Journal of Educational Psychology
2014, Vol. 106, No. 3, 881–899

0022-0663/14/$12.00 DOI: 10.1037/a0035833

Testing the Theory of Successful Intelligence in Teaching Grade 4 Language Arts, Mathematics, and Science

Robert J. Sternberg
Cornell University

Linda Jarvin
Paris College of Art

Damian P. Birney
University of Sydney

Adam Naples
Yale University

Steven E. Stemler
Wesleyan University

Tina Newman
Center for Children With Special Needs, Glastonbury, Connecticut

Renate Otterbach
University of San Francisco

Carolyn Parish
SRA International, Fairfax, Virginia

Judy Randi
University of New Haven

Elena L. Grigorenko
Yale University

This study addressed whether prior successes with educational interventions grounded in the theory of successful intelligence could be replicated on a larger scale as the primary basis for instruction in language arts, mathematics, and science. A total of 7,702 4th-grade students in the United States, drawn from 223 elementary school classrooms in 113 schools in 35 towns (14 school districts) located in 9 states, participated in the program. Students were assigned, by classroom, to receive units of instruction that were based either upon the theory of successful intelligence (SI; analytical, creative, and practical instruction) or upon teaching as usual (weak control), memory instruction (strong control), or critical-thinking instruction (strong control). The amount of instruction was the same across groups. In the 23 comparisons across 10 content units in 3 academic domains, there were only a small number of instances in which students in the SI instructional groups generally performed statistically better than students in other conditions. There were even fewer instances where the different control conditions outperformed the SI students. Implications for the future of SI theory and the scalability of research efforts in general are discussed.

Keywords: successful intelligence, critical thinking, memory, instruction, scalability

Throughout the first decade of the 21st century, educational researchers and policymakers have placed an increased emphasis on the twin goals of (a) using experimental designs to evaluate educational interventions and (b) gaining a greater understanding of the issues related to the scalability of educational interventions. The value placed on interventions that have been experimentally tested is highlighted by repositories such as the U.S. Department of Education's "What Works" clearinghouse (http://ies.ed.gov/ncee/wwc/). Projects related to issues of scalability were funded by the Department of Education in the early to mid-2000s, and the results of these projects

This article was published Online First April 7, 2014.

Robert J. Sternberg, Department of Human Development, Cornell University; Linda Jarvin, Paris College of Art; Damian P. Birney, School of Psychology, University of Sydney; Adam Naples, Child Development Center, Yale University; Steven E. Stemler, Department of Psychology, Wesleyan University; Tina Newman, Center for Children With Special Needs, Glastonbury, Connecticut; Renate Otterbach, Department of General Education, University of San Francisco; Carolyn Parish, SRA International, Fairfax, Virginia; Judy Randi, Department of Education, University of New Haven; Elena L. Grigorenko, Child Study Center and Department of Psychology, Yale University.

This research was supported primarily by National Science Foundation Grant REC-9979843 with additional support from the Javits Act Program (Grant No. R206R000001). We are grateful to Sig Abeles, Jill Citron-Pousty, William Disch, Tona Donlon, Sarah Duman, Rebecca Felton, PJ Henry, Alex Isgut, Steve Leinwand, Delci Lev, Donna Macomber, Mari Muri, Nefeli Misuraca, Paul O'Keefe, Alina Reznitskaya, Robyn Rissman, Morgan Reynolds, Christina Schwarz, Emma Seppala, Gregory Snortheim, Heidi Soxman, Cheri Stahl, and Olga Stepanossova for their invaluable assistance on this project.

Correspondence concerning this article should be addressed to Robert J. Sternberg, Department of Human Development, Cornell University, B44 MVR Hall, Ithaca, NY 14853. E-mail: robert.sternberg@cornell.edu

are beginning to receive increased attention in the empirical literature (Constas & Sternberg, 2006; McKenna & Walpole, 2010).

In the present article, we report on a large-scale empirical field study that also sought to address issues related to scalability. We examined whether applying the theory of successful intelligence to instruction and assessment in Grade 4 language arts, mathematics, and science would result in superior learning outcomes relative to alternative instructional methods, in particular, memory-based instruction and critical-thinking based instruction (strong comparison/control conditions) and teaching as usual—whatever it happened to be (weak comparison/control condition). The study involved the participation of 7,702 fourth-grade students in 113 elementary schools and 223 classrooms across the United States in 35 towns (14 school districts) located in nine states (Alabama, California, Connecticut, Massachusetts, Minnesota, Kansas, North Carolina, South Carolina, and West Virginia), in order to determine whether prior successes with the theory's instructional application could be replicated at scale.

Background

There is evidence to suggest that teaching and assessment may be more effective when they are based in part on cognitive-psychological theories that have been applied to education (Bruning, Schraw, & Norby, 2010; Corno, Cronbach, Kupermintz, & Lohman, 2001). Certainly, this has been a major claim of researchers as well as textbook authors in educational psychology (e.g., Ormrod, 2010; Slavin, 2008; Woolfolk, 2009). One such cognitive-psychological theory is the theory of successful intelligence.

The theory (Sternberg, 1997, 2005, 2010) argues that *successful intelligence* is a person's ability to achieve his or her goals in life, within his or her sociocultural context, by capitalizing on strengths and correcting or compensating for weaknesses, in order to adapt to, shape, and select environments through a combination of analytical, creative, and practical skills (Sternberg, 2003b, 2009; Sternberg, Grigorenko, & Jarvin, 2007; Sternberg, Jarvin, & Grigorenko, 2009). Different students have different combinations of these skills. The theory is based on the notion that students learn in different ways and that they have different strengths in learning (Sternberg, Grigorenko, & Zhang, 2008a, 2008b), just as teachers have different strengths in teaching (Spear & Sternberg, 1987). Our goal is to assist teachers in balancing their teaching in such a way that each of the abilities can be addressed, exercised, and given a chance to develop (Sternberg & Grigorenko, 2000; Sternberg et al., 2007, 2009).

Teaching for analytical thinking means encouraging students to (a) analyze, (b) critique, (c) judge, (d) compare and contrast, (e) evaluate, or (f) assess. When teachers refer to teaching for "critical thinking," some of them may mean teaching for analytical thinking. Examples of exercises designed to develop such skills might ask students to (a) analyze a political speech, (b) critique a work of art, (c) judge the value of a social program, (d) compare and contrast two works of literature, (e) evaluate the conclusions drawn from a scientific experiment, or (f) assess the rationale for a cultural custom.

Teaching for creative thinking means encouraging students to (a) create, (b) invent, (c) discover, (d) imagine if . . . , (e) suppose that . . . , (f) predict . . . , or (g) design. Teaching for creative thinking requires teachers not only to support and encourage creativity but also to role-model it and to reward it when it is displayed (Sternberg & Lubart, 1995; Sternberg & Williams, 1996). Examples of such teaching activities might ask students to (a) create a work of art, (b) invent an alternative ending for a story they read, (c) discover the principle behind a natural phenomenon, (d) imagine what life would be like if global warming continued unabated, (e) suppose that they grew up alingual—having no language at all, (f) predict what will happen in the current civil war in Syria, or (g) design a psychological experiment to test a hypothesis about human behavior.

Teaching for practical thinking means encouraging students to (a) apply, (b) use, (c) put into practice, (d) implement, (e) employ, or (f) persuade someone of something. Such teaching must relate to the real practical needs of the students, not what would be practical for individuals other than the students (Sternberg et al., 2000). Examples might include asking students to (a) apply what they have read in a story to their life, (b) use their knowledge of mathematics to balance a checkbook, (c) put theory into practice in exercising defensive driving, (d) implement a plan for losing (or gaining) weight, (e) employ the rules of haiku and write one, (f) or persuade someone that an argument is sound.

Measurement Research Support for the Theory of Successful Intelligence

A number of different studies have been conducted that validate the premise of the theory of successful intelligence in the field of assessment and measurement. Here we present them only selectively and briefly.

First, assessments based on the theory of successful intelligence appear to map onto skills that are relevant, broadly speaking, to success in life and various indicators of well-being (e.g., Grigorenko & Sternberg, 2001; Sternberg et al., 2000). Second, these assessments have demonstrated adequate psychometric properties (e.g., Kornilov, Tan, Elliott, Sternberg, & Grigorenko, 2012; Sternberg, Castejón, Prieto, Hautamäki, & Grigorenko, 2001). Third, measurements of different kinds of skills (analytical, creative, and practical) can be done relatively independently of each other (e.g., Grigorenko et al., 2009). Fourth, successful-intelligence assessments can improve prediction of grade-point average as well as prediction of success in extracurricular and leadership activities; such assessments also can reduce ethnic-group differences in performance (Sternberg, 2010; Sternberg, Bonney, Gabora, Karelitz, & Coffin, 2010; Sternberg & The Rainbow Project Collaborators, 2006). Finally, as illustrated in Advanced Placement Psychology, Statistics, and Physics tests, the inclusion of creative and practical assessments in addition to memory and analytical ones can reduce ethnic-group differences while increasing construct validity (Stemler, Grigorenko, Jarvin, & Sternberg, 2006; Stemler, Sternberg, Grigorenko, Jarvin, & Sharpes, 2009).

Thus, there is evidence that assessments based on the theory of successful intelligence can provide valuable concurrent and predictive information about cognitive functioning at various stages of the life span and in various settings.

Instructional Research Support for the Theory of Successful Intelligence

A number of instructional studies have been conducted with students in different age groups and in various subjects to validate the relevance of the theory of successful intelligence in the classroom (for more detail and other research support for the theory, see Sternberg, 1985, 1997, 2003b; Sternberg, Jarvin, & Grigorenko, 2011). Here we briefly exemplify two types of relevant studies: aptitude–treatment interaction (ATI) and main effect studies of the theory.

An example of the ATI approach is a study (Sternberg, Grigorenko, Ferrari, & Clinkenbeard, 1999) in which the Sternberg Triarchic Abilities Test (STAT; Sternberg, 2003a) was used to assess analytical, creative, and practical skills through multiple-choice and essay items. The test was administered to 326 children around the United States and in some other countries who were identified by their schools as gifted by any standard whatsoever. Children were selected for a summer program in (college-level) psychology if they fell into one of five ability groupings: high analytical, high creative, high practical, high balanced (high in all three abilities), or low balanced (low in all three abilities). The high-school students ($n = 199$) who came to Yale were then divided into four instructional groups. Students in all four instructional groups used the same introductory-psychology textbook, a preliminary version of Sternberg (1995), and listened to the same psychology lectures, by a Yale professor teaching the introduction to psychology course at Yale College. What differed among the four groups was the type of afternoon discussion section to which students were randomly assigned. They were assigned to an instructional condition that emphasized memory, analytical, creative, or practical instruction. The discussion sessions were taught by qualified instructors with no particular training in, or commitment to, the theory of successful intelligence. Instructors were assigned to the instructional conditions at random and were required to use differential teaching approaches. The instructors were unaware of students' patterns of abilities as revealed by the STAT. Consider examples of instruction. In the memory condition, the participants might be asked to recall the originator of a major theory of depression. In the analytical condition, they might be asked to compare and contrast two theories of depression. In the creative condition, they might be asked to formulate their own theory of depression. In the practical condition, they might be asked how they could use what they had learned about depression to help a friend who was depressed. Students in all four instructional conditions were evaluated in terms of their performance on homework, a midterm exam, a final exam, and an independent project. Each type of work was evaluated for analytical, creative, and practical quality. Thus, all students were evaluated in exactly the same way. The results indicated the presence of an aptitude–treatment interaction whereby students who were placed in instructional conditions that better matched their pattern of abilities outperformed students who were mismatched. For all performance assessments combined, for better matched versus mismatched groups, Cohen's *d*s were 0.343, 0.195, and 0.255 for analytical, creative, and practical, respectively. In other words, when students are taught at least some of the time in a way that fits how they think, they do better in school. These results suggest that the negative Cronbach and Snow (1977) results for aptitude–treatment interactions may have been due to lack of theoretical basis for instruction or of theoretical match between instruction and assessment. Pashler, McDaniel, Rohrer, and Bjork (2008), however, have argued that there is still only weak evidence for aptitude–treatment interactions, and the interested reader can refer to Sternberg et al. (2008b) for an alternative point of view.

Subsequently, a main-effect study of the theory (Sternberg, Torff, & Grigorenko, 1998) examined the learning of social studies and science by third graders and eighth graders. The 225 third graders were students in a very low income neighborhood, and the 142 eighth graders were students who were largely middle to upper middle class. Classroom teachers, and consequently their students, were assigned to one of three instructional conditions pseudorandomly so as to balance the number of students and classrooms in each condition. In the first condition, they were taught the course that they would have learned had there been no intervention (i.e., the emphasis was on memory). In a second condition, teaching emphasized critical (analytical) thinking. In the third condition, students were taught in a way that emphasized a balance of analytical, creative, and practical thinking. All students' performance was assessed for memory learning through multiple-choice assessments as well as for analytical, creative, and practical learning through performance assessments. As expected, students in the successful-intelligence (analytical, creative, practical) condition on average outperformed the other students in terms of the performance assessments. In particular, third graders from the successful-intelligence instructional conditions did better in four out of four comparisons with the standard teaching condition (mean Cohen's $d = 1.082$ for $n = 4$) and in three out of four comparisons with the critical thinking condition (mean Cohen's $d = 0.510$ for $n = 3$). Eighth graders in the successful-intelligence condition did better in seven out of seven comparisons with the standard teaching condition (mean Cohen's $d = 0.842$ for $n = 7$) and in three out of seven comparisons with the critical thinking condition (mean Cohen's $d = 1.332$ for $n = 3$). One could argue that this result merely reflected the way they were taught. Nevertheless, the result suggested that teaching for these kinds of thinking succeeded. More important, however, was the result that children in the successful-intelligence condition outperformed the other children even on the multiple-choice memory tests (Cohen's *d*s were 0.289 and 0.383, and 1.283 and 0.833 for standard and critical thinking instructional conditions in the third- and eighth-grader studies, respectively). In other words, even when the goal is simply to maximize children's memory for information, teaching for successful intelligence is still superior. It enables children to capitalize on their strengths and to correct or to compensate for their weaknesses, allowing them to encode material in a variety of interesting ways.

These results were extended to reading curricula at the middle-school and high-school levels (Grigorenko, Jarvin, & Sternberg, 2002). To illustrate, at the middle-school level ($n = 871$), language arts were taught explicitly for successful intelligence. At the high-school level ($n = 432$), successful intelligence instruction for reading comprehension was infused into instruction in mathematics, physical sciences, social sciences, English, history, foreign languages, and the arts. As in previous studies, each assignment contained analytical, creative, and practical tasks. At both middle- and high-school levels students who were taught for successful intelligence outperformed students who were taught in standard

ways (mean Cohen's $d = 0.483$ for middle-school level and mean Cohen's $d = 0.238$ for high-school level).

Ideally, schools might utilize a uniform broad-based, construct-valid, theoretical model in their instruction and assessment and even in admissions, where relevant (Sternberg, 2010). Of course, the model need not be the theory of successful intelligence. Certainly, there are other models (Gardner, 1993, 2006; Mayer, 2011).

The fundamental difference between the current study and the studies discussed above is its scope and specific characteristics. Unlike previous studies, which were framed as either development and narrowly focused efficacy evaluations (Sternberg et al., 1999) or efficacy and replication studies of the theory of successful intelligence in the classroom (Grigorenko et al., 2002; Sternberg et al., 1998), the present study was conceived of as a scaling-up (Sternberg et al., 2006), main-effect evaluation of the utility of the theory of successful intelligence in actual classrooms.

Scaling up Educational Interventions

Educational research is replete with studies of new and exciting interventions that have been shown to work in one particular context or another. One of the biggest challenges facing the field of educational research, however, is the search for effective interventions (e.g., curricular) that yield similar effects across diverse contexts (Elmore, 1996). In other words, are there interventions that can be successfully scaled up? The concept of "upscaling" is derived from economic theories that are currently pervasive in discussions surrounding education reform in the United States. Specifically, the microeconomic concept of "economies of scale" suggests that certain work can be done more efficiently by increasing the size of operation (Folland, Goodman, & Stano, 2013). In light of such reasoning, several funding agencies, including the National Science Foundation, the Institute for Educational Sciences, and the National Institutes for Health, have been engaged in funding research that has been demonstrated to work in more limited contexts in order to determine whether the results can be replicated on a broader scale (e.g., http://www.nsf.gov/pubs/2002/nsf02062/nsf02062.pdf). Their support has funded research by several teams (e.g., Clements, 2005; Francis, 2011; Fuchs, 2004; Hurtig, 2004; Pane, 2007; Starkey, 2004) as well as the present study.

In their book, Glennan, Bodily, Galegher, and Kerr (2004) comprehensively examined the lessons learned from 15 different curricular programs that attempted to go to scale. Generally speaking, the results of this and other research have found three major factors affecting successful scale up (Glennan et al., 2004). First, if the intervention is developed externally, by sources other than teachers themselves, it is often less costly for schools and districts (Nunnery, 1998). This is not to say that teachers have no input. Rather curricula are often co-constructed, with teachers deciding which components to emphasize. Ultimately, however, the easier and less costly it is to implement a design, the more likely it is to be adopted (Glennan et al., 2004). The successful intelligence intervention in the current study was developed externally, although evaluation input from teachers was central to the process (Randi & Jarvin, 2006).

The second factor affecting the success of educational interventions is whether they involve whole-school reform or targeted reform, in which only some classrooms or student populations receive the intervention. Some evidence suggests that when the whole school is involved, there is greater buy-in across the board, which in turn leads to a greater likelihood of success. The current study represents not only an effort at scaling up but also a large-scale experimental study of different educational interventions. As such, it was neither a targeted reform, per se, nor a traditional whole-school reform.

The third factor impacting the successful scale-up of educational reforms is whether they relate to structural changes, teacher knowledge, or curriculum content. Specifically, prior research has shown that structural changes (e.g., classroom size, student groupings, team teaching) tend to have smaller impacts on educational outcomes than teacher knowledge or curriculum content changes. Within the context of the current study, the focus was on teacher knowledge and curriculum content.

Elias, Zins, Graczyk, and Weissberg (2003) have argued that there is "a need to better document the stories of educational innovation and scaling up efforts so that contextual details can enrich an understanding of what is required for success" (p. 303). The current study is aimed at not only understanding the factors associated with going to scale but also attempting to simultaneously run a large-scale experimental study.

With this work, we attempted to (a) explore whether a curriculum based on the theory of successful intelligence is effective when implemented under conditions that would be typical if a district were to implement it on its own (i.e., without special support from the developer or research team)[1] across a variety of circumstances (e.g., different student populations, different types of schools) and (b) provide an estimate of the robustness of the successful-intelligence instruction. In other words, the main question we sought to address was whether a curriculum based on the theory of successful intelligence would continue to be more effective than instructions relying mostly on memory and/or analytical skills, when implemented on a large scale, with different types of students, school, and teachers. Notably, teacher training and ongoing support provided by the research team were much more limited than in previous studies.

Method

Participants

Given the scope of this study, our aim was to recruit schools representing a wide range of geographical locations (i.e., different states and different student populations: urban, suburban, and rural), ethnic-minority representation, and socioeconomic profiles. In total, 3,270 school districts across the United States were contacted about the program. The final sample included schools from 35 towns located in 11 counties of nine states (Alabama, California, Connecticut, Massachusetts, Minnesota, Kansas, North Carolina, South Carolina, and West Virginia). We worked in 14 school districts represented by 113 elementary schools, 223 teachers, and 223 classrooms. We entered information on 7,702 student participants, and obtained usable data (i.e., complete pre- and

[1] Of course, only agreeable teachers of classrooms in volunteering schools within the district participated in each condition, and thus the ideal—district implementation—is approximated to varying extents.

posttest data) from 7,574 students. Some students received more than one unit of instruction, but the number of units administered and the order of the units were not fixed and varied depending on the fit between each school district's prescribed content and the topics covered in our units. Correspondingly, here we present the analyses unit by unit, with the total number of observations at n = 10,845. All students were fourth graders.

Parents and caregivers were informed of the instructional intervention being implemented in their children's classrooms and that the intervention had been endorsed by the school's district superintendent, principal, and classroom teacher. To facilitate broad acceptance, we kept the information collected on individual students to a minimum. Demographic information was thus obtained at the school level. In total, 49.6% of the students in the schools who participated were girls and 27.8% were underrepresented minorities. A further breakdown of the distribution of demographic information across schools by condition is provided in Table 1 (Table 5 provides additional demographic information as it relates to specific units). Study conditions (successful intelligence = SI, critical thinking = CT, memory = M, and teaching as usual = TAU-control) were randomly assigned to schools. Random assignment at the school level was chosen to avoid contamination within a same school building, in which teachers and students naturally talk to each other and share learning materials. In larger districts equal numbers of schools were assigned to each condition, and in small districts, with fewer than three schools participating, the assignment was random.

The guiding principles behind the critical thinking and memory conditions were drawn, respectively, from the education literature (for an introduction, see Halpern, 1996) and research on memory and mnemonic techniques (for an overview, see Baddeley, Eysenck, & Anderson, 2009). As illustrated in Table 2, there is some overlap of activities across conditions, because critical thinking and analytical thinking in the theory of successful intelligence are similar constructs, and because the corresponding condition included memory activities. The main difference between SI, CT, and M curricula, then, is that the first balances an array of activities whereas the latter two focus on one particular approach (CT or M). Overall, the three versions (SI, CT, and M) have comparable amounts of student activities and require the same duration of classroom time and student time on task to cover the content. Thus, in one case (SI) there is a mixture of different types of activities. In the two other conditions (CT and M) there are more CT and M activities, respectively, and the creative and practical activities that were present in SI are absent.

The study's material development and data collection phase took five years to complete.[2]

Table 1
School-Based Demographic Information Across Intervention Conditions

	Intervention condition			
Variable	SI	CT	M	TAU-control
% female				
M	49.63	48.21	48.71	47.85
SD	3.71	3.06	2.43	
% Asian				
M	3.83	2.56	3.89	7.38
SD	4.36	3.56	5.12	
% Black				
M	20.39	10.71	28.46	11.38
SD	15.40	14.11	27.01	
% Hispanic				
M	8.97	15.50	9.58	10.77
SD	12.28	29.79	17.11	
% White				
M	66.80	71.23	58.06	70.46
SD	22.24	28.74	29.13	
No. schools	43	40	30	1
No. classes	100	65	55	3

Note. SI = successful intelligence; CT = critical thinking; M = memory; TAU-control = teaching as usual control.

Materials

Teaching units. Lesson materials (hereafter, units) were developed for three academic domains (language arts, mathematics, and science) and for different content (e.g., within science there were units on ecology, electricity, light, and magnetism) in a similar manner, equalizing the engagement of targeted skills across the experimental conditions. Each unit was preceded and followed by unit-specific pre- and posttests. The content was based on a thorough review of the standards of each participating state at the time of the creation of the curriculum. We focused on those content topics that a majority of states suggested should be covered in their curriculum at the fourth-grade level. In some cases we selected a topic that was targeted in Grade 4 in one state but in Grade 3 in another state. There was never more than a one-year discrepancy between the topics, however, and, when present, the discrepancy did not influence participation in the study.

The curricula in each of the three instructional treatment conditions (SI, CT, and M) were similar in that they covered the same concepts (e.g., magnetism), contained equal amounts of student activities, and required exact or comparable amounts of classroom instruction and student engagement. They were different, however, in the manner that the concepts were approached, presenting student activities that combined analytical, creative, practical approaches to learning (in the SI condition); or a majority of analytical approaches (in the CT condition); or a majority of activities encouraging memorization (in the M condition). Because the SI instructional approach also engages students in critical thinking, there were some activities that were offered both in the SI curriculum and in the CT curriculum. The same holds true for the memory-based activities that were offered in all three instructional approaches. Table 2 provides an example of how the activities differed in the three instructional approaches: Students in the SI condition had an analytical activity and one practical activity, students in the CT condition had two analytical activities, and

[2] Due to the magnitude and duration of the study, various preliminary reports of the data were produced. These reports included different subsamples of the study or presented analyses of the data in a variety of different ways (e.g., year by year of the study), using different data-analytic approaches, or with a variety of different software. Inevitably, there are differences between the obtained results, although all of the previous analyses have pointed to the advantage of the successful intelligence condition. This is the first presentation of the whole sample, where the analyses were carried out in the most conservative way, unit by unit, across all years of the implementation, utilizing a single analytic framework.

Table 2
Illustration of How Units in the Three Instructional Conditions Covered the Same Content for Students but With Different Instructional Approaches and Activities in the Language Arts Unit on Biography as a Literary Genre (1 Day)

Successful intelligence	Critical thinking	Memory
	Objectives	
Students will be able to (a) explain what a biography is, (b) identify and interpret life events, given a biographical statement, (c) compose (orally and in writing) one-sentence biographical statements.	Students will be able to (a) explain what a biography is, (b) identify and interpret life events, given a biographical statement, (c) compose (orally and in writing) one-sentence biographical statements.	Students will be able to (a) define what a biography is, (b) identify life events, given a biographical statement, (c) compose (orally and in writing) one-sentence biographical statements.
	Activities	
• [Analytical activity]: Given a short biography, identify and categorize life events using a graphic organizer • [Practical activity]: Write biographical statements about friend or family members	• [Analytical activity]: Given a short biography, identify and categorize life events using a graphic organizer • [Memory activity]: Write biographical statements about the subject of a biography	• [Memory activity]: Given a short biography, recall life facts, using notes and a frame as memory aids • [Memory activity]: Write biographical statements about the subject of a biography
	Skills	
• Genre: Biography • Description and interpretation of text • Sentence writing	• Genre: Biography • Description and interpretation of text • Sentence writing	• Genre: Biography • Description of text • Sentence writing

students in the M condition had two memory activities. Table 3 provides the specific activity instructions for the classroom teacher.

Language arts curriculum units. Five thematic language-arts units were completed by the students in each of the three conditions (SI, CT, and M). These five units were titled (a) How and Why Nature Tales (*Wonders of Nature*); (b) Informative Nonfiction (*True Wonders*); (c) Biography (*Lively Biographies*); (d) Quest Literature (*Journeys*); and (e) Mystery (*It's a Mystery*). Thus, in total there were 15 instructionally customized newly developed units (5 content units × 3 treatment conditions). Although the content and duration of each unit were identical, within each condition, each unit was taught with different techniques based on the SI, CT, and M specifications. Students across the three conditions received the same, unit-specific pre- and posttest assessments. That is, there were five pre–posttest pairs corresponding to the five content units.

Intended as an introductory unit, *The Wonders of Nature* introduced students to two short poems about nature, which served to motivate students to "wonder" about the natural phenomena explained in *pourquoi* ("how and why") tales. Students were taught to identify the characteristic elements of pourquoi tales, including the concept of cause and effect. As a culminating activity, students were expected to write their own pourquoi tale.

In *True Wonders,* students learned library research skills. They were expected to develop an understanding of research methods, understand the difference between fiction and nonfiction, and learn to use reading strategies to synthesize information from nonfiction sources.

In *Lively Biographies,* students were exposed to biography as a genre. They engaged in a series of activities that helped them to develop a working knowledge of the nature of the genre, the sequencing events in chronological order, and the use of graphic organizers in the recording of events. Students then interviewed someone and produced a photo-biography.

In *Journeys,* students were engaged in the reading of quest tales and, through a series of activities, gained an understanding of the elements of the quest tale. Students were expected to articulate universal themes, identify and articulate qualities of quest heroes, and demonstrate knowledge of the above through the writing process.

Finally, in the *It's a Mystery* unit, students listened to a read-aloud mystery and at the same time independently read a mystery of their choice. Through activities based on the readings, students gained an understanding of the mystery genre, including how suspense and intrigue are built. For example, students identified the setting, characters, plot development, conflict, and resolution; learned vocabulary common to the genre; discussed human experiences and motives; and followed clues to solve the mystery. Usable data were collected for all five units (see the Note on missing data section below).

Mathematics curriculum units. Five mathematics units, including pre- and postintervention assessments, were completed in each of the four conditions (SI, CT, M, and TAU-control)[3]: (a) *Equivalent Fractions*; (b) *Measurement*; (c) *Geometry*; (d) *Data Analysis and Representation*; and (e) *Number Sense and Place Value.* Thus, in total there were 20 customized instructional units

[3] In our work with multiple districts and schools around the country, we established, due to the wide range of content covered in the various curricula used across the country, that the only domain in which we could implement a TAU condition was Mathematics. The diversity of curricula, pedagogies, and standards was too great in the domains of Language Arts and Science to justify a homogeneous TAU condition.

Table 3
Detailed Descriptions of the Analytical, Practical, and Memory Activities Cited in Table 2

Analytical activity: Listening for life facts—Biographical statement model	*After students demonstrate an understanding of biography, move on to an example of a biography. Ask the students to listen to the biography and try to identify life facts, such as date and place of birth, what the person looks like, or what the person accomplished. Write the biography on chart paper so the children can follow along while you read.* Sample biography: Mrs. Murray was born in San Diego, California, and learned to swim almost before she learned to walk. Her older brother Peter taught her to swim at the marina where their dad worked as a lifeguard. As a youngster, Mrs. Murray liked to race her brother and the other children who swam at the marina. She was a tall, athletic youngster who kept her long, blond hair tied back in a ponytail. Her family was not at all surprised when she joined the high school swim team and won many medals. Today when she is not teaching her fourth grade class, Mrs. Murray still enjoys swimming and teaching her own children to swim at the local beach. *Then ask the children to share "life facts" they learned about the person from hearing the brief biography. For example, they might share that Mrs. Murray is a good swimmer or that she was born in California. You might want to point out that biographies are usually written in the third person because they are about someone else's life story. As the children share what they can remember about your life, write the "life fact" under the appropriate heading on the tag board chart. Use the category labels to prompt the children to remember life facts they heard. Tell them they can use the BIOgraphic organizer as a guide while they are reading.* Note to teacher: A classroom wall chart—a BIOgraphic organizer—can be made out of tag board or flannel board for repeated use throughout the lesson. Ideally, it should be created as a pocket chart so that students can post their sentence strips to sort life facts throughout this unit. Category headings (e.g., accomplishments, appearance, family, friends, occupation) may be changed to fit reading passages. A similar matrix will be used as an advance organizer throughout the unit to assist students in reading biographies for life facts.
Practical/creative activity: Writing biographical statements about friends and family	*After reading and discussing the model biography, tell the students they will finish a brief practical activity in which they will become a biographer. Tell the students that they will be doing a short activity in which they will select something memorable about a person they know well and write one biographical statement about that person. Ask the students to think of someone they know well. You may want to prompt the students to remember different aspects of the person's life by slowly asking them a series of "remember" questions. Tell the students to close their eyes and try to remember what the person looks like, how old the person is, what the person wears, what the person likes to do, where the person works or goes to school, what friends and family members the person has, and what interesting or memorable things the person has done.* *Then ask the students to select one interesting memory and write one statement about this person. Students should write their sentences on a sentence strip/oak tag so that the sentences can be saved and referred to in future lessons, as necessary. Classroom paraprofessionals may be involved in helping the students write a complete sentence and/or checking for correct spelling and punctuation. These sentences will serve as models of short biographical statements. They will also serve as examples of "life facts" or the kinds of information a biography typically tells about a person's life.*
Memory activity: Recall life facts	*After students demonstrate an understanding of biography, move on to an example of a biography. Ask the students to listen to the biography and take notes to memorize life facts, such as date and place of birth, what the person looks like, or what the person accomplished. Write the biography on chart paper so the children can follow along while you read.* Sample biography: Mrs. Murray was born in San Diego, California, and learned to swim almost before she learned to walk. Her older brother Peter taught her to swim at the marina where their dad worked as a lifeguard. As a youngster, Mrs. Murray liked to race her brother and the other children who swam at the marina. She was a tall, athletic youngster who kept her long, blond hair tied back in a ponytail. Her family was not at all surprised when she joined the high school swim team and won many medals. Today when she is not teaching her fourth grade class, Mrs. Murray still enjoys swimming and teaching her own children to swim at the local beach. *Remove the biography and ask students to recall the main life facts they just heard. Ask students to review their notes, set them aside, and then recall the main facts about the person in the biography.*

Note. Text in italics is for the teacher.

(5 content units × 4 treatment conditions); within each condition, each unit was taught using different techniques based on the SI, CT, M, and TAU-control specifications. However, there were only 5 pre–posttest pairs, as students across the four conditions received the same pre- and posttest assessments.

The *Equivalent Fractions* unit was intended as a follow-up to an introductory fractions unit. In it students developed an understanding of the concept of equivalence, modeled equivalent fractions with concrete manipulatives, identified and generated equivalent fractions (denominators less than 12), and applied the concept of equivalent fractions in practical and problem-solving situations.

In the *Measurement* unit, students learned to measure quantities (including time, length, perimeter, area, weight, and volume) in everyday and problem situations. They compared, contrasted, and converted within systems of measurements (customary and metric) and estimated measurements in everyday and problem situations. In addition, students learned about the use of appropriate units and instruments for measurement.

In *Geometry*, students engaged in the identification and modeling of simple two-dimensional and three-dimensional shapes and developed an understanding of their properties (reviewing perimeter, area, and volume). Students were expected to understand and identify geometric concepts such as "congruent," "similar," and "symmetric." Finally, students combined, rotated, reflected, and translated shapes.

In *Data Analysis and Representation*, students were given an opportunity to collect, organize, and display data from surveys, research, and classroom experiments. They used the concepts of range, median, and mode to describe a set of data and to interpret data in the form of charts, tables, tallies, and graphs. They learned about the use of bar graphs, pictographs, and line graphs and the advantages and disadvantages of each.

In the *Number Sense and Place Value* unit, students used number lines to identify and understand negative numbers and the ordering of numbers. They were led to an understanding of how to use the place-value structure of the Base 10 number system and how to identify factors and generate equivalent representations of numbers to use in problem solving. In addition, students explored even/odd numbers, square numbers, and prime numbers.

Usable data were collected only from three of the units: *Equivalent Fractions*, *Measurement*, and *Geometry* (see the Note on missing data section below).

Science curriculum units. Four science units, including pre- and postintervention assessments, were completed in each of three conditions (SI, CT, and M): (a) *The Nature of Light*; (b) *Magnetism*; (c) *Electricity*; and (d) *Ecology*. In total, there were 12 customized instructional units (4 content units × 3 treatment conditions); within each condition, each unit was taught with different techniques based on the SI, CT, and M specifications. There were only 4 pre–posttest pairs, as students across the three conditions received the same pre- and posttests.

The Nature of Light unit introduced the concepts of light, reflection, and refraction. By the end of this unit, students were able to show that light travels in straight lines; give examples illustrating that visible light is made of different colors; list colors of visible light; explain how a prism can separate visible light into different colors; explain how mirrors can be used to reflect light; give examples of absorption; describe and give examples of reflection; give examples and describe refraction; and describe the similarities and differences between absorption, reflection, and refraction.

In the *Magnetism* unit, students learned the properties and uses of magnets. By the end of this unit, students were able to explain the difference between magnetic and nonmagnetic objects; give examples of magnetic and nonmagnetic objects; define magnetism; predict whether two magnets will attract or repel each other; describe the effects of a magnet on a compass; explain the difference between temporary and permanent magnets; define the terms *lodestone* and *keeper* as they apply to magnetism; illustrate that the magnetic force is strongest at the poles; and identify materials that may interfere with a magnetic field.

In the *Electricity* unit, students were engaged in hands-on activities relating to electrical circuits. By the end of this unit, students were able to explain that static electricity occurs when charges are moved from one object to another; give examples of static electricity; explain how an object can become charged; define what a cell is; explain the relationship between a cell and a battery; explain what current electricity is; list the essential components of a series circuit; explain how a series circuit works; explain how a parallel circuit works; explain the difference between a series circuit and a parallel circuit; explain what conductors are; explain what insulators are; and give examples of insulators.

In the *Ecology* unit, students were provided with a basic understanding of the interdependence of organisms and their environments through a series of activities focusing on environmental factors and their impact on animals and people, respectively, and the interdependence of plants and animals. In addition, students developed the skills necessary to conduct scientific investigations and gain an appreciation for science as a discipline. By the end of the unit, students were able to explain what a terrarium is; describe some environmental factors that are important to the growth and survival of plants and animals; give examples of how environmental factors affect the growth and survival of plants; explain how animals depend on the nonliving environment to survive; describe some environmental factors that affect animals' ability to survive and grow; give examples of the effect of the same environmental factor on different animals; describe an ecosystem; explain some of the relationships between plants, animals, and the physical environment; explain how energy passes through an ecosystem; describe the conditions that are necessary for an ecosystem to function; explain how people depend on their environment; give examples of how people can have a positive or negative effect on their environment; and understand why it is important to use natural resources wisely. Only two units, *The Nature of Light* and *Magnetism*, produced usable data (see the Note on Missing Data section below).

Assessments. Unit-specific assessments were developed to capture mastery in the content area of each unit but were generated in such a way that equal numbers of items tapped into the four key abilities at which the intervention conditions were aimed—that is, memory, analytical, creative, and practical abilities (Randi & Jarvin, 2006). Each pre- and posttest had 20–22 items. In order to equalize test difficulty statistically and place pre- and posttest scores on the same measurement scales, we included 3–7 items that were common to both pre- and posttest in each unit. These items were used to obtain ability scores (see below).

Initial rubrics were developed by the research team for all of the units' pre- and postintervention assessments; they were then refined in collaboration with several raters once initial student data had been collected. All the student data were then rated with the final rubrics. The items were roughly equally divided between multiple-choice (scored 0–1) and open-ended (scored 0–5) formats, with 40% to 59% identified as multiple-choice items, depending on the test. Students in all conditions received identical, unit-specific pre- and posttests. Table 4 presents the Cronbach's alpha internal consistency reliability estimates, for pre- and posttests, for both multiple-choice and open-ended questions simultaneously (Rizopoulos, 2006).

Procedures

Assignment to experimental groups. Recruitment efforts were targeted at school districts rather than at individual schools, and we sought permission and buy-in from district superintendents before reaching out to principals. Depending on the size of the district and the number of schools judged by the superintendent to be candidates for

Table 4
Internal Consistency (α) and Construct Reliability (Con. r_{xx}) of Curriculum Unit Pretests and Posttests

Curriculum units	Pretest			Posttest			Common items	*n*
	Items	α	Con. r_{xx}	Items	α	Con. r_{xx}		
Language Arts								
How and Why Nature Tales (*Wonders of Nature*)	22	0.767	0.991	22	0.826	0.995	6	1,626
Informative Nonfiction (*True Wonders*)	22	0.786	0.993	22	0.793	0.995	4	1,233
Biography (*Lively Biographies*)	22	0.845	0.992	22	0.778	0.990	7	752
Quest Literature (*Journeys*)	22	0.783	0.990	22	0.832	0.991	3	520
Mystery (*It's a Mystery*)	22	0.813	0.992	22	0.803	0.988	7	549
Mathematics								
Equivalent Fractions	22	0.816	0.997	22	0.748	0.994	5	1,735
Measurement	22	0.698	0.992	22	0.739	0.993	3	1,550
Geometry	22	0.659	0.990	22	0.775	0.992	3	545
Science								
The Nature of Light	20	0.876	0.991	20	0.848	0.980	6	1,328
Magnetism	20	0.762	0.986	20	0.646	0.982	2	917

Note. Con. r_{xx} = construct reliability of factors jointly estimated with common item anchoring.

participation, one or more experimental conditions were implemented in the district. In all cases, teachers within a given school were assigned the same condition to avoid within building contamination. In other words, in small districts there might be only one participating school, so that the district is confounded with the experimental condition, whereas in a larger district, all conditions might be assigned, always to different schools. Within these constraints, the allocation to experimental condition was random. This design reflects the challenges and constraints of large-scale implementation in diverse settings, where districts and schools need to have voices in making decisions about the experimental interventions they are interested in considering. In other words, administrators decided if the district should participate, and if so, which schools should be involved, but they did not select the experimental condition(s) to be implemented. Although it is difficult to ascertain the full impact of the final allocation of schools to condition, our analyses include pretest scores as a covariate. This is in part to address concerns that even perfectly random allocation does not ensure a balance of student attributes across conditions. Yet another challenge was to get all of the participating teachers to implement all of the instructional units. Although upon recruitment districts committed to working with the whole curriculum (i.e., all units), the delivery of the full curriculum across all participating schools proved impossible due to differences between schools in terms of the content that they wanted to prioritize at the given grade level, scheduling issues due to local tests and other required activities, as well as differences among classrooms in terms of student level and speed of progression through instructional materials.

Teacher training. A 2-day, 12-hour in-service training program was developed and implemented by members of the research team for all the participating teachers. The workshop was tailored to the experimental condition that the participating teachers had been assigned to (i.e., SI, CT, or M).

Day 1 focused on (a) the program design, teacher requirements, and other logistics and (b) the theoretical principles of teaching and instruction for each one of the three experimental conditions. After introductions, teachers were presented with a program overview and the timeline and expectations for participation were reviewed and discussed as a group. The researchers then presented the theoretical underpinnings and prior empirical evidence for the effectiveness of the approach (SI, CT, or M). Teachers in the SI condition thus learned about the previous studies on the effectiveness presented in the introduction to this article; teachers in the CT condition were given examples of critical thinking based instructional interventions, and teachers in the M condition were taught about the effectiveness of different mnemonic strategies for learning material. In addition to learning about earlier work, participants got to practice activities that had proven successful. Again, specific activities practiced varied between the SI, CT, and M groups. Finally, teachers practiced hands-on use of the CORE system. CORE (Collaborative Online Research Environment) is a software package that was designed specifically for this program to allow teachers to access, download, and print curriculum materials, as well as to provide a discussion board allowing them to chat both with other teachers enrolled in the same condition (SI, CT, or M) and with the curriculum developers and content specialists involved in the program.

Day 2 focused on modeling the units in each subject area and provided teachers with an opportunity for hands-on experience with the unit format. Materials distributed included a teacher guide containing instructional material, background information, resource materials reflecting print and nonprint sources, and student workbooks. Teachers received only materials relevant to the instructional approach they were to implement in their classroom. In other words, a teacher trained to implement the SI instructional approach was trained with other teachers implementing the SI approach and saw only the SI instructional materials. Teachers also were introduced to the instructional strategies particular to the condition. An overview of the pre- and postintervention assessments concluded the sessions.

Fidelity monitoring. Fidelity monitoring was carried out in two ways: through the CORE system and by collecting and reviewing all student workbooks to track the level of completion. As mentioned above, the CORE system is a Java-based collaborative environment designed to establish and promote long-distance collaborations with teachers. It was designed, created, and maintained by the Yale University Information Technology Department for the purposes of this program and

enabled the research team to stay in touch with implementing teachers throughout the school year. Because all electronic conversations between teachers and between teachers and research-team members were recorded and stored, the system provided data to measure fidelity of implementation. A second measure was provided by the collected student workbooks, which contained information on which part(s) of a curriculum unit and what activities had been completed by the students in a given classroom. Both teacher logs and student workbooks were analyzed for indicators of fidelity; only those teachers whose students completed all homework assignments, and whose CORE logs were indicative of both understanding of and adherence to the program, were included in the data analyses. We did not have reason to expect (and did not observe) any differences in the usage of the CORE system and in the utilization of the workbooks across instructional treatments (SI, CT, and M). In other words, there were differences across classrooms, with some but not other teachers utilizing the CORE system regularly and some teachers returning student workbooks where every activity had been completed and other teachers returning student workbooks where entire sections were blank, but these differences were observed within each instructional treatment condition. Student workbooks were used as indicators that permitted a participating classroom to be entered in the study database. If the workbook contained less than 70% of the activities completed, the data from a given teacher were not entered into the database. Altogether, ~10% of the participating classrooms in each instructional condition did not meet this criterion.

Data processing. All data processing was carried out at the Center for the Psychology of Abilities, Competencies, and Expertise (PACE Center) at Yale University. Details regarding the management of the data can be found in the Appendix. Close to one hundred casual employees were hired in addition to permanent research-assistant staff to assist with data entry (multiple-choice questions) and coding of open-ended questions. The open-ended questions were coded with a detailed rubric developed by the curriculum developer, and coders were trained to reach satisfactory interrater reliability levels (i.e., the correlations between the pair's open-ended item ratings had to be greater than .70) before they were allowed to start coding materials.

Statistical Analyses

Note on missing data. As we worked with a large number of schools and districts, we could exercise only limited control over what and how many units were selected by teachers to be administered. Buy-in required a commitment to the whole program, but teachers needed to map their preferences for particular units onto their school calendars and other administrative demands. In turn, to include a unit into the analyses, we had to have a reasonable number of students receiving the unit across all study conditions. Unfortunately, this did not happen for two Mathematics and two Science units. Due to small or distinctly uneven distributions of the number of participants across conditions within certain units, the corresponding data were not analyzed for those four units.

Attrition. Extending our reporting of fidelity monitoring, a certain degree of student attrition is also expected as students come and go throughout the school year due to illness and the like. Some students may also not be available for testing at one or the other assessment or may have joined the class part way through the training. An analysis of attrition revealed statistical differences in six of the nine units,[4] although effect sizes (η^2) are small with no consistent pattern for any one condition. There was statistically less attrition in the SI condition for three units (η^2: *Equivalent Fractions* = .005; *Measurement* = .006; and *Magnetism* = .047), less attrition in the M and CT conditions for three units (η^2: *True Wonders* = .042; *The Wonders of Nature* = .015, and mysteries .028), and no statistical differences in attrition for the remaining three units. Of importance, these differences were not related to the intervention differences to be reported shortly. Only students who were available for assessment at both time points were included in the analyses.

Overview of analyses. The analyses we report here were conducted in two stages. First, we derived performance measures for each unit. Second, we ran unit-specific analyses that included a set of covariate and interaction terms.[5] The rationale and general approach for these are described next.

Derivation of performance measures. To combine multiple-choice (binary) and open-ended (ordinal) items into a single ability score, we used Samejima's graded response model (Samejima, 1997), as implemented in Mplus (Muthén & Muthén, 2005), for both pre- and posttest data (such scores have a range of approximately −3 to 3). For the overlapping items that were presented both at pre- and posttest, their loading and threshold (i.e., their discrimination and difficulty parameters) were constrained. This allowed for the statistical equating of pre- and posttest item difficulty. As recommended in the literature (Geiser, Eid, Nussbeck, Courvoisier, & Cole, 2010), scores were calculated for only those individuals with both pre- and posttest data. We do not elaborate on these analyses here; however, details can be obtained from the authors. Traditional internal consistency measures for the pre- and posttest assessments of each unit are provided in Table 4, along with construct reliability estimates (Gefen, Straub, & Boudreau, 2000).

Unit-specific analyses with covariates. As we have described, the students who participated in the current study were sampled from a large and diverse population. One of the touted benefits of a cognitive approach to educational interventions is the real possibility of capturing a much broader and diverse range of approaches to learning. This has certainly been our general experience in the smaller scaled applications of the theory of successful intelligence. One difficulty we faced in the current study is that student-level diversity (e.g., gender and ethnicity) was not collected for reasons described previously. We attempt to capture this diversity and the differential extent that it may impact performance across condition by using a number of school- and classroom-level covariates. The diversity we are capturing is thus in terms of the educational environment, not the child's specific circumstances.

[4] Attrition here is defined as data not available at either pretest or posttest.

[5] These analyses are the culmination of a comprehensive series of analytics conducted in a number of passes across this large database. We acknowledge the reviewers' significant input in shaping the final set we report here.

Following the derivation of measures for each educational unit (5 Language Arts, 3 Mathematics, and 2 Science units), a series of mixed-effects (multilevel) regressions was fit to estimate the effect of intervention condition on the posttest performance. The pretest was always included as a covariate in the regressions. To evaluate the robustness of the obtained results, we repeated the analyses using, inter alia, alternative centering (group mean), a different random clustering variable (school rather than teacher), and the propensity scores approach to match the experimental groups as closely as possible (Dehejia & Wahba, 2002; Ho, Imai, King, & Stuart, 2007). Although there was some variability in the findings (i.e., the magnitude of effects), the pattern of results was generally consistent.[6] The approach we used for the analyses reported here is as follows: There were two levels in the multilevel analysis: students at Level 1 clustered within classroom teachers at Level 2. That is, random effects (covariates and intervention conditions) were estimated at the teacher level (Level 2) to account for classroom level clustering. Students' posttest and pretest performances were modeled at Level 1. Where statistically possible, all models included critical classroom- and school-level demographic variables and their interaction with experimental condition. *Title I status,*[7] *gender* (defined as the proportion of the school population that was male; i.e., *% male*) and *% White* (proportion of the school population that was White) were school-level variables, and *giftedness* (whether the class was identified as a regular or gifted-education classroom) was a classroom-level variable. The % male and % White variables were grand-mean centered for entry alone and as part of interaction terms. It is conceivable to introduce school variability as a third level in the model by clustering classrooms within schools. However, the distribution of the number of classes across schools and intervention conditions was quite broad—on average there were only 1.63 classrooms per school (standard deviation = .45). This suggested to us (and was supported by our preliminary analyses) that the school-level variables would provide little additional statistical information (in relation to their association with student performance) if they were modeled at the school level, rather than at the classroom/teacher level. Furthermore, given the limited variability in number of classrooms per school, a three-level model would be unstable. As such, the decision was made to stay with the simpler two-level model. The regression models were fit in R with the nonlinear mixed effects models (NLME) package (Pinheiro, Bates, DebRoy, Sarkar, & the R Core team, 2009). Note that the NLME package accommodates both linear and nonlinear models; however, in the present study only linear models were run. We treated intervention conditions as multiple, dummy coded variables (with SI as the reference group) in the analyses for each unit. In one or more conditions of some units, covariates were constants or zero. They were excluded from analyses when this occurred.

Results

Sample Data

Table 5 presents descriptive statistics of the unadjusted pretest and posttest performance scores, and the characteristics of the sample by study condition. Of note is the large variability in sample characteristics among the different conditions and different units. This reflects the realities of conducting research during real-time classroom teaching using intact classrooms. To control statistically for this variability, we fit regressions separately for each unit and included pretest as a Level 1 covariate and demographic variables (Title I, % male, % White, and giftedness) as Level 2 covariates. Interaction terms were also entered when possible to capture (in part) variability in the differential functioning of covariates between conditions. All models were run with the same set of covariates first, and for those models that would not statistically converge with all covariates, the models were modified. Covariates not able to be included for a particular model are represented with a dash in Table 6. Regressions were fit with varying intercepts and were grand-mean centered (*Title I* and *giftedness* indicators were not centered because these are binary variables). The analyses, which included the intervention condition coded into multiple dummy-variables with SI as the reference group (i.e., CT vs. SI, and M vs. SI, and, in addition for Mathematics units, TAU vs. SI), revealed the following results. First, all unit analyses included the student-level pretest score as a covariate, and in all cases, as would be expected, it was a statistically significant predictor of posttest performance. We report unstandardized regression coefficients in Table 5 and the graphical representation of this data in Figure 1 (along with 95% confidence intervals). Below is a summary of the results for each academic domain.

Units

Language arts units. There were five language arts units that had analyzable data. Three of the five had a statistically significant effect for intervention condition. Controlling for student pretest score and school-level covariates (gender, % White, and Title I, and their interaction with condition) there was a statistically significant advantage to the SI condition over the CT condition in *Wonders of Nature* ($b = -0.86$, $p = .05$) and *Journeys* ($b = -0.29$, $p = .02$). CT was superior to SI in *Mysteries* ($b = 0.81$, $p = .01$). There were no statistically significant intervention effects for any of the other Language Arts units.

Mathematics units. Three mathematics units had analyzable data. Two of the three had statistically significant effect for intervention condition. Controlling for pretest performance and Level 2 covariates, statistically significant intervention effects were observed for *Equivalent Fractions* in favor of SI over TAU ($b = -0.27$, $p = .01$) and for *Measurement* in favor of Memory over the SI intervention ($b = 0.28$, $p < .04$). There were no statistically significant intervention effects for *Geometry*.

Science units. There were two science units that had analyzable data, and both had statistically significant effect for intervention condition. For *The Nature of Light* unit, there was a significant intervention effect in favor of SI over Memory ($b = -0.78$, $p < .01$). For *Magnetism,* there was a significant

[6] All of these results, as well as the details of the results presented in this article, are available from the authors upon request.

[7] We used Title I data (http://nces.ed.gov/) for each school as a proxy for socioeconomic status.

Table 5
Descriptive Characteristics of the Study Groups

Curriculum units	Test results[a]					Demographic characteristics[b]			
	Pretest		Posttest						
	M	*SD*	*M*	*SD*	*N*	% girls	% White	Title I	Giftedness
Language arts									
How and Why Nature Tales (*Wonders of Nature*)									
SI	0.01	0.87	0.54	1.35	703	51.1	63.2	48.1	26.9
CT	−0.02	0.87	0.43	1.13	542	47.5	88.2	30.8	39.1
M	0.18	0.97	−0.02	1.30	436	49.1	63.6	88.1	24.5
Informative Nonfiction (*True Wonders*)									
SI	0.03	1.03	0.11	0.84	519	50.2	73.2	31.4	34.9
CT	−0.08	0.69	0.02	0.63	377	47.8	88.2	29.2	34.0
M	−0.24	0.95	−0.09	0.81	337	49.2	64.6	82.8	28.2
Biography (*Lively Biographies*)									
SI	−0.09	0.98	0.00	0.41	340	53.6	72.4	69.7	0.0
CT	−0.09	0.87	−0.04	0.43	220	48.2	79.7	56.4	0.0
M	−0.20	0.91	−0.02	0.39	192	48.8	59.1	100.0	0.0
Quest Literature (*Journeys*)									
SI	0.03	0.91	0.20	0.82	322	55.0	68.8	56.8	0.0
CT	−0.25	1.04	−0.25	1.05	144	49.2	75.3	45.1	0.0
M	−0.11	0.72	0.12	0.74	89	52.5	83.8	100.0	0.0
Mystery (*It's a Mystery*)							100.0		
SI	−0.16	0.75	0.08	0.41	232	52.1	62.9	90.5	0.0
CT	−0.59	1.15	−0.05	0.42	157	48.8	88.2	32.5	0.0
M	−0.31	0.68	−0.02	1.30	160	50.0	76.9	100.0	0.0
Mathematics									
Equivalent Fractions									
SI	−0.19	0.89	−0.06	0.46	663	50.5	74.8	24.1	57.5
CT	−0.31	0.94	−0.06	0.47	585	48.8	67.9	21.4	65.5
M	−0.40	0.81	−0.03	0.43	451	50.3	74.5	36.8	47.5
TAU	−1.09	0.57	−0.70	0.34	36	47.9	70.2	100.0	0.0
Measurement									
SI	0.16	0.81	0.09	0.92	548	50.4	78.0	19.0	59.7
CT	−0.03	0.95	0.02	1.04	485	48.3	77.4	27.2	67.0
M	−0.12	0.86	0.01	0.92	485	49.8	69.4	45.4	47.2
TAU	−0.85	0.93	−0.89	1.09	32	47.9	70.2	100.0	0.0
Geometry									
SI	−0.24	0.89	−0.20	0.69	284	50.1	54.1	68.7	0.0
CT	0.65	0.68	0.65	0.55	128	50.1	80.2	4.7	100.0
M	−0.10	0.69	−0.07	0.68	103	47.7	67.2	100.0	0.0
TAU	−1.03	0.60	−0.56	0.75	30	47.9	70.2	100.0	0.0
Science									
The Nature of Light									
SI	0.09	0.94	0.01	0.31	617	49.9	69.7	20.5	76.3
CT	0.17	0.86	0.08	0.34	444	49.1	72.7	5.6	81.3
M	−0.36	0.84	−0.16	0.27	267	47.3	62.7	30	63.7
Magnetism									
SI	0.05	0.83	−0.08	0.72	345	52.5	65.4	0.0	84.6
CT	−0.24	0.86	0.18	0.60	453	47.7	69.4	0.0	100.0
M	−0.38	0.98	0.03	0.57	119	47	79.7	0.0	100.0

Note. SI = successful intelligence; CT = critical thinking; M = memory; TAU = teaching as usual control.
[a] The pretest and posttest scale is a function of Samejima's graded response model (Samejima, 1997); 0 is defined as the average ability level for individuals as measured by the test. [b] School-level data (average of the percentage of students in the school for a given characteristic).

advantage for the critical thinking condition over SI ($b = 0.32$, $p = .04$).

Summary of Analyses

In sum, the analyses, which included the intervention condition coded into multiple dummy-variables with SI as the reference group, revealed 7 effects (out of 23) of mention. There were four cases where SI was advantageous (*Wonders of Nature, Journeys, Equivalent Fractions,* and *Light*), one case where Memory was advantageous (*Measurement*), and two cases in favor of Critical Thinking (*Mysteries* and *Magnetism*). This is not substantially different from what we might expect by chance. The SI intervention did not lead to an overall advantage as expected, but equally it did not lead to a disadvantage.

Table 6
Regression Coefficients (and p Values in Parentheses) for Multilevel Analyses of Curriculum Units

Effects	*Wonders of Nature*	*True Wonders*	*Lively Biographies*	*Journeys*	*It's a Mystery*	*Equivalent Fractions*	*Measurement*	*Geometry*	*The Nature of Light*	*Magnetism*
Conditions vs. SI										
CT	**−0.86 (0.05)**	0.09 (0.68)	−0.06 (0.80)	**−0.29 (0.02)**	**0.81 (0.01)**	−0.02 (0.84)	0.08 (0.59)	0.23 (0.20)	−0.11 (0.55)	**0.32 (0.04)**
M	−0.68 (0.21)	0.32 (0.39)	0.11 (0.14)	0.48 (0.26)	−0.11 (0.34)	0.06 (0.66)	**0.28 (0.04)**	−0.02 (0.93)	**−0.78 (0.00)**	0.15 (0.46)
TAU	—	—	—	—	—	**−0.27 (0.01)**	0.12 (0.30)	0.21 (0.15)	—	—
Covariates										
Pretest	**1.01 (0.00)**	**0.64 (0.00)**	**0.35 (0.00)**	**0.82 (0.00)**	**0.33 (0.00)**	**0.40 (0.00)**	**0.94 (0.00)**	**0.70 (0.00)**	**0.21 (0.00)**	**0.42 (0.00)**
% White	**0.30 (0.00)**	0.12 (0.17)	−0.06 (0.28)	−0.04 (0.63)	0.02 (0.72)	**0.14 (0.00)**	0.06 (0.29)	−0.05 (0.62)	−0.04 (0.55)	**0.17 (0.01)**
% male	0.02 (0.77)	0.03 (0.65)	−0.05 (0.10)	0.03 (0.61)	0.03 (0.47)	−0.06 (0.21)	0.01 (0.83)	0.04 (0.70)	**−0.07 (0.03)**	−0.01 (0.94)
Title I	−0.10 (0.71)	0.18 (0.32)	0.01 (0.85)	−0.03 (0.80)	**0.52 (0.00)**	0.06 (0.47)	0.00 (0.98)	0.12 (0.48)	−0.16 (0.16)	—
Giftedness	0.18 (0.54)	0.26 (0.18)	—	—	—	−0.04 (0.68)	**0.35 (0.01)**	—	−0.17 (0.32)	—
Interactions										
CT × % White	0.32 (0.46)	−0.23 (0.40)	0.12 (0.55)	0.10 (0.27)	−0.25 (0.27)	−0.09 (0.14)	−0.12 (0.06)	0.19 (0.55)	0.01 (0.93)	—
M × % White	**−0.50 (0.00)**	−0.14 (0.19)	0.09 (0.16)	−0.99 (0.34)	−0.11 (0.34)	−0.04 (0.77)	−0.07 (0.42)	0.12 (0.53)	**0.27 (0.01)**	—
CT × % Male	0.25 (0.25)	−0.12 (0.55)	0.08 (0.87)	—	**−0.99 (0.05)**	0.07 (0.25)	0.08 (0.56)	—	0.11 (0.06)	—
M × % Male	−0.09 (0.69)	−0.04 (0.77)	−0.10 (0.48)	—	−0.05 (0.79)	0.03 (0.53)	−0.08 (0.29)	—	**1.01 (0.00)**	—
CT × Title I	0.07 (0.83)	−0.11 (0.67)	—	—	—	0.13 (0.34)	−0.18 (0.37)	—	**0.40 (0.04)**	—
M × Title I	0.52 (0.37)	−0.46 (0.24)	—	—	—	−0.12 (0.33)	−0.06 (0.67)	—	**0.81 (0.01)**	—
CT × Giftedness	0.11 (0.80)	0.14 (0.63)	—	—	—	0.13 (0.31)	−0.07 (0.69)	—	0.16 (0.49)	—
M × Giftedness	0.19 (0.63)	−0.07 (0.80)	—	—	—	0.24 (0.31)	−0.18 (0.35)	—	0.02 (0.92)	—
Intercept	**0.66 (0.00)**	−0.07 (0.53)	**0.66 (0.00)**	**0.19 (0.01)**	**−0.31 (0.03)**	−0.02 (0.78)	**−0.22 (0.01)**	−0.12 (0.38)	0.14 (0.32)	−0.08 (0.43)

Note. Dependent variable is the scaled posttest performance score. Details of covariates and analyses are provided in the text. Bold values represent results that are statistically significant at $p < .05$. Dashes represent covariates that were not able to be included for a model. SI = successful intelligence; CT = critical thinking; M = memory; TAU = teaching as usual.

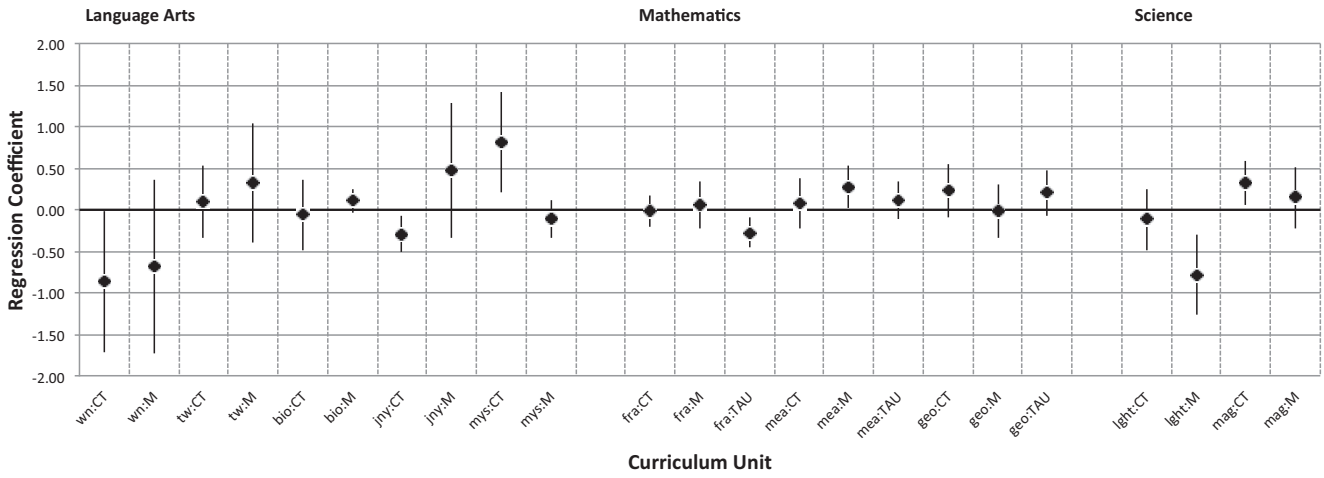

Figure 1. Regression coefficients and 95% confidence intervals of students in the SI condition relative to experimental conditions for each curriculum unit. The SI group is set at 0.00. Correspondingly, all units with coefficients below the 0 line indicate an advantage of the SI condition (and conversely for coefficients above the 0 line). Conditions: SI = successful intelligence; CT = critical thinking; M = memory; TAU = teaching as usual. Language Arts units: wn = *Wonders of Nature*; tw = *True Wonders*; bio = *Lively Biographies*; jny = *Journeys*; mys = *It's a Mystery*. Mathematics units: fra = *Equivalent Fractions*; mea = *Measurement*; geo = *Geometry*. Science units: lght = *The Nature of Light*; mag = *Magnetism*.

The pattern of influence of the covariates, both alone and as interactions, is varied across interventions (see Table 6, covariates). This pattern attests to the diversity of variables that influence, in complex ways, attempts to scale experimental investigations of intervention effects into everyday contexts. Controlling for these demographic characteristics of the schools and classrooms using the data we have access to, the SI intervention was advantageous in each domain (Language Arts, Mathematics, and Science) but weakly and inconsistently so.

Discussion

Based on the data collected in previous studies and discussed in the introduction, teaching for successful intelligence has been shown to help improve instruction and assessment in a variety of disciplines at diverse grade levels (Grigorenko et al., 2002; Sternberg et al., 1998, 2011). Most important, SI research has helped to provide a way of showing that if students are taught in a way that fits their ability profiles, they will achieve at higher levels and be better able to leverage their diverse skills (Sternberg et al., 1999).

The range of results found in the current study across all units and conditions are discordant with our previous findings. That is, regardless of (a) the rigorous research design, (b) the substantial resources invested by our team of highly skilled researchers drawn from around the world, as well as the numerous classroom teachers who invested time and energy to be involved, (c) the infrastructure available from one of the very best universities in the world in which the project was hosted, and of course (d) the recognition and support of the National Science Foundation (NSF) granting committee who invested in the SI theory to fund this large-scale research project, the results are sobering, especially in light of our previous successes. Because of the investments of the many stakeholders involved with the project, it is incumbent on us to reflect on the implications of these findings in relation to the future of SI theorizing and for educational research that aims to scale up interventions that have previously demonstrated advantages in small, controlled studies. In this regard we first consider the future utility of the "economy of scale" argument, on which large-scale intervention studies are often grounded, and second reflect on the specific implications of scaling the SI intervention relative to the strong control interventions in regard to implementation fidelity.

Economies of Scale: Is It a Viable Approach?

One potential explanation for the observed results is that the attempt to apply economic theories and models to education may be fundamentally misguided. Many policymakers endorse a factory metaphor for thinking about education, in which students are the "products" to be filled with knowledge and teachers are a means of production (see Madaus, Haney, & Kreitzer, 1992, for a description). The microeconomics concept of economies of scale, upon which the notion of scaling up educational interventions rests, has been demonstrated to be highly effective in the manufacturing world (e.g., Henry Ford's assembly line). However, Seddon (2010) has argued convincingly that economies of scale are not applicable in the context of human service professions, and educational delivery is arguably much more closely aligned with human services than with a factory metaphor. Further, as Elias et al. (2003) noted, one of the reasons that scaling up educational interventions is challenging is because educational interventions primarily rely on human operators rather than technologies. Teachers are not automatons that execute a standardized curriculum in a standardized way. Rather, Elias et al. suggest that a more useful metaphor for thinking about scaling up interventions is a sailing analogy in which various elements of the environment can take a

toll on a successful voyage and thus call to the forefront the skill of the sailors in navigating the environment. In addition, given the long history of local control of education in the United States, each state, district, and even school may have a unique cultural, organizational, and educational context (Stemler & Bebell, 2012). Although there is currently a movement toward the development of Common Core Standards in education in an effort to reduce some of the variability in curricular issues, this will not address all of the systemic variability that can impact efforts to scale up educational interventions.

Given the rigor strived for but not necessarily fully attained in the current investigation, our data suggest that it may be time to abandon the illusion that economies of scale should be pursued in the context of educational interventions. Instead, alternate models such as those being embraced by various teacher education programs throughout the country currently appear to us to be more promising. These models take a very different approach in which the implementation is tightly monitored and supported and in which new organizations wishing to join must be evaluated for the relevance of their contextual characteristics.

Implementation Fidelity at Scale: SI Dynamics

Traditional higher level teaching interventions, like training for memory skills, are formidable interventions against which to pit new teaching approaches for a number of reasons. First, traditional, memory-based strategies are the ones teachers may be expected to revert to in uncertain situations (e.g., when attempting to implement a new teaching philosophy for the first time). It takes time for teachers to acclimate themselves to a new philosophy, and two days of teacher training, although the most we could request, simply may not be sufficient. Second, given that the SI condition includes traditional memory and critical thinking aspects, as well as creative and practical ones, it may be possible for teachers to focus on more traditional aspects and still feel they are appropriately adhering to the SI condition. Third, it is important to remember that the unit content was identical across all conditions. The differences between intervention conditions were in the framing of the teacher training, which included differential instruction in the underlying philosophy of SI, M, or CT, as appropriate. Furthermore, the curriculum content across all units and conditions was strong and well structured enough to provide engaging activities aimed at facilitating knowledge acquisition in the specific domain regardless of the intervention framing. Fourth, just as it is expected to take time for teachers to acclimate themselves to the SI philosophy, students also need time to adjust to differences in instruction (Jeltova et al., 2011). Finally, many of the content areas chosen for the units inherently required analytic skills and the memorization of facts. This is certainly true for the Mathematics units and to a lesser extent the Science units. However, it is also true of the Language Arts units.

It also is possible either that the SI model does not work effectively for all the conditions we studied or that our realization of it was less than fully effective. It would take further research to elicit a more definitive answer to such questions.

Limitations

A study such as this one obviously has its limitations. We consider population issues, cost–benefit issues, and teacher and student issues that impact on fidelity of implementations.

Population issues. All students were fourth graders, and only three academic subjects were used. The sheer scale of the study practically ensured that some implementation sites would have higher fidelity than others. In addition, given that the study unfolded in nine states across the country, it was impossible to utilize a single standardized achievement test across all study groups and all domains. A measure of overall achievement (i.e., an end-of-year standardized achievement test) would have provided an alternative test of effectiveness.

Cost–benefit analyses. As innovation and change are costly, a fair question to consider is the cost–benefit analyses that compare the obtained gain in achievement to the costs of introducing a change in instruction. This question has not been the focus of investigation in studies introducing cognitive theories of learning-based approaches to classrooms, and the theory of successful intelligence is not to be excluded. Nevertheless, we are not prepared to conclude just yet that cognitive-based interventions, including those grounded in the theory of successful intelligence, generally do not lead to sufficient enhanced student achievement to be worth the effort. This is in part because the specific advantages of cognitively based interventions may interact with content, school-level variables and the scale of the implementation in complex and dynamic ways.

Insights from the present efforts to upscale an instructional intervention within the context of an experimental study are consistent with those stated in the literature. First, teacher buy-in plays a critical role in the success of any curricular intervention. Throughout the year, teachers inevitably faced many external demands that compromised their ability to complete all of the intended units. Second, when working with intact classrooms, there are potential confounds that can creep into study design. In the case of the current study, there are examples in which the instructional condition is confounded with a particular type of classroom (e.g., gifted classrooms that received memory-based instruction). Such anomalies cannot be co-varied out.

Teacher and student issues. We found perhaps the most challenging aspect of the study to be teachers' differential comfort levels with various instructional methods. Even though teachers were trained in the teaching method they would use, when under stress, we might expect some teachers to revert to what is easiest and most familiar. Under the pressures of day-to-day teaching over the long term, which poses different demands than either teaching for a laboratory experiment or teaching for a short-term study, even teachers who are well trained in a new method may find themselves reverting to older, more familiar methods that they can use without the constant vigilance and concentration required of new interventions. They revert because they are under so many other pressures: classroom management issues, parental pressures, and administrative mandates that they need to confront at the same time. Fidelity to treatment method thus becomes an issue, and such violations of fidelity are particularly difficult to control in the context of a large-scale study such as ours.

A further issue is students' own comfort with different methods of instruction. Students, like teachers, are simply much more familiar with memory-based instruction than with other methods used in the teaching/learning process. Because the students' mental

resources often are split between listening to the teacher, thinking about and planning for events going on in their extracurricular lives, and engaging in the social context of the classroom, they as well may find it easier to relate to traditional teaching than to novel methods of instruction.

Summary

In sum, the results of this large-scale, multistate study suggest that there are difficulties associated with scaling up educational interventions that have been demonstrated to be effective in smaller contexts. Implementation of the curricular materials was designed and implemented with a minimal level of support from the research team, and the student achievement results revealed that the impact of the curriculum on student performance, when compared with strong pedagogical approaches involving teaching for memory and/or critical thinking, as well as with "teaching-as-usual" approaches, was heterogeneous. The results suggest that SI instruction does lead to student achievement outcomes that are, at a minimum, generally equivalent to those associated with other strong instructional interventions. Overall, the effects were weak, and the pattern of influence of the school and classroom covariates on posttest performance differed across interventions and units. Across the domains of literature, mathematics, and science, enhanced student performance was observed in only 7 out of 23 comparisons. SI was advantageous in four cases. There was one case where M was advantageous and two cases in favor of CT.

The traditional approach would be to conduct more rigorous, lab-like investigations into SI effectiveness; consequently, smaller replications of this study in different contexts might be called for. Or, it might be suggested that we investigate our critical thinking and memory interventions more rigorously. However, it is important to recognize that such rigor, by definition, introduces into the investigation constraints that are not feasible in real, intact classrooms—constraints we specifically set out to free in the current study.

It is important to place the data, results, and related discourse presented here in the larger context of the relevant literatures and question whether we as a research group, and the discipline in general, are going about such investigations the wrong way. Should implementation of interventions be tightly monitored and supported and participation eligibility be evaluated for relevance of contextual characteristics? These questions need deep reflection. The following observations seem to be important.

First, even if a particular instructional approach has generated robust evidence pertaining to its efficacy and replication, this does not mean that the scaling it up will be as effective as its more controlled, smaller scale evaluations. We argue that such a diffusion of the promise of an intervention is linked, primarily, to contextual factors, both systematic and random, influencing the context in which the intervention is scaled. This observation is relevant not only to the work presented here but to many other educational interventions. Second, it appears that scaled-up interventions may be characterized by a decrease of effect sizes observed in more controlled evaluations of the efficacy and robustness of an experimental pedagogy. Third, systematic efforts are needed (a) to characterize and parameterize contextual factors that threaten the consistency of an intervention when scaling up and (b) to quantify the expected decrease on previously reported intervention effect sizes. These issues should be factored into the cost–benefit analyses of implementing change in education and should inform policy decision making. In such analyses and decisions, the empirical challenges to an innovation should be considered along with the humanistic and societal values and the ever-changing demands of the labor market. Factors such as these often do not wait for the relevant rigorous studies to be completed in a time comparable to the dynamics of real life.

References

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. New York, NY: Psychology Press.

Bruning, R. H., Schraw, G. J., & Norby, M. M. (2010). *Cognitive psychology and instruction* (5th ed.). Upper Saddle River, NJ: Prentice-Hall.

Clements, D. (2005). Scaling up TRIAD: Teaching early mathematics for understanding with trajectories and technologies [Grant proposal]. Retrieved from the Institute for Educational Sciences website: http://ies.ed.gov/funding/grantsearch

Constas, M., & Sternberg, R. J. (Eds.). (2006). *Translating educational theory and research into practice*. Mahwah, NJ: Erlbaum.

Corno, L., Cronbach, L. J., Kupermintz, H., & Lohman, D. F. (2001). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. New York, NY: Routledge.

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions*. New York, NY: Irvington.

Dehejia, R., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and Statistics, 84,* 151–161. doi:10.1162/003465302317331982

Elias, M. J., Zins, J. E., Graczyk, P. A., & Weissberg, R. P. (2003). Implementation, sustainability, and scaling up of social-emotional and academic innovations in public schools. *School Psychology Review, 32,* 303–319.

Elmore, R. (1996). Getting to scale with good educational practice. *Harvard Educational Review, 66,* 1–26.

Folland, S., Goodman, A., & Stano, M. (2013). *The economics of health and health care* (7th ed.). New York, NY: Pearson.

Francis, D. (2011). Scale-up evaluation of reading intervention for first grade English learners [Grant proposal]. Retrieved from the Institute for Educational Sciences website: http://ies.ed.gov/funding/grantsearch

Fuchs, D. (2004). Scaling up peer assisted learning strategies to strengthen reading achievement [Grant proposal]. Retrieved from the Institute for Educational Sciences website: http://ies.ed.gov/funding/grantsearch

Gardner, H. (1993). *Multiple intelligences: The theory in practice*. New York, NY: Basic Books.

Gardner, H. (2006). *Multiple intelligences: New horizons in theory and practice*. New York, NY: Basic Books.

Gefen, D., Straub, D., & Boudreau, M. (2000). Structural equation modeling and regression: Guidelines for research practice. *Communications of the Association for Information Services, 1*(7), 1–78.

Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S., & Cole, D. A. (2010). Analyzing true change in longitudinal multitrait-multimethod studies: Application of a multimethod change model to depression and anxiety in children. *Developmental Psychology, 46,* 29–45. doi:10.1037/a0017888

Glennan, T. K., Bodily, S. J., Galegher, J. R., & Kerr, K. A. (2004). *Expanding the reach of educational reforms: Perspectives from leaders in the scale-up of educational interventions*. Santa Monica, CA: RAND Corporation.

Grigorenko, E. L., Jarvin, L., Diffley, R., Goodyear, J., Shanahan, E. J., & Sternberg, R. J. (2009). Are SSATs and GPA enough? A theory-based approach to predicting academic success in secondary school. *Journal of Educational Psychology, 101,* 964–981. doi:10.1037/a0015906

Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2002). School-based tests of the triarchic theory of intelligence: Three settings, three samples, three syllabi. *Contemporary Educational Psychology, 27*, 167–208. doi:10.1006/ceps.2001.1087

Grigorenko, E. L., & Sternberg, R. J. (2001). Analytical, creative, and practical intelligence as predictors of self-reported adaptive functioning: A case study in Russia. *Intelligence, 29*, 57–73.

Halpern, D. F. (1996). *Thought and knowledge: An introduction to critical thinking*. Mahwah, NJ: Erlbaum.

Ho, D., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*, 199–236. doi:10.1093/pan/mpl013

Hurtig, R. (2004). Breakthrough to literacy in the Chicago public schools: A large-scale evaluation of the effectiveness of a reading comprehension interventions [Grant proposal]. Retrieved from the Institute for Educational Sciences website: http://ies.ed.gov/funding/grantsearch

Jeltova, I., Birney, D., Fredine, N., Jarvin, L., Sternberg, R. J., & Grigorenko, E. L. (2011). Making instruction and assessment responsive to diverse students' progress: Group-administered dynamic assessment in teaching mathematics. *Journal of Learning Disabilities, 44*, 381–395. doi:10.1177/0022219411407868

Kornilov, S. A., Tan, M., Elliott, J. G., Sternberg, R. J., & Grigorenko, E. L. (2012). Gifted identification with Aurora: Widening the spotlight. *Journal of Psychoeducational Assessment, 30*, 117–133. doi:10.1177/0734282911428199

Madaus, G. F., Haney, W., & Kreitzer, A. (1992). *Testing and evaluation: Learning from the projects we fund. Policy issues in the conduct of corporate support for education*. Washington, DC: Council for Aid to Education.

Mayer, R. E. (2011). Intelligence and achievement. In R. J. Sternberg & S. B. Kaufman (Eds.), *Cambridge handbook of intelligence* (pp. 738–747). New York, NY: Cambridge University Press.

McKenna, M. C., & Walpole, S. (2010). Planning and evaluating change at scale: Lessons from Reading First. *Educational Researcher, 39*, 478–483. doi:10.3102/0013189X10378399

Muthén, L. K., & Muthén, B. O. (2005). *Mplus user's guide*. Los Angeles, CA: Author.

Nunnery, J. A. (1998). Reform ideology and the locus of development problem in educational restructuring: Enduring lessons from studies of educational innovation. *Education and Urban Society, 30*, 277–295. doi:10.1177/0013124598030003002

Ormrod, J. E. (2010). *Educational psychology: Developing learners* (7th ed.). Uppers Saddle River, NJ: Prentice-Hall.

Pane, J. (2007). Effectiveness of cognitive tutor Algebra One implemented at scale [Grant proposal]. Retrieved from the Institute for Educational Sciences website: http://ies.ed.gov/funding/grantsearch

Pashler, H., McDaniel, M., Rohrer, D., & Bjork, R. (2008). Learning styles: Concepts and evidence. *Psychological Science in the Public Interest, 9*, 106–119.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & the R Core team. (2009). nlme: Linear and nonlinear mixed effects models. R package version 3.1-92 [Computer software].

Randi, J., & Jarvin, L. (2006). An "A" for creativity: Assessing creativity in the classroom. *The Thinking Classroom, 7*(4), 26–32.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory. *Journal of Statistical Software, 17*, 1–25.

Samejima, F. (1997). Graded response model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.

Seddon, J. (2010). *Why do we believe in economies of scale?* Retrieved from http://www.vanguard-consult.dk/wp-content/uploads/2011/10/whydowebelieveineconomiesofscale.pdf

Slavin, R. E. (2008). *Educational psychology: Theory and practice*. Needham Heights, MA: Allyn-Bacon.

Spear, L. C., & Sternberg, R. J. (1987). Teaching styles: Staff development for teaching thinking. *Journal of Staff Development, 8*, 35–39.

Starkey, P. (2004). Scaling up the implementation of a pre-kindergarten mathematics curriculum in public preschool programs [Grant proposal]. Retrieved from the Institute for Educational Sciences website: http://ies.ed.gov/funding/grantsearch

Stemler, S. E., & Bebell, D. (2012). *The school mission statement: Values, goals, and identities in American education*. Larchmont, NY: Eye on Education.

Stemler, S. E., Grigorenko, E. L., Jarvin, L., & Sternberg, R. J. (2006). Using the theory of successful intelligence as a basis for augmenting AP exams in psychology and statistics. *Contemporary Educational Psychology, 31*, 344–376. doi:10.1016/j.cedpsych.2005.11.001

Stemler, S. E., Sternberg, R. J., Grigorenko, E. L., Jarvin, L., & Sharpes, D. K. (2009). Using the theory of successful intelligence as a framework for developing assessments in AP Physics. *Contemporary Educational Psychology, 34*, 195–209. doi:10.1016/j.cedpsych.2009.04.001

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York, NY: Cambridge University Press.

Sternberg, R. J. (1995). *In search of the human mind*. Orlando, FL: Harcourt Brace College.

Sternberg, R. J. (1997). *Successful intelligence*. New York, NY: Plume.

Sternberg, R. J. (2003a). *Sternberg Triarchic Abilities Test*. Unpublished manuscript, Yale University.

Sternberg, R. J. (2003b). *Wisdom, intelligence, and creativity synthesized*. New York, NY: Cambridge University Press.

Sternberg, R. J. (2005). The theory of successful intelligence. *Revista Interamericana de Psicología, 39*, 189–202.

Sternberg, R. J. (2009). WICS: A new model for liberal education. *Liberal Education, 95*(4), 20–25.

Sternberg, R. J. (2010). *College admissions for the twenty-first century*. Cambridge, MA: Harvard University Press.

Sternberg, R. J., Birney, D., Jarvin, L., Kirlik, A., Stemler, S., & Grigorenko, E. L. (2006). From molehill to mountain: The process of scaling up educational interventions (First-hand experience upscaling the theory of successful intelligence). In M. Constas & R. J. Sternberg (Eds.), *Translating educational theory and research into practice* (pp. 205–221). Mahwah, NJ: Erlbaum.

Sternberg, R. J., Bonney, C. R., Gabora, L., Karelitz, T., & Coffin, L. (2010). Broadening the spectrum of undergraduate admissions. *College and University, 86*(1), 2–17.

Sternberg, R. J., Castejón, J. L., Prieto, M. D., Hautamäki, J., & Grigorenko, E. L. (2001). Confirmatory factor analysis of the Sternberg Triarchic Abilities test in three international samples: An empirical test of the triarchic theory of intelligence. *European Journal of Psychological Assessment, 17*, 1–16. doi:10.1027//1015-5759.17.1.1

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J., Snook, S., Williams, W. M., . . . Grigorenko, E. L. (2000). *Practical intelligence in everyday life*. New York, NY: Cambridge University Press.

Sternberg, R. J., & Grigorenko, E. L. (2000). *Teaching for successful intelligence*. Chicago, IL: Skylight.

Sternberg, R. J., Grigorenko, E. L., Ferrari, M., & Clinkenbeard, P. A. (1999). Triarchic analysis of an aptitude–treatment interaction. *European Journal of Psychological Assessment, 15*, 3–13. doi:10.1027//1015-5759.15.1.3

Sternberg, R. J., Grigorenko, E. L., & Jarvin, L. (2007). *Teaching for successful intelligence* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Sternberg, R. J., Grigorenko, E. L., & Zhang, L. (2008a). A reply to two stylish critiques. *Perspectives on Psychological Science, 3*, 516–517. doi:10.1111/j.1745-6924.2008.00092.x

Sternberg, R. J., Grigorenko, E. L., & Zhang, L. (2008b). Styles of learning and thinking matter in instruction and assessment. *Perspectives on*

Psychological Science, 3, 486–506. doi:10.1111/j.1745-6924.2008.00095.x

Sternberg, R. J., Jarvin, L., & Grigorenko, E. L. (2009). *Teaching for wisdom, creativity, and success*. Thousand Oaks, CA: Corwin.

Sternberg, R. J., Jarvin, L., & Grigorenko, E. L. (2011). *Explorations of the nature of giftedness*. New York, NY: Cambridge University Press.

Sternberg, R. J., & Lubart, T. I. (1995). *Defying the crowd*. New York, NY: Free Press.

Sternberg, R. J., & The Rainbow Project Collaborators. (2006). The Rainbow Project: Enhancing the SAT through assessments of analytical, practical, and creative skills. *Intelligence, 34,* 321–350. doi:10.1016/j.intell.2006.01.002

Sternberg, R. J., Torff, B., & Grigorenko, E. L. (1998). Teaching triarchically improves school achievement. *Journal of Educational Psychology, 90,* 374–384. doi:10.1037/0022-0663.90.3.374

Sternberg, R. J., & Williams, W. M. (1996). *How to develop student creativity*. Alexandria, VA: Association for Supervision and Curriculum Development.

Woolfolk, A. H. (2009). *Educational psychology* (11th ed.). Upper Saddle River, NJ: Prentice-Hall.

Appendix

Technical Issues in the Handling of Data

This Appendix describes the technical details regarding the handling of data. Participating teachers were instructed to label and package all student materials in a particular way and send the materials to Yale. A set of materials from one test (pre- or post-) from one teacher from one unit was called a "package." For a package to be processed and entered into the database, the student workbook had to meet the fidelity standards (see above). In collaboration with the Yale University Social Science Statistical Laboratory, an ACCESS database template was developed. This template was used to build separate databases for each of the 4 years of data collection. Each database was used to (a) inventory, or log, the materials received from the teachers; (b) track the materials as they were sent to coders to score; and (c) store test data and demographic information. The four databases were housed on the central PACE server, with file access restricted to members of the project team.

Database Structure

The structure of the database was rather complex and contained several types of tables, as described below.

Participant information tables. Four tables contained nontest information about the different participants in the study: students, teachers, schools, and districts. Unique ID numbers were given to each element within a table (e.g., each packet was given a unique ID in the packet table). Each teacher was given a different teacher ID number for each school at which he or she taught during that year; hence, some teachers were given more than one unique ID. In most cases, the information in these tables was entered into the database before any assessment data were collected.

Coding administration tables. Tracking test materials was a particularly challenging part of administering a large-scale project that involved continuous receipt and scoring of tests. Two important database processes were involved. The first process, material logging, was used to inventory the completed assessments received by the PACE Center. The second process, material checkout, was used to track the assessments as they were given to coders to score. Four Access tables were involved, and information was continually added to these tables as part of the material logging and checkout processes. First, upon receiving student materials from the schools and/or teachers, PACE research assistants "logged in" the materials to the Access database for the appropriate year of data collection. The logging process consisted of assigning tests of similar type (e.g., Geometry pretests) from a single teacher to a "packet" and creating inventories of materials received by the PACE Center.

A packet was considered both (a) an envelope containing a collection of one particular test type for all the students associated with one teacher and (b) an Access database unit that identified this collection of student tests. A system of Access forms was developed to allow a research assistant simultaneously to add information to two tables that inventoried and tracked the packets and tests. The first table was a "material" table used as an inventory of test materials received by the center. The second table was a "packet" table that was used to assign a packet number to a packet and track it as it was sent to coders to score. A packet that was successfully logged in was then ready to be rated by a coder. Second, via a system of queries, packets were selected and assigned to coders to rate. A "checkout" table tracked when each coder checked out and returned each packet. In addition, a "coder" table was maintained that contained a list of each coder, his or her unique coder ID, and notes about the coder. Queries and forms were used simultaneously to update these tables as coders completed the agreement process and as packets were assigned to coders.

Data tables. When a coder was assigned a packet to score, she or he was also given a scoring template for the packet: an Excel spreadsheet used to record multiple-choice item responses and open-ended item ratings for each student. These templates contained a row for each student, with columns corresponding to the ratings needed for each item and additional

(*Appendix continues*)

columns containing identifying information (IDs for the packet, student, type of test, and coder). Each scoring template contained columns for only one test type (e.g., Geometry pretests) that corresponded to the columns of an Access data table; data for each test type were stored in separate Access data tables. The coder returned an electronic (Excel) and/or a paper copy of the completed scoring template to the National Science Foundation (NSF) team. If the coder submitted only a paper copy, it was given to data entry personnel to enter into the Excel template (with this latter option reserved for skilled coders who had little computer access or expertise). Information from the completed Excel scoring template was then directly uploaded into an Access data table by copying the data cells of the Excel sheet and pasting them into the Access table.

Quality control. Measures were taken to monitor the quality of the ratings during data collection. These measures included limiting database access to a small number of the most experienced personnel, using data-validation controls to prevent the entry of out-of-range values, supervising the coders carefully after their training was completed, and maintaining problem logs in the database.

Limiting the number of Access users. Access to the database was limited to only a small core of management personnel to ensure participant confidentiality and to minimize the possibility of human error. The databases were stored on a central server and required network permissions to be viewed or modified. For most of the study, only a small number of our most technologically sophisticated personnel were allowed access to the database to check, upload, and clean data. At times, the number of people working simultaneously on coding exceeded 20 trained coders. Rather than having all of these coders enter their ratings into the Access data tables directly, we introduced a middle step between rating and Access data entry. Coders' ratings were entered into Excel, as described above, and then given to the core database managers to upload.

Excel template and Access table validations. Two related measures were taken to prevent the entry of out-of-range values into the Access databases. First, cell validations were used in Excel that would allow coders to enter only legitimate ratings. Legitimate ratings included codes used to designate an omitted or illegible response to an item (i.e., 6 or e for omitted responses, and 7 or f for illegible responses). A second layer of protection was also used to prevent the uploading of empty (unrated or unrecorded) data cells into the Access database and to serve as a second check for out-of-range values. Validation rules were eventually implemented in all Access data tables to prevent the uploading of missing or out-of-range values. When a core NSF research assistant could not upload the data from a coder's template because of an out-of-range or missing value, the paper copy of the template and/or the coder was consulted to find the true value of that rating.

Coder supervision. Coders were not permitted to score tests until they reached an acceptable level of initial interrater reliability with their coding partner (i.e., the correlations between the pair's open-ended item ratings were greater than 0.70). Coders who reached this criterion then began coding tests independently from their partners; coders who were not able to establish acceptable levels of interrater reliability were not permitted to continue on the project. Of the 90 coders who began the training and agreement process, only 76 were permitted to score tests for the study. Core personnel maintained weekly contact with active coders after initial interrater reliability was reached. They maintained the quality of the ratings by being available to answer coders' questions about scoring and by reminding coders of the scoring guidelines when the coder's ratings were discovered to have violated validation rules. The design of the study also allowed for the discovery of coder irregularities throughout the scoring process. As a quality control check, over thirty percent of the tests each coder scored were also scored by another coder. These overlapping ratings were used to detect discrepancies between coders and to flag coders who were having particular difficulty. Data from two of the 76 coders were deleted due to continued discrepancies with other coders. In addition, during the final stage of data cleaning before analysis, a random 10% of codings were spot-checked against the hard copies of the assessments. The 74 remaining coders were diverse with respect to their genders, ages, educational backgrounds, and test coding experience. They ranged in age from 18 to 66 and included research assistants, undergraduate student workers, temporary part-time employees, and PhD-level research scientists. Many coders had previous experience scoring tests, and some had experience creating scoring rubrics for tests. More than 20,000 pre- and posttests including over 400,000 items were read and rated by these raters. All pre- and posttests packets had two raters, who used written rubrics to evaluate the quality of children's responses. Each pair trained together on one packet: The two raters in the pair rated identical tests, their scores were then compared to establish interrater reliability, differences in scoring were pointed out and the rater pair discussed responses until agreement was reached. They were then sent back with the packet and their new ratings checked for reliability. Training was conducted until pairs reached an agreement of .70, which was treated as the minimum acceptable level; when the agreement was reached, the raters read and rated a common overlapping set of materials (representing 1/3 of all the tests rated by the pair) and then each rater read and rated separate sets. The quality of the data and interrater agreement was monitored in an ongoing fashion. Rater biases were carefully evaluated.

Problem note tables. During the last year of data collection, tables were created in Access to centralize notes made on test administration quirks (e.g., a teacher photocopying all but a page of a test). One table was used to record problems with a particular student; another table was used to record quirks that affected the entire classroom. Both tables made note of what was done to correct the problem. These notes were used to ensure that common problems were treated consistently. These tables were used to determine the usability of the data for final analyses.

Received July 27, 2011
Revision received November 4, 2013
Accepted December 29, 2013 ■