of testing should demonstrate that any negatives associated with taking a test are outweighed by positive outcomes; furthermore, evidence of this type should demonstrate the likelihood of intended benefits actually occurring (AERA, APA, NCME, 1999). The inclusion of testing consequences as a source of validity is described as the "most contested validity territory" (Cizek, Rosenbertg, & Koons, 2008, p. 398). Some scholars (e.g., Kane, 2001; Linn, 1997; Messick, 1995; Shepard, 1997) argue for including this in conceptions of validity because "negative consequences can render a score's use as unacceptable" (Kane, 2013, p. 1). However, others (e.g., Borsboom, Mellenbergh, & van Heerden, 2004; Cizek et al., 2008; Cizek, 2012; Dwyer, 2000; Popham, 1997) have argued against its inclusion in validity theory for reasons such as "the social consequences of score use do not bear on the validity of score interpretations" (Cizek, 2012, p. 3).

Fifth, we have pragmatic concerns about the length of the SLA-D and the feasibility of students' completing this measure in one 40-minute class session. To address this concern, we recommend shortening the SLA-D to 19 items (2 minutes per item). Following the same iterative analyses and processes described in the pilot studies as well as contextual factors such as amount of reading required for an item (we took on a "less is better" approach) and an adequate distribution of items across components of our scientific literacy framework, we reviewed the 41 items of the SLA-D and identified seven items to delete from each version of the SLA-D. If these deletions are made, our final recommended prototype measure includes two versions of the SLA-D, each with 19 multiple-choice items (nine shared items

between versions for a total of 29 unique items) and all 25 items from the motivation and belief scales. The statistical analyses on the items from these reduced SLA-D versions based on the 264 responses to the 26-item tests indicated results very similar to the results presented in our Prototype test (data not shown). Therefore, we feel that it is possible to make reasonable inferences about middle schools students' scientific literacy based on the reduced measure. In Appendices 1 and 2 in the Supporting Information, we indicate which items to redact from the SLA-D versions.

DISCUSSION OF THE SLA

The SLA is intended to assess middle schools students' sense of field/discipline general scientific literacy. The SLA is designed to be administered in one class period (40–50 minutes) via a paper and pencil format. The SLA has two parts: the SLA-D that assesses five components of demonstrated scientific literacy and the SLA-MB modified from three existing scales that measure scientific literacy motivation and beliefs. There are two versions of the SLA-D portion. Each includes 26 multiple-choice items (11 shared and 15 unique items on each version), presented in Appendices 1 and 2 in the Supporting Information, including directions for shortening the measure if needed. The SLA-D items are written at, or below, the sixth-grade level according to the Flesch–Kincaid index. The SLA-MB is composed of three subscales for a total of 25 Likert items that include the three motivation and belief scales (Appendix 3 in the Supporting Information).

Our findings support using the SLA-D and SLA-MB to assess middle school students' scientific literacy. Through careful attention to the standards for developing validity arguments (AERA, APA, NCME, 1999), we have provided comparative validity evidence related to test content, response process, and internal structure. The results of our iterative process of item construction, administration, and revision provide support that the SLA-D and SLA-MB align with the underlying conceptualization of scientific literacy that we sought to assess. In addition, the development of this measure was guided by experts from SEPA, an interdisciplinary research team, and a sound conceptualization of scientific literacy based on the extant literature.

The SLA-D items in both versions demonstrate good reliability, and the items on each adhere to recommended guidelines for percent correct, discrimination index, item-total correlation coefficients, and frequency distribution of distractors selected, all of which provide evidence for the strong internal structure of this measure. Furthermore, the lack of statistical or practical difference in scores from students responding to the two versions of the SLA-D suggests that these versions are assessing equivalent information. For these reasons, we recommend the use of this measure with middle school students and encourage users to evaluate the reliability in their data and consider the appropriateness of this tool for providing valid evaluations of scientific literacy in the contexts in which it is used.

The correlations among the SLA-D score and the SLA-MB scales further inform our understanding of scientific literacy as abilities, motivation, and beliefs. The interesting distinctions in correlations among value of science, self-efficacy for scientific literacy, and personal epistemology indicate a potentially developing sense of scientific literacy in these students. For instance, students who value science would likely feel that they are good at science, but their personal epistemology may not yet be well formed so it is unrelated to appreciation and ability. The low correlation between total item score and the value of science score reflects the possible disconnect between appreciation and ability; the higher correlation between total item score and self-efficacy reflects the link of self-assessment and external assessment. The moderate correlation between total item score and the personal

epistemology score is an indication that those who understand the nuance of science will also be better at science as measured by an external evaluation.

IMPLICATIONS AND CONCLUSIONS

While the current tool is still a work in progress, we see implications for the present work to inform both pedagogical practice and theoretical development of the construct of scientific literacy. Pedagogically, the relations among the SLA-D and SLA-MB illustrate the importance of teaching not just the content of science literacy but also the need to allow students opportunities to develop beliefs and values that support the use of science in their lives. Classroom teachers could use the SLA as a formative assessment at the beginning of the academic year to target aspects of scientific literacy (knowledge, beliefs, and values) for instruction during the school year. Furthermore, teachers could use the SLA-MB to begin a conversation with their students about the students' personal epistemology, value, and motivation for science. Exposing such beliefs could be a first step in helping students to better understand themselves in relation to science.

Theoretically, the SLA provides a measure of demonstrated knowledge as well as students' beliefs and motivation. To achieve the goal of a scientifically literate society, individuals need to be more than knowledgeable of the science content, they must also value that content and be open to it as a source of information for decision making. The correlational results presented here, while still tentative, indicate that a relationship between demonstrated knowledge and motivation and beliefs exist. Furthermore, we identified three key areas of motivation and belief for inclusion in conceptions of scientific literacy: personal epistemology, self-efficacy, and value. The field of motivation offers a variety of other constructs that may also prove informative. Thus, these findings tentatively suggest that further theoretical and empirical investigation into the nature of knowledge, motivation, and a belief as part of scientific literacy is warranted.

The concept of scientific literacy "has become an internationally well-recognized educational slogan, buzzword, catchphrase, and contemporary educational goal" (Laugksch, 2000, p. 71) despite the lack of agreement on just what it *is* (see Dillon, 2009; Holbrook & Rannikemae, 2009; Laugksch, 2000; Roberts, 2007). We have developed a measure of scientific literacy that is appropriate for middle school students. It is not designed to assess specific content knowledge, such as Newton's law of gravity or Boyle's law of thermodynamics, but measures a functional understanding and appreciation of science.

While the SLA has met standard measures of internal structure and reliability, we do not consider the test to be fully validated in either the technical or the vernacular sense. We see the studies presented here as the intermediate steps of a work in progress and would like interested groups to use and evaluate this test to develop a wide group validation along the lines of *crowd sourcing.* We hope that through this mechanism a reasonable and useful test of scientific literacy can be fully developed from our work. We believe that such a tool is necessary to the promotion of scientific literacy that in turn can aid in combating ignorance about the importance of science and promote rational scientific policy decision making in a democratic society.

The authors would like to thank Mark Kaelin under whose guidance the project was conceived and the team assembled, and Tony Beck of the SEPA program for his encouragement and support. We appreciate Lisa Abrams, Mike Kennedy, Marian Passannante, Kristin Bass, and Ron Vangi for their expert advice, and Doug Larking for reading a prior version of this paper. Finally, we thank the middle school science teachers and students who good-naturedly participated in our testing and understood the importance of their contribution to the research.

REFERENCES

AAAS Project 2061 Science Assessment Website. (2011). American Association for the Advancement of Science, Project 2061. Retrieved April 9, 2011, from http://assessment.aaas.org

Aikenhead, G. (2011). Towards a cultural view on quality science teaching. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), The professional knowledge based of science teaching (pp. 107–127). New York: Springer.

Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). How subject-matter knowledge affects recall and interest. American Educational Research Journal, 31(2), 313–337.

American Association for the Advancement of Science. (1989). Project 2061: Science for all Americans. Washington, DC: Author. Retrieved February 12, 2011, from http://www.project2061.org/publications/sfaa/online/sfaatoc.htm.

American Association for the Advancement of Science. (1993). Benchmarks for scientific literacy. Project 2061. New York: Oxford University Press.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Psychological Association.

Arons, A. B. (1983). Achieving wider scientific literacy. Daedalus, 112(2), 91–122.

Ausubel, D. P. (1977). The facilitation of meaningful verbal learning in the classroom. Educational Psychologist, 12(2), 162–178.

Baldwin, R. S., Peleg-Bruckner, Z., & McClintock, A. H. (1985). Effects of topic interest and prior knowledge on reading comprehension. Reading Research Quarterly, 20(4), 497–504.

Bandura, A. (1997). Self-efficacy: The exercise of control. New York: Freeman.

Berry, A., Loughran, J., & Mulhall, P. (2007). Values associated with representing science teachers' pedagogical content knowledge. In D. Corrigan, J. Dillon, & R. Gunstone (Eds.), The re-emergence of values in science education (pp. 149–163). Rotterdam, The Netherlands: Sense Publishers.

Bøe, M. V. (2012). Science choices in Norwegian upper secondary schools: What matters? Science Education, 96, 1–20.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. Psychological Review, 111, 1061–1071.

Braten, I., Stromso, H., & Samuelson, M. S. (2008). Are sophisticated students always better? The role of topic-specific personal epistemology in the understanding of multiple expository texts. Contemporary Educational Psychology, 33, 814–840.

Britner, S. L., & Pajares, F. (2001). Self-efficacy beliefs, motivation, race, and gender in middle school science. Journal of Women and Minorities in Science and Engineering, 7, 271–285.

Bryan, R. R., Glynn, S. M., & Kittleson, J. M. (2011). Motivation, achievement, and advanced placement intent of high school students learning science. Science Education, 95(6), 1049–1065.

Business Higher Education Forum (BHEF). (2011). Creating the workforce of the future: The STEM interest and proficiency challenge (2011). Retrieved August 31, 2012, from http://www.bhef.com/publications/documents/BHEF_Research_Brief-STEM_Interest_and_Proficiency.pdf.

Bybee, R. W. (2008). Scientific literacy, environmental issues, and PISA 2006: The 2008 Paul F-Brandwein lecture. Journal of Science Education and Technology, 17, 566–585.

Chen, J. A., & Pajares, F. (2010). Implicit theories of ability of Grade 6 science students: Relation to epistemological beliefs and academic motivation and achievement in science. Contemporary Educational Psychology, 35(1), 75–87.

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. Psychological Methods, 17(1), 31–43.

Cizek, G. J., Rosenbertg, S. L., & Koons, H. H. (2008). Sources of validity evidence for educational and psychological tests. Educational and Psychological Measurement, 68, 398–412.

Conley, A. M., Pintrich, P., Vekiri, I., & Harrison, D. (2004). Changes in epistemological beliefs in elementary science students. Contemporary Educational Psychology, 29, 186–204.

Cross, R. T., & Price, R. F. (1992). Teaching science for social responsibility. Sydney, Australia: St Louis Press.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. Journal of Research in Science Teaching, 37, 582–301.

Dillashaw, F. G., & Okey, J. R. (1980). Test of the integrated science process skills for secondary science students. Science Education, 64(5), 601–608.

Dillon, J. (2009). On scientific literacy and curriculum reform. International Journal of Environmental & Science Education, 4, 201–213.

Drennan, J. (2003). Cognitive interviewing: Verbal data in the design and pretesting of questionnaires. Journal of Advanced Nursing, 42(1), 57–63.

Duit, R., & Treagust, D. F. (2003). Conceptual change: A powerful framework for improving science teaching and learning. International Journal of Science Education, 25, 671–688.

Duncan, D. B. (1975). t-tests and intervals for comparisons suggested by the data. Biometrics, 31, 339–359.

Dwyer, C. A. (2000). Excerpt from validity: Theory into practice. The Score, 22, 6–7.

Eccles, J., Barber, B., & Jozefowicz, D. (1999). Linking gender to educational, occupational, and recreational choices: Applying the Eccles et al. model of achievement-related choices. In W. B. Swann, Jr., J. H. Langlois, & L. A. Gilbert (Eds.), Sexism and stereotypes in modern society: The gender science of Janet Taylor Spence (pp. 153–192). Washington, DC: American Psychological Association.

Eccles, J., & Wigfield, A. (2002). Motivational beliefs, values, and goals. Annual Review of Psychology, 53(1), 109–132.

Fuchs, B. A. (2008).Teaching Science as Inquiry: Successes and Challenges in the U.S. Presented at NIH Blueprint K-12 Neuroscience Research–K-12 Education Workshop. Rockville, MD.

Gauld, C. (1982). The scientific attitude and science education: A critical reappraisal. Science Education, 66, 109–121.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education, 15, 309–334.

Hamm, M. (1992). Achieving scientific literacy through a curriculum connected with mathematics and technology. School Science and Mathematics, 92, 6–9.

Hazen, R. M., & Trefil, J. (1991). Science matters: Achieving scientific literacy. New York: Doubleday.

Hodson, D. (1999). Going beyond cultural pluralism: science education for socio-political action. Science Education, 83, 775–796.

Hofer, B. K. (2000). Dimensionality and disciplinary differences in personal epistemology. Contemporary Educational Psychology, 25, 378–405.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. Review of Educational Research, 67, 88–140.

Holbrook, J., & Rannikemae, M. (2007). The nature of science education for enhancing scientific literacy. International Journal of Science Education, 29, 1347–1362.

Holbrook, J., & Rannikemae, M. (2009). The meaning of scientific literacy. International Journal of Environmental & Science Education, 4(3), 275–288.

Hopkins, K. D. (1998). Educational and psychological measurement and evaluation (8th ed.). Boston: Allyn and Bacon.

Jarman, R., & McClune, B. (2007). Developing scientific literacy: Using news media in the classroom. Maidenhead, England: McGraw-Hill International.

Jenkins, E. (2003). School science: Too much, too little, or a problem with science itself? Canadian Journal of Science, Mathematics and Technology Education, 3(2), 269–274.

Johnson, O. K. (1951). The effect of classroom training up on listening comprehension. Journal of Communication, I, 58.

Kane, M. T. (1994). Validating interpretative arguments for licensure and certification examinations. Evaluation & the Health Professions, 17, 133–159.

Kane, M. T. (2001). Current concerns in validity theory. Journal of Educational Measurement, 38, 319–342.

Kane, M. T. (2012). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50, 1–73.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(1), 1–73.

Ketelhut, D. J. (2010). Assessing gaming, computer and scientific inquiry self-efficacy in a virtual environment. In L. A. Annetta & S. Bronack (Eds.), Serious educational game assessment: Practical methods and models for educational games, simulations and virtual worlds. Amsterdam, The Netherlands: Sense Publishers.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160.

Lambert, J. (2006). High school marine science and scientific literacy: The promise of an integrated science course. International Journal of Science Education, 28, 633–654.

Laugksch, R. C. (2000). Scientific literacy: A conceptual overview. Science Education, 84(1), 71–94.

Laugksch, R. C., & Spargo, P. E. (1996). Construction of a paper-and-pencil *Test of Basic Scientific Literacy* based on selected literacy goals recommended by the American Association for the Advancement of Science. Public Understanding of Science, 5(4), 331–359.

Lent, R. W., Brown, S. D., & Gore, P. A., Jr. (1997). Discriminant and predictive validity of academic self-concept, academic self-efficacy, and mathematics-specific self-efficacy. Journal of Counseling Psychology, 44, 307–315.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. Educational Measurement: Issues and Practice, 16, 14–16.
Liu, X. (2009). Beyond science literacy: Science and the public. International Journal of Environmental & Science Education, 4(3), 301–311.
Messick S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. American Psychologist, 50, 741–749.
Millar, R., & Osborne, J. (1998). Beyond 2000: Science education for the future. The report of a seminar series funded by the Nuffield Foundation. London: School of Education, King's College London.
Miller, J. D. (1983). Scientific literacy: A conceptual and empirical review. Daedalus, 112, 29–48.
National Assessment Governing Board (NAGB). (2010). Science framework for the 2011 National Assessment of Educational Progress. ED 512544. Washington, DC: U.S. Department of Education.
National Research Council. (1996). National Science Education Standards. Washington, DC: National Academy of Science Press.
National Research Council. (2002). Scientific research in education (Committee on Scientific Principles for Education Research. In R. J. Shavelson & L. Towne (Eds.), Center for Education, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.
National Research Council. (2012). A framework for K-12 science education: Practices, crosscutting concepts, and core ideas. Committee on a conceptual framework for new K-12 science education standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.
National Science Teachers Association. (1991). Position statement. Washington, DC: Author.
Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. Science Education, 87(2), 224–240.
Organisation for Economic Co-operation and Development. (2003). The PISA 2003 assessment framework–mathematics, reading, science and problem solving: Knowledge and skills. Paris: Author.
Organisation for Economic Co-operation and Development. (2006). Assessing scientific, reading, and mathematical literacy. Paris: Author.
Organisation for Economic Co-operation and Development. (2007). PISA 2006: Science competencies for tomorrow's world, volume I analysis. Paris: Author.
Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. International Journal of Science Education, 25(9), 1049–1079.
Pajares, F., & Valiante, G. (1997). Influence of self-efficacy on elementary students' writing. Journal of Educational Research, 90, 353–360.
Perkins, D. N., Jay, E., & Tishman, S. (1993). Beyond abilities: A dispositional theory of thinking. Merrill-Palmer Quarterly, 39, 1–21.
Popham, W. J. (1997). Consequential validity: Right concern–wrong concept. Educational Measurement: Issues and Practice, 16, 9–13.
Presser, S., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., Rothgeb, J. M., et al. (2004). Methods for testing and evaluating survey questions. The Public Opinion Quarterly, 68, 109–130.
Reynolds, C. R., Livingston, R. B., & Wilson, V. (2006). Measurement and assessment in education. Boston: Pearson.
Roberts, D. A. (2007). Scientific literacy/science literacy. In S.K. Abell & N.G. Lederman (Eds.), Handbook of research on science education (pp. 729–780). Mahwah, NJ: Erlbaum.
Ryder, J. (2001). Identifying science understanding for functional scientific literacy. Studies in Science Education, 36, 1–44.
Shamos, M. (1995). The myth of scientific literacy. New Brunswick, NJ: Rutgers University Press.
Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52, 591–611.
Shaw, D., & Young, S. (2004). Revised guidelines for conducting item analyses of classroom tests. The Researcher, 18, 15–22.
Shell, D. F., Colvin, C., & Bruning, R. H. (1995). Self-efficacy, attribution, and outcome expectancy mechanisms in reading and writing achievement: Grade-level and achievement-level differences. Journal of Educational Psychology, 87, 386–398.
Shen, B. S. P. (1975). Science literacy and the public understanding of science. In S. B. Day (Ed.), Communication of scientific information (pp. 44–52). Basel, Switzerland: S. Karger A.G.
Shepard, L. A. (1997). The centrality of test use and consequences for test validity. Educational Measurement: Issues and Practice, 16, 5–8, 13, 24.
Showalter, V. M. (1974). What is united science education? Part 5. Program objectives and scientific literacy. Prism II, 2, 3–4.

Sireci, S. G. (2013). Agreeing on validity arguments. Journal of Educational Measurement, 50, 99–104.

Songer, N. B., & Linn, M. C. (1991). How do students' views of science influence knowledge integration? Journal of Research in Science Teaching 28: 761–784.

United Nations Educational, Scientific and Cultural Organization. (1993). Final report: International forum on scientific and technological literacy for all. Paris: Author.

Wenning, C. J. (2006). Assessing nature-of-science literacy as one component of scientific literacy. Journal of Physics Teacher Education Online, 3(4), 3–10.

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. Journal of Physics Teacher Education Online, 4, 21–24.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. Contemporary Educational Psychology, 25, 68–81.

Yore, L. D., Pimm, D., & Tuan, H. (2007). The literacy component of mathematical and scientific literacy. International Journal of Science and Mathematics Education, 5(4), 559–589.

Zimmerman, C., Bisanz, G. L., Bisanz, J., Klein, J. S., & Klein, P. (2001). Science at the supermarket: a comparison of what appears in the popular press, experts' advice to readers, and what students want to know. Public Understanding of Science, 10, 37–58.