The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant R305F100013 to The University of Texas at Austin as part of the Reading for Understanding Research Initiative. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education.

Educ Psychol Rev (2015) 27:31–50
DOI 10.1007/s10648-014-9274-2

REPLICATION

Improving Middle-School Students' Knowledge and Comprehension in Social Studies: a Replication

Sharon Vaughn · Greg Roberts · Elizabeth A. Swanson · Jeanne Wanzek · Anna-Mária Fall · Stephanie J. Stillman-Spisak

Published online: 26 June 2014

Abstract This study aimed to replicate findings that demonstrated impact on students' reading comprehension and social studies content learning. Using a randomized control trial, intervention, and outcome measures, this study was replicated in 85 8th-grade social studies classes with 19 teachers. Teachers were provided professional development on comprehension canopy, essential words, knowledge acquisition, and the use of team-based learning. Measures of reading comprehension administered at pre- and post-testing and measure of vocabulary and knowledge acquisition was administered at pre-, post-, and two follow-up times, 4 and eight weeks following treatment. On the measure of vocabulary and knowledge acquisition, students in the treatment condition outperformed those in the comparison condition at all time points. There were no statistically significant differences for reading comprehension.

Keywords Middle school · Reading comprehension · Social studies · Content learning · Replication

S. Vaughn (✉) · G. Roberts · E. A. Swanson · A.-M. Fall · S. J. Stillman-Spisak
The Meadows Center for Preventing Educational Risk, The University of Texas at Austin, 1912 Speedway, D4900, Austin, TX 78712-1284, USA
e-mail: srvaughn@austin.utexas.edu

G. Roberts
e-mail: gregroberts@austin.utexas.edu

E. A. Swanson
e-mail: easwanson@austin.utexas.edu

A.-M. Fall
e-mail: amfall@austin.utexas.edu

S. J. Stillman-Spisak
e-mail: stephstillman@austin.utexas.edu

J. Wanzek
Florida Center for Reading Research, Florida State University, 1107 W. Call Street, Tallahassee, FL 32306, USA
e-mail: jwanzek@fcrr.org

Background and Purpose

Perhaps the most significant curriculum challenge for secondary social studies/history teachers is the implementation of the Common Core State Standards (CCSS; www.corestandards.org). While content knowledge acquisition has always been the driving focus of teachers' instruction, the CCSS require that teachers broaden the lens to include reading text and improving comprehension in the social studies discipline. Students in the middle grades are expected to: (a) cite text evidence from primary and secondary sources; (b) determine the meaning of words and phrases as used in text; (c) identify author's points of view or purpose; (d) distinguish among facts, opinions, and reasoned judgments in a text; and perhaps the greatest challenge, (e) read and comprehend social studies/history texts in their grade complexity band independently and proficiently.

Accomplishing the goals of the CCSS requires teachers to exercise an increased emphasis on improving students' understanding and learning from complex texts. To determine the extent to which middle-grade teachers typically provide opportunities for students to read and understand text, Swanson et al. (2014) conducted an observation study of the amount and type of text reading that occurred in 137 English language arts and social studies classes. Their findings revealed that text reading occurred in only 38 % of all middle and secondary social studies class periods observed and less than 20 % of the middle-school social studies classes. Furthermore, the amount of time dedicated to text reading consumed 10.4 % of social studies instructional time. We interpret these findings as revealing that middle-school teachers will require considerable adjustments to their current instructional practices to provide the type and breadth of reading opportunities and instruction needed to assure middle-school students meet the expectations of the CCSS.

Assisting middle-school teachers in identifying instructional practices associated with improved outcomes for their students in reading also requires that the practices support the primary function of assuring that students acquire the content knowledge necessary to succeed. Recently, Vaughn and colleagues (2013) described findings from a multicomponent treatment, promoting acceleration of comprehension and content through text (PACT), aimed at meeting the dual purpose for middle-school students of improving content area knowledge acquisition in social studies/history and improving reading comprehension. Social studies teachers implemented PACT in three distinct topical units (Colonial America, Road to Revolution, and Revolutionary War) that were designed to be taught in 10 days but varied somewhat in the actual amount of time it took teachers to cover each unit. Students in the treatment classes performed significantly higher than students in the typical practice comparison classes on measures of content acquisition (ES=0.17), content reading comprehension (ES=0.29), and standardized reading comprehension (ES=0.20). Thus, the study provided evidence that the PACT treatment could meet the purpose of improving content knowledge as well as reading comprehension. The What Works Clearinghouse (WWC) identified this study as a well-implemented, randomized control trial that met the highest possible rating for a study reviewed by WWC—meeting evidence standards without reservations.

The purpose of this study is to replicate the study conducted by Vaughn et al. (2013). We chose to replicate this study for several reasons: (a) the CCSS require middle-school teachers to implement both knowledge acquisition and content reading practices in their instruction, and there are few validated approaches to this integrated instruction, (b) replication studies are of high importance in all scientific inquiry and yet are rarely conducted in education (Yong 2012), and (c) the original study was rated by WWC as high quality and thus provides the opportunity for a high-quality replication that can determine the extent of evidence for the PACT instruction.

Components of the PACT Intervention

Beck and Eno (2012) provided a review of research on social studies instruction identifying two primary ways that the social studies are taught. One prominent instructional approach they identified, mainstream signature pedagogy, relies primarily on the teacher (e.g., using lecture), and the materials (e.g., textbook) as the sources for acquiring knowledge about the content area. The authors also indicated that there was an emerging signature pedagogy, less often practiced in the middle- and high-school grades, which is more student centered and engages learners in historical inquiry. Both of these instructional approaches, mainstream and emerging, require students to use text as sources for better understanding the social studies and form the foundation for the components of the PACT intervention—the treatment that we propose to replicate.

The PACT instructional components are embedded into the teachers' content instruction that is aligned with district and state standards. The components include: (a) a comprehension canopy (CC) to set the stage for the unit, (b) instruction and built in review of essential words (EW) and concepts for the content, (c) knowledge acquisition through critical reading of primary and secondary texts and (d) team-based learning (TBL; Michaelsen and Sweet 2011). The TBL component allows for small teams of students (four to five students) to engage in collaborative discourse around the content and texts in order to construct knowledge and apply learned content in problem-solving or perspective taking activities. There are two elements to the TBL implementation in PACT: comprehension checks and knowledge application.

Comprehension Canopy The CC provides teachers an opportunity to build background knowledge as well as offering a motivating springboard and overarching comprehension question to guide instruction throughout the 10-day unit. On day 1 of the unit, teachers build students' background knowledge by showing a short, high-interest video clip. Prior to viewing the clip, teachers provide students with a purpose for viewing (e.g., "As you watch the video, write two reasons why the colonists called the First Continental Congress."). After the video clip, students engage in either class-wide or small group discussion that addresses the purpose for viewing as well as additional questions posed by the teacher designed to connect prior learning to new content. Second, teachers provide a complex CC question related to the content of the unit (e.g., "Was the American Revolution inevitable? Why or why not?"). For the remaining 9 days of the cycle, lessons begin with a brief review of the CC question, and a short discussion of the new information that informs their answer to the question. The review process for the CC provides students with an opportunity to consider new knowledge gained from the instruction and forms a guide for knowledge not yet obtained. At the end of the 10-day unit, students provide an answer to the CC question.

Essential Words For each 10-day unit, a set of four to five high-utility, high-frequency concepts that address overarching ideas related to the content of the unit (e.g., revenue, protest, natural rights) are covered. On day 1, the teacher introduces each EW to the students using a simplified definition, visual representation, related words, sentences with the EW in context, and question prompts for brief discussion of the word meaning in context (turn-and-talk prompts). Throughout the unit, lessons begin with a brief review of one or more EW. EW are also integrated into texts, TBL comprehension checks, and TBL knowledge application (TBLK) activities.

Knowledge Acquisition During each 10-day unit, teachers provide three knowledge acquisition sessions with text. During these sessions, students are taught key information aligned with district and state standards through the use of primary and secondary text sources. Teachers engage students in text reading for approximately 20 min/session. Text reading takes place in whole class, small group, paired, or individual silent reading arrangements and is coupled with checks for understanding classroom discourse and notetaking that facilitates connections to the CC, EW, and previously learned material.

TBL Comprehension Checks To assess understanding of content and provide teachers with data-based information to guide subsequent instruction, teachers ask students to complete two short comprehension checks (five multiple-choice questions and one open-ended writing question) on days 4 and 6 of the 10-day unit. TBL comprehension checks are designed to assess students' understanding of key content while promoting both individual and team accountability. For each comprehension check the following procedure is used. First, students complete the check individually with no access to text or notes and submit it to the teacher for a grade. This provides the teacher with an opportunity to assess individual student progress. Next, students move into their pre-established, heterogeneous teams of four or five students. This time, students complete the check again with their text and notes, but with two requirements: (1) the team must agree on the same answer, and (2) evidence from the text or notes must be used to support the team's decision. Scratch off answer sheets are used to provide immediate feedback for the team. If the team scratched the correct answer, a star is revealed. If the answer is incorrect, there is no star and the team returns to discuss the question and use text sources to make an alternative selection. During the team-completed comprehension check, teachers monitor progress and provide feedback to teams, encouraging the use of text evidence, active participation from each team member, and productive discourse. The teacher also identifies the content that students misunderstood. In the final phase of the TBL comprehension checks, teachers spend about 10 min reteaching content that students did not previously understand.

TBL Knowledge Application On day 9 of the 10-day unit, students work in their heterogeneous teams of four to five to complete a knowledge application activity designed to clarify, apply, and extend understanding of text and content. These activities require students to articulate new perspectives, solve problems, and present conclusions. First, students read a short text and are given a related assignment that connects to the reading, the content they have learned, and the CC question used throughout the unit. For example, students read “Letter to the Free Society of Traders” by William Penn and list the reasons Penn gave to encourage people in England to settle in the new colony. Students provide two reasons related to geography/climate, economic opportunities, the native population, and the government. Second, students are asked to use the text from the unit as well as their notes to choose a colony and prepare a series of statements to persuade settlers to choose their colony. As groups work, teachers encourage teams to engage in discourse, form their claims, and prepare to present their work to the class. During their presentation, students answer questions from the teacher and peers that further extend their understanding of the content. To close the lesson, the teacher guides the students to address the CC question for a final time and facilitates discussion of the final answer.

Importance of Replication Studies

Replication studies are of high value because they increase confidence and generalizability of findings—a necessary step to assure the practical value of findings. Yong (2012) argues that replication studies are essential as they reduce the likelihood of researcher error, which is increasingly common in applied research. Perhaps the biggest concern about not replicating findings is that research errors can readily go undetected and then become part of research-based educational practice. In fact, the What Works Clearinghouse (2011) requires at least two studies of an intervention to identify whether the intervention has positive effects.

If replication studies are so valuable, why are there so few of them in education? Considerable attention has been focused on the inappropriate reasons why replication studies are not well embraced including: (a) a positive bias towards new findings, (b) scholars are disinterested in confirming previous work, and (c) replication studies may be difficult to publish as they lack the glamour appeal of novel findings (Yong 2012). Makel et al. (2012) argue there are inadequate numbers of replication studies in published research and documented this view by determining that approximately 1 % of studies published in the top 100 psychology journals were replications. These authors, as well as others, are establishing a wave of concern about the need for additional research studies that address replication of findings (Burman et al. 2010).

Research Question

This study aimed to replicate findings from a study of the PACT intervention that demonstrated impact in comprehension and content learning (Vaughn et al. 2013). Specifically, we examined the research question: What are the effects of the PACT intervention in social studies on eighth-grade students' content acquisition, reading comprehension of content text, and general reading comprehension?

Methods

Study Design

A randomized control trial was implemented with randomization occurring within teacher at the class level. An independent researcher who was not part of the research team conducted the randomization of all social studies classes to either treatment or comparison condition blocking on teacher. This design aligned with the one use in the original study (Vaughn et al. 2013) and is considered a very robust investigation of the effectiveness of the treatment because teacher effects are controlled. All students participated in the same social studies content learning in both conditions, and the only variation between treatment and comparison was the application of the PACT instructional practices within the treatment condition.

Participants and Setting

This study was implemented with 19 US History teachers in seven diverse middle schools located in five large school districts in the Southeast and Southwest USA. Each teacher taught between two and six classes of US History for a total of 85 class sections in the study. Included in this study were 47 treatment classes and 38 comparison classes (for teachers with an odd number of classes, the additional class was assigned to the treatment condition).

Teachers The 19 participating US History teachers (ten females and nine males) possessed a bachelor's degree; six teachers also held a master's degree. Teaching experience ranged from 2 to 38 years (M=15.47 years; SD=13.1). Teachers' ethnicity was either White (89.5 %) or Hispanic (10.5 %).

Students Of the 1,487 students (male=712) who consented to participate in the study, 39 % qualified for free or reduced lunch, 4.8 % were classified as limited English proficient (LEP), and 7.9 % of students qualified for special education services. Students' average age was 13.16

Table 1 Student demographics

	Treatment		Comparison	
	n	%	*n*	%
Gender				
Male	399	48.8	313	46.8
Female	391	47.8	339	50.7
Ethnicity				
White	500	61.1	419	62.6
African American	149	18.2	124	18.5
Hispanic	192	23.5	150	22.4
Asian	30	3.7	24	3.6
Native American	77	9.4	61	9.1
Two or more race specified	32	3.9	22	3.3
Free or reduced lunch	319	39	261	39
English language learners	43	5.2	28	4.1
Special education	67	8.2	51	7.6

in the treatment condition and 13.16 in the comparison condition. See Table 1 for additional student demographic data.

Pretest scores for treatment and comparison classes were compared on the outcome measures (Gates–MacGinitie reading comprehension subtest and the assessment of social studies knowledge (ASK)) to determine the effectiveness of randomization. No significant differences between conditions on either of the measures were found at pretest (see Tables 2 and 3).

Intervention Procedures

Teachers provided all students (treatment and comparison) instruction during their regularly scheduled eighth grade US History classes. Teachers delivered three consecutive instructional units to both treatment and comparison classes, each lasting 10 days (Colonial America, the Road to Revolution, and the Revolutionary War). Classes met daily for 50–55 min or every other day for 90-min periods. Teachers implemented the 30 classes over the course of 6 to 10 weeks. Students in comparison classes received instruction that would typically occur in an

Table 2 Pretest and posttest means, standard deviations, and ranges for reading outcomes

	Pretest				Posttest			
Measures	M	SD	*n*	Range	M	SD	*n*	Range
Gates–MacGinitie reading comprehension								
Treatment	103.71	15.2	766	65–135	101.61	16.1	756	65–135
Comparison	105.49	14.9	635	65–135	103.14	15.4	599	65–135
ASK reading comprehension in social studies								
Treatment	10.84	4.37	775	1–21	11.84	4.73	743	1–21
Comparison	10.85	4.29	633	2–21	11.90	4.60	605	2–21

Note. ASK assessment of social studies knowledge

Table 3 Means, standard deviations, and ranges for ASK knowledge acquisition

	Pretest (n=1,418)			Posttest (n=1,374)			Follow-up 1 (n=1,280)			Follow-up 2 (n=1,288		
	M	SD	Range	M	SD	Range	M	SD	Range	M	SD	Range
Treatment	19.68	7.18	2–40	26.85	8.91	5–41	26.19	9.22	5–42	26.37	9.28	2–42
Comparison	20.59	7.47	1–40	25.02	8.38	4–41	24.37	8.87	1–41	24.46	8.94	6–41

Note. ASK assessment of social studies knowledge

eighth-grade US History class addressing the same material as the units taught in the treatment classes. The same content, aligned with standards identified by the school districts, was covered over the same period of time in both treatment and comparison classes. Therefore, all students were provided the same opportunity to learn the content from within the standard curriculum regardless of random assignment to treatment or comparison classes. The contrast of interest was in the delivery of the content and not in the content/curriculum provided. Thus, treatment classes were provided the content using the PACT components described previously in the introduction and the comparison classes received the same content with a business as usual implementation.

Description of the Treatment Instruction

The components of the treatment are described in the original study (Vaughn et al. 2013) and in the introduction to this paper. In summary, components include: (a) a CC to provide an overall organization for understanding each unit, (b) instruction and review of EW and concepts for understanding text and learning the content, (c) knowledge acquisition through text reading of primary and secondary sources, and (d) TBL (Michaelsen and Sweet 2011), including comprehension checks and knowledge application.

Professional Development and Teacher Support Before the school year started, teachers participated in an intensive 1-day (8 h) professional development workshop covering implementation of the PACT intervention and study design features related to maintaining a firewall between treatment and comparison conditions. The following topics were covered during professional development: (a) implementation of intervention components, including CC, EW, knowledge acquisition, TBL comprehension checks, and TBLK activities, (b) procedures to facilitate student use of discourse and text evidence to support claims, (c) a detailed description of the experimental study design with emphasis on the importance of fidelity to intervention materials, and (d) a discussion of ways to ensure treatment instruction was not implemented in the comparison classes.

In addition to the PACT workshop, teachers were provided in-class coaching support in their treatment classes during implementation of the three units. Coaching support was based on need but was provided at minimum once per week for each teacher. Support took the form of modeling, co-teaching, monitoring student work during teacher instruction, observation and feedback, and support in planning for lessons.

Implementation of the Treatment Teachers were asked to deliver three, 10-day units of study that consisted of the five related intervention components (CC, EW, knowledge acquisition, TBL comprehension check, and TBLK) embedded within their content instruction. Teachers were provided with semi-scripted lesson plans and a daily schedule identifying the components to be delivered on each day of the 10-day unit. Students received a set of student materials that included EW logs and copies of the reading passages with embedded stopping points to take notes. They also received all student materials for the TBL comprehension checks and knowledge application activities.

Observation of Fidelity of Treatment and Comparison Classes

Traditionally, observations of fidelity are used to report the extent to which the intervention was implemented as intended (Swanson et al. 2013). However, fidelity measures can be used to investigate the extent to which the components of a treatment are evident across conditions (e.g., Swanson et al. 2013) providing an estimate of "achieved relative intervention strength" (Hulleman and Cordray 2009, p. 88). In this study, fidelity is incorporated into multilevel analyses to more fully document the relationship between intervention implementation and student outcomes.

Fidelity of Treatment Treatment classes were audio recorded to measure fidelity of intervention implementation. A multistep procedure was used to collect and code audio recordings. First, one treatment class per teacher was randomly selected. During the identified class periods, the teacher audio recorded every lesson that took place over the span of the three, 10-day units. This totaled 30 treatment audio recordings per teacher. A trained research assistant listened to each audio to identify the components contained within each recording. Two recordings of each intervention component were randomly selected, resulting in a set of ten treatment audio recordings per teacher that were evenly distributed over the three units of instruction.

The fidelity tool was designed to determine the extent to which teachers implemented the key components of the treatment: CC, EW, knowledge acquisition, TBL comprehension check, and TBLK. Coders who were blind to treatment and comparison conditions, reviewed the tapes and assigned a rating for each component to identify the alignment of the implementation with the intended intervention. In addition, three global quality ratings were assigned based on the entire observation: (1) overall instructional quality, (2) overall classroom management, and (3) overall implementation of the PACT intervention. A Likert-type scale ranging from 1 (very low alignment with intervention procedures or very low quality) to 5 (very high alignment with intervention procedures or very high quality) was used for each rating.

Coder Training and Reliability Inter-rater reliability on the fidelity tool was established using a gold standard method (Gwet 2001). Four researchers independently coded a pre-identified audio-recorded lesson using the fidelity tool. Codes were compared with a gold standard that was established on the same audio recording by two senior researchers on the team. Subsequent audios were coded until inter-rater agreement of codes was 90 % or higher.

Table 4 presents fidelity data for each of the components observed in the treatment classes. The data indicate that components were implemented with relatively high levels of fidelity. Teachers experienced most difficulty reaching a level of "high implementation" in the knowledge acquistion and TBLK portions of the intervention. Overall, these data suggest that the intervention was implemented with at least medium–high implementation in the treatment classes.

Instruction in Comparison Classes Members of the research team assured that there was a firewall between the instruction provided in treatment vs comparison classes by providing checks and feedback to teachers. This ongoing process of clarification facilitated the integrity

Table 4 Frequency for fidelity observations in treatment and comparison classrooms

Implementation	CC				WU				TBLC				EW				KA				TBLK	
	n (35)	%	*n* (167)	%	*n* (45)	%	*n* (167)	%	*n* (37)	%	*n* (167)	%	*n* (37)	%	*n* (167)	%	*n* (37)	%	*n* (167)	%	*n* (34)	%
Treatment classrooms																						
4=high	12	34.3			24	53.3			2	5.4			16	43.2			0	0			1	2.9
3=mid-high	11	31.4			9	20.0			26	70.3			12	32.4			13	34.1			10	29.4
2=mid-low	9	25.7			7	15.6			8	21.6			6	16.2			19	51.4			19	55.88
1=low	3	8.6			5	11.1			1	2.7			3	8.1			5	13.5			4	11.8
Comparison classrooms																						
4=high			0	0			0	0			0	0			0	0			0	0	0	0
3=mid-high			0	0			0	0			0	0			0	0			0	0	0	0
2=mid-low			0	0			0	0			0	0			1	.60			0	0	0	0
1=low			4	2.4			14	8.38			0	0			8	4.79			16	9.58	4	2.4
0=not observed			163	97.6			153	91.6			167	100			158	94.6			151	90.42	163	97.6

Note. CC comprehension canopy, *WU* warm-up, *TBLC* team-based learning comprehension check, *EW* essential words, *KA* knowledge acquisition, *TBLK* team-based learning knowledge acquisition of the treatment and the separation of comparison class practices. Comparison classes were also audio recorded using the same procedures as the treatment classes to identify evidence of use of treatment components in comparison classes. One comparison class per teacher was identified and audio recorded over the course of the three, 10-day treatment units. From the set of 30 comparison audio recordings per teacher, ten consecutive days of instruction provided in comparison classrooms were chosen for coding. The same fidelity tool used to code treatment audio recordings was used to code recordings of comparison classes. Inter-rater reliability on using the fidelity tool for comparison class coding was established using the same procedures as well.

Table 4 presents data collected from comparison audio recordings. Data indicate that intervention components were rarely observed in comparison classrooms. When they were observed, the component was recognizable at the very lowest level of implementation. For example, all or almost all elements of the CC were observed in 65.7 % of treatment classes and were never observed in comparison classes. This trend of relatively high levels of implementation in treatment classes and almost no implementation in comparison classes is repeated for all components of the intervention.

Measures

We used the same measures of impact that were used in the original study (Vaughn et al. 2013). The ASK (Vaughn et al. 2013) and the Gates–MacGinitie reading comprehension subtests (MacGinitie et al. 2006) were administered to students in the treatment and comparison groups prior to and immediately following treatment. Trained research personnel who were uninformed of the condition (treatment or comparison) to which students were assigned administered all assessments.

ASK (Vaughn et al. 2013) The ASK assessment is a researcher-developed measure that includes two subtests. The first is a 42-item, four-option, untimed multiple-choice test that measures content knowledge in the three units that comprised the intervention (Colonial America, Road to Revolution, and Revolutionary War). With permission, items were collected from released Texas state social studies tests (Texas Assessment of Knowledge and Skills), released Massachusetts state social studies tests (Massachusetts Comprehensive Assessment System), and released advanced placement tests in social studies from the College Board. Researcher-developed vocabulary items were also included in the item set. The ASK knowledge acquisition measure was administered at pretest, at posttest, 4 weeks subsequent to posttest, and again 4 weeks later.

The second subtest is a 20-item, four-option, untimed multiple-choice test that measures reading comprehension. The assessment consists of three reading passages (Lexile range= 1,090–1,140; word count range=312–349), each of which is related to content covered in the three 10-day cycles. Students read each passage silently and immediately answer seven multiple-choice questions about the passage. Reading comprehension items were researcher-developed and measured students' ability to identify main ideas, understand vocabulary in context, identify cause and effect, and summarize. The ASK reading comprehension measure was administered at pretest and posttest.

Alpha coefficients for the ASK knowledge acquisition and reading comprehension measures were 0.89 and 0.85, respectively. However, dimensionality is better indicated by the fit of the underlying confirmatory model in an item response context. Classical test statistics, like alpha coefficients, are most meaningful in a classical framework. In this case, questions about the extent to which a set of items measure the same underlying construct is directly addressed by the model-fitting process described in Vaughn et al. (2013) and replicated in this study.

Gates–MacGinitie Reading Comprehension Subtest (4th Edition) (MacGinitie et al. 2006) The Gates–MacGinitie reading comprehension subtest is a group-administered, timed (35 min) assessment of reading comprehension. The assessment consists of expository and narrative passages ranging in length from 3 to 15 sentences. Students read each passage silently and answer three to six multiple-choice questions related to the most recently read passage. As the students progress through the assessment, items increase in difficulty. Internal consistency reliability ranges from 0.91 to 0.93, and alternate form reliability is reported as 0.80 to 0.87.

Analysis

We fit a preliminary series of multilevel confirmatory factor models to check validity of the ASK measures. Factor indicators were binary (correct/incorrect item response); variance for the latent variable was fixed at 1; and item/factor loadings and threshold estimates (parallel to intercepts in a categorical model) were freely estimated across individuals using a robust maximum likelihood estimator according to missing data theory (Muthen et al. 1987). This specification is analogous to a two-parameter logistic (2PL) item response model which yields indices of overall fit as well as estimates of item difficulty and item discrimination that are useful for comparing the obtained and implied covariance matrices, evaluating the maximum likelihood minimization function, and weighting items based on difficulty. We also fit preliminary factor models for the implementation data to establish estimates of treatment fidelity and treatment crossover. Protocols for coding fidelity data were based on the PACT program model (*intended model*), which represents teacher behaviors that are necessary and unique to PACT as well as the causal mechanisms that hypothetically link activities to anticipated student outcomes. The intended model is the PACT intervention program referred to as the treatment program. The implemented (or enacted) model represents variations of the PACT program presented to students in classrooms. Recognizing that elements of PACT could crossover into comparison classes, threatening internal validity, we evaluated teacher behaviors in both treatment and comparison classes in terms of PACT program elements. Scoring audio files against a common benchmark (the PACT intended model, in this case) provided a vehicle for explicitly indexing the extent of fidelity and crossover. Correspondence between the intended and enacted program in treatment classes was the basis for conceptualizing fidelity, addressing the question: To what extent did the teacher implement program elements as intended in the treatment classes? Within a given teacher, the degree of alignment between the intended program and the enacted program in the comparison classes represented treatment crossover (To what extent did the teacher implement program elements in the comparison condition?), which was a primary threat to the design's internal validity.

The confirmatory models for the ASK subtests and for the model of implementation (i.e., fidelity and crossover) were conceptualized in terms of a larger multilevel model predicting student outcomes, represented in Fig. 1. Implementation (see "Results" for additional details) is modeled as mediating the effect of treatment on students' acquisition of content and reading comprehension. We assumed that assignment to condition predicted student's exposure to treatment and that level of treatment (i.e., enacted program) caused differences in average outcomes across the two conditions. Model parameters were fit to reflect the nested nature of the data and stratification of the overall design (i.e., students within classes, classes within

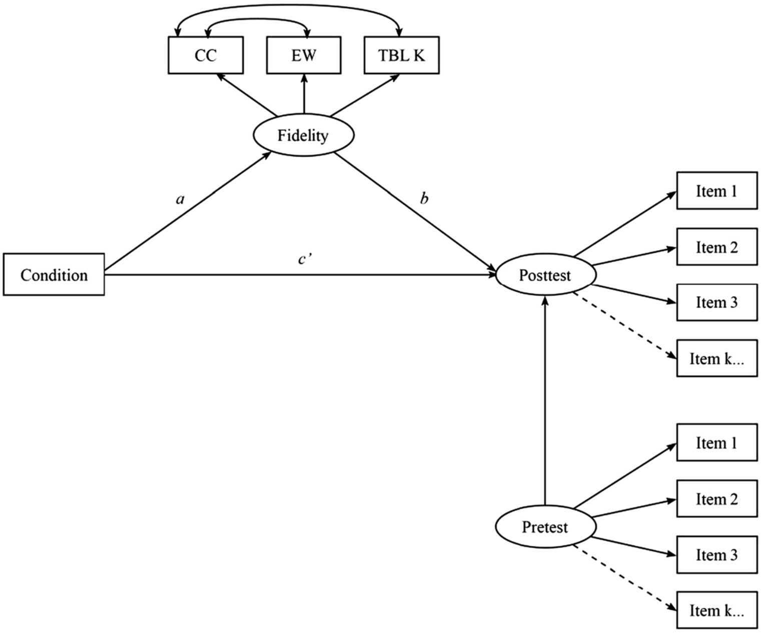

Fig. 1 Hypothetical model for mediating effects of implementation. *CC* comprehension canopy, *EW* essential word introductory routine, *TBLK* team-based learning knowledge application. The model was fit as a MIMIC model with condition coded as 0=comparison and 1=treatment

teachers). Latent factors were estimated as continuous, with means of 0. Effect sizes for group differences (Choi et al. 2009) were calculated as follows:

$$\widehat{d} = \frac{\left|\widehat{K}_1 - \widehat{K}_2\right|}{\sqrt{\left(\frac{n_1}{n_1+n_2}\right)\widehat{\phi}_1 + \left(\frac{n_1}{n_1+n_2}\right)\widehat{\phi}_2}} \tag{1}$$

where $\widehat{K}_1$ and $\widehat{K}_2$ represent latent means for the intervention and comparison groups; n_1 and n_2 the intervention and comparison group sample sizes; and $\widehat{\phi}_1$ and $\widehat{\phi}_2$ the variance of the latent variable for the intervention and comparison groups, respectively. The estimate of effect can be interpreted as in univariate analyses. For example, a value of $\widehat{d}$ = 0.25 indicates that the two population means on the latent construct η are estimated to be one fourth of an error-free (i.e., latent) standard deviation apart along the latent η continuum (Choi et al. 2009).

Finally, in randomized designs, measurement invariance is a particular concern at pretest, as a means of unconfounding structural differences, and as a check on group comparability and the success of the randomizing procedure. Temporal measurement invariance may be of interest, as well, when multiple data points are involved, as in a pretest/posttest design. However, in intervention studies, measurement invariance across time is complicated by the presence of treatment and the possibility that its effect may be differentially evident in parameters of the posttest measurement models for treatment and comparison groups, particularly in cases where test items are aligned with elements of intervention. To address this possibility and to avoid the potential confound it suggests, we constrained measurement parameters as equal across time and across groups at posttest, effectively "forcing" differential variance, related presumably to treatment, into the structural model at time 2. Parameter differences were evaluated using nested model comparisons of scaled log likelihood values (Muthén and Muthén 2007), according to the following:

$$\mathrm{TRd} = -2(\mathrm{LL}_0 - \mathrm{LL}_1)/\Delta_\mathrm{c} \qquad (2)$$

where the difference in log likelihood (TRd) equals the difference in values for the nested (H_0) and comparison (H_1) models, respectively, multiplied by −2, and standardized according to the estimated difference in scaling correction factors (Δ_c). TRd is distributed as χ^2.

Results

We present findings by type of effect, first describing the main effects on the ASK and on the Gates–MacGinitie, then describing the mediating effect of implementation.

Main Effects on ASK and Gates–MacGinitie

Full-sample confirmatory factor analyses (i.e., the two-parameter logistic item response theory models) were fit on the pretest data to evaluate the degree to which the hypothesized models represent the observed data (i.e., absolute fit). In each, latent variables were estimated as continuous with a mean of 0 and variance of 1.0. All item/factor loadings and thresholds were estimated across individuals. The fit was very good for the ASK knowledge acquisition test: χ^2=967.389, *df*=819, *p*=.002, comparative fit index (CFI)/Tucker–Lewis index (TLI)=0.97/0.96, root mean square error of approximation (RMSEA)=.011, 90 % confidence interval (CI)=0.008–0.014. The model for the ASK proximal reading comprehension test also fit the data extremely well: χ^2=312.855, *df*=189, *p*=.001, CFI/TLI=0.96/0.96, RMSEA=0.022, 90 % CI=.017–0.026. These results suggest that the investigator-developed measures represent the hypothesized latent factors (knowledge acquisition and reading comprehension in social studies) per the pattern of observed responses. These results also support the use of scaled log likelihoods to compare nested models and evaluate parameter differences. Adequate absolute fit minimizes the possibility of basing inferences on statistically significant differences between two poorly fitting models.

ASK Knowledge Acquisition in Social Studies To evaluate invariance at pretest, models were refit as latent-class mixture models to accommodate the presence of missing item-level data in multigroup structure. Model 1a in Table 5 represents the configural model for the ASK knowledge acquisition test with no equality constraints across the treatment and comparison groups. The metric model (Model 2a in Table 5), including constraints on item/factor loadings across groups, did not differ statistically significantly from the configural model (TRd=56.66, *p*=0.06), suggesting no group differences in measurement-level factor loadings. Differences in the fit of the metric and scalar models (TRd=41.21, *p*=0.51) also did not differ statistically from 0, which indicates that intercept (or threshold) differences did not differ statistically significantly across groups per each item and, combined with the relative fits for the configural and metric models, establishes full measurement invariance at pretest on the ASK knowledge acquisition. With the measurement models constrained as equal, pretest latent means did not differ across groups (*p*=0.46), confirming pretreatment comparability and supporting the internal validity of the design, at least for the purposes of the ASK knowledge acquisition test.

Table 5 Measurement invariance testing at pretest across groups

Model	−LL	*P*	*c*	Model comparison	*TRd*	Δ*P*	*p* value
ASK knowledge acquisition							
1a. Configural invariance	−36,039.99	170	1.3457				
2a. Metric invariance	−36,073.81	128	1.3956	2a. vs 1a.	56.66	42	0.06
3a. Scalar invariance	−36,100.26	87	1.4486	3a vs 2a.	41.21	41	0.51
ASK reading comprehension in social studies							
1b. Configural invariance	−18,218.36	86	1.4536				
2b. Metric invariance	−18,227.75	65	1.5383	2b. vs 1b.	15.75	21	0.78
3b. Scalar invariance	−18,234.94	45	1.7200	3b vs 2b.	12.74	20	0.88

Note. ASK assessment of social studies knowledge. *c* scaling correction factor, −*LL* log likelihood value, *P* number of parameters, *TRd* log likelihood Chi-square difference test

Main treatment effects on the ASK knowledge acquisition test were estimated by fitting the model in Fig. 2, with students nested in class, classes blocked on teacher, and measurement model parameters constrained as equal across groups and across time. Pretest latent means were fixed at 0 across groups given the presence of structural invariance. Pretest variance was fixed at 1.0 in both groups, and posttest latent means and variances were freely estimated. At posttest, the latent means were 0.787 for the comparison and 1.33 for the treatment, a difference ($\widehat{K}_1 - \widehat{K}_2 = 0.55$) that was statistically significantly greater than 0 ($z=6.33$, $p=0.000$). The latent mean effect size (d) was 0.32, suggesting an average improvement among students in treatment of about 32 % of a standard deviation. At the first follow-up, time 3 in Fig. 3, the difference in latent means was 0.50 ($p<0.001$) and the effect size was 0.29. At the second follow-up, group means differed by 0.46 ($p<0.001$). The effect

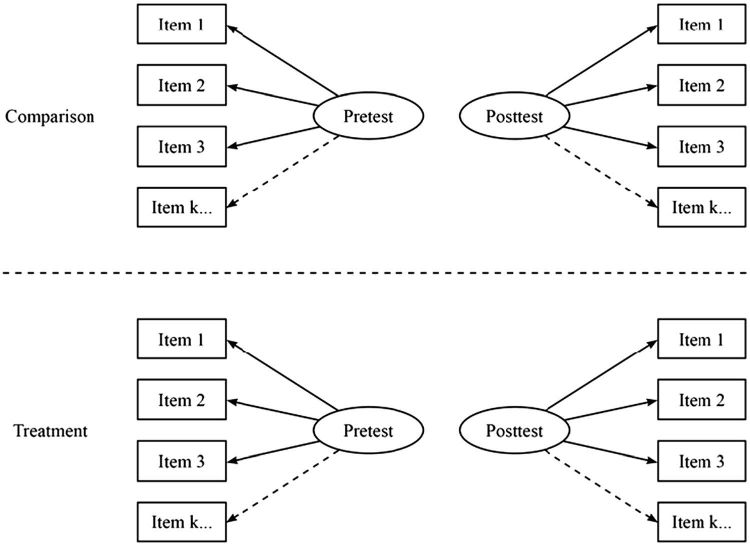

Fig. 2 Hypothetical model for main effects of treatment. The model was fit as a multigroup model

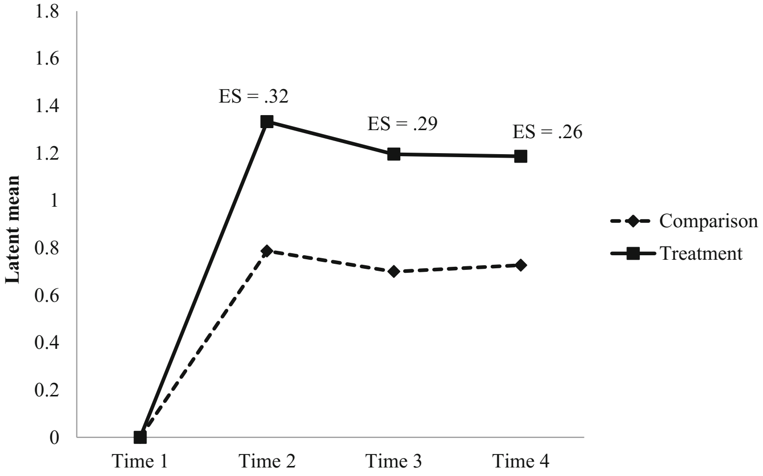

Fig. 3 Latent mean differences in ASK knowledge acquisition at three time points. The vertical axis represents the estimated number of additional items scored correctly at each time point compared with time 1, which is set at 0. Effect sizes reflect group differences in latent means at each time point when estimated with time 1 as 0

size was 0.26. The time 1 latent mean is estimated as 0. Values at subsequent time points, including means for each group and differences in means across the two groups, should be interpreted in relation to time 1.

ASK Reading Comprehension in Social Studies Full measurement invariance was also evident for the ASK reading comprehension test (p values of 0.78 for metric/configural and 0.88 for scalar/metric comparisons; see Table 5), and there were no group differences on pretest latent means (p=0.82). The latent posttest means were 0.276 and 0.307 for comparison and treatment, respectively. The difference ($\widehat{K}_1 - \widehat{K}_2 = 0.031$) was not statistically significant (z= 0.46; p=0.64), and the standardized effect size was 0.02.

Gates–MacGinitie Reading Comprehension Subtest We fit a hierarchical linear model (MLwIN 2.23; Rasbash et al. 2004) with students nested in classes and classes nested in teachers to estimate the main effect of treatment on the Gates–MacGinitie reading comprehension subtest. Pretest scores at level 1 of the model were grand-mean centered (Enders and Tofighi 2007). The effect of treatment (0=comparison; 1= intervention) was modeled on level 2. We calculated effect sizes as Hedges' g, using the coefficient corresponding to the intervention effect as the numerator and the posttest unadjusted pooled standard deviation as the denominator. The Gates–MacGinitie is a standardized, normed measure, and we administered alternate forms at pretest and posttest.

There were no statistically significant pretest differences between students in the treatment and comparison groups, suggesting pretreatment equivalence on the Gates–MacGinitie (β=−3.85, SE=3.68, p=0.30). Average posttest scores did not differ across condition (β=.25, SE=1.54; p=.87). The effect size was 0.01. Random effects in the model differed statistically significantly from 0 at all three levels, with 13.9 % of the total variance at the teacher level, 17.9 % at the class level, and 68.2 % at the student level.

Mediation Analysis

Figure 1 depicts the mediating effect of implementation on the relationship of treatment and ASK knowledge acquisition and reading comprehension. Within each class, mean scores were calculated for each of the program elements across occasions and conditions and these values were used as input in the confirmatory model. Three program elements comprised the structural model for implementation: CC, EW, and TBLK. Models with the other program elements included (or subsets of these elements) did not converge or were less reliably aligned with the observed data. The bivariate correlations between this latter set of elements (those not included in the model in Fig. 1) and improved student knowledge acquisition or reading comprehension from pretest to posttest were not statistically significant from 0 in the total sample, in the comparison group, or in the treatment group, supporting the notion that they were less necessary to estimates of fidelity or to explanations of observed treatment effects.

In Fig. 1, the path labeled *a* represents the effect of experimental condition on fidelity, and the *b* path corresponds to fidelity's effect on posttest scores when controlling for pretest differences. Assuming *a* and *b* are unstandardized regression weights, the indirect effect of condition on posttest scores through fidelity can be represented as *ab*, the product of the *a* and *b* paths. The total effect of condition on posttest scores can be expressed as the sum of the direct and indirect effects: $c=c'+ab$. Hypothetically, we expect implementation to fully mediate the effect of treatment on knowledge acquisition and reading comprehension outcomes to the extent that the intended or normative program is in fact effective. Complete mediation would be indicated by a nonsignificant c' path in Fig. 1 when *ab* is added to the model (i.e., the total effect is explained by treatment's effect through fidelity). An absence of mediation or evidence of incomplete mediation would suggest one of several possibilities: a poorly implemented program model, a well-implemented but ineffective program model, or a combination of the two. Results for the mediation analysis are in Table 6. We calculated a 95 %

Table 6 Estimates and standard errors for paths a and b and the indirect effect

Effect	Unstandardized coefficient (SE)	*p* value	95 % CI
ASK knowledge acquisition			
a	2.97 (0.10)	0.000	(2.77, 3.17)
b	0.57 (0.25)	0.025	(0.07, 1.06)
c′	−1.14 (0.77)	0.14	(−2.64, 0.37)
a × *b*	1.68 (0.76)	0.027	(0.24, 3.15)[a]
c	0.54 (0.09)	0.000	(0.37, 0.71)
ASK reading comprehension			
a	2.98 (0.10)	0.000	(2.78, 3.18)
b	0.10 (0.05)	0.055	(−0.00, 0.19)
c′	−0.26 (0.14)	0.076	(−0.54, 0.03)
a × *b*	0.28 (0.15)	0.057	(−0.01, 0.58)[a]
c	0.03 (0.07)	0.69	(−0.10, 0.14)

Note. SE standard error, *CI* confidence interval

$*p<0.05$; $**p<0.01$; $***p<0.001$

[a] Confidence intervals for the indirect effect are based on the Monte Carlo method (available at http://www.quantpsy.org)

CI for indirect effects using a Web-based Monte Carlo calculator (Selig and Preacher 2008) available at http://www.quantpsy.org based on 20,000 simulated draws from the distributions for the a and b parameters.

For ASK knowledge acquisition, the direct effect c was significant ($p<0.001$) in the model without the indirect path through fidelity. When added, the indirect effect was statistically significant ($p=0.027$) and the direct effectc' was not ($p=0.14$), suggesting that fidelity, as conceptualized and measured here, mediated the effect of treatment; elements of the intended program were evident in the treatment but not in the comparison classes and the enacted program had an impact on students' content acquisition. Table 6 also presents results of the mediation analysis for ASK reading comprehension, although there is no statistically significant main effect for the mediation analysis of reading comprehension.

Discussion

Identifying instructional practices that are feasible to implement, align with the CCSS, and most importantly improve content learning and reading for understanding are essential for improving overall access to post secondary opportunities. In 2013, Vaughn and colleagues reported findings from a randomized control trial examining the effects of a multi-component set of instructional practices (e.g., CC, EW, knowledge application, and TBL) on the content learning and reading comprehension of eighth grade social studies students. Findings indicated that students who were provided the treatment demonstrated statistically significantly improved outcomes in reading comprehension and content learning when compared with students not participating in the treatment but provided typical instruction over the same content.

We conducted a replication study recognizing that replication studies are essential in improving confidence about instructional practice and are infrequently available in educational research. Our goal was to determine whether we could replicate the findings and establish the effects of the PACT intervention on the content acquisition and reading comprehension of eighth grade students in social studies.

This replication reported statistically significant findings for knowledge acquisition in favor of the treatment condition. The latent mean effect size (d) was 0.32 indicating that students in the treatment, on average, outperformed students in the comparison condition by about 32 % of a standard deviation. This effect is considered robust particularly as the same content was provided to students in both the treatment and comparison conditions and was provided by the same teachers (classes were randomly assigned to treatment and comparison conditions). The effect size for knowledge acquisition in this study compares favorably with the previous study with this replication study yielding overall greater impact on treatment participants (i.e., study 1, ES=0.17; study 2, ES=0.32). The consistent effects in two randomized studies of two different samples, suggests confidence in the positive effects of PACT on students' vocabulary and content knowledge in social studies.

Another question of interest is the extent to which participants maintained this learning over time. To address this question we provided follow-up testing at two times: within 4 weeks of posttesting and then again approximately 8 weeks after posttesting to determine overall long-term retention of knowledge. The significant and sustained effects of the treatment on knowledge acquisition at both a 4-week (ES=0.29) and an 8-week (ES=0.26) follow-up assessment are meaningful. Notably, both the treatment and comparison groups maintained their learning from the three units over time, with the treatment group maintaining their advantage from the PACT implementation. The fact that students were generally able to retain the knowledge gained during the three units, may be attributed to distributed practice, a feature of effective instruction known to boost learning (Bjork et al. 2013; Cepeda et al. 2006; Dunlosky 2013). Distributed practice is part of traditional history instruction whereby critical constructs and ideas that are taught in previous units are revisited and connected in the story and sequence of teaching history. Thus, it is likely students continued to revisit in subsequent units the vocabulary, concepts, and knowledge they learned in the initial units allowing for retention of the material over time. The treatment group had gained significantly more knowledge during the units, and retained this advantage over time. Furthermore, the long-term findings in two separate studies (Vaughn et al. 2013; this study) provide practitioners with additional evidence that PACT implementation can improve student content knowledge in social studies.

Unlike the original study (Vaughn et al. 2013), this study did not yield statistically significant differences on either measure of reading comprehension. As a result, there is not yet consistent and generalizable evidence that the PACT intervention significantly improves students' reading comprehension over typical instruction. Further research, including perhaps more extensive implementations, are needed to confirm for whom and under what conditions the PACT intervention has positive effects in the area of reading comprehension.

We also used the fidelity documentation in both the treatment and comparison classes to index treatment implementation and crossover between classes (i.e., the extent to which teachers used elements of the PACT treatment in comparison classes). These data suggested the enacted treatment was implemented within a middle–high range, and there was little to no crossover of the PACT elements in the comparison classes confirming the internal validity of the design. These data and their relationship with student outcomes were further examined in a mediation model. The three PACT treatment elements that best fit the model and were reliably aligned with the observed data were: CC, EW, and TBLK. This model confirmed that the enacted treatment (i.e., CC, EW, and TBLK) caused improvement in students' content learning.

The emerging pedagogy aspects of the PACT intervention may be less familiar and comfortable for social studies teachers than mainstream pedagogy (Beck and Eno 2012). In a survey of social studies teachers' opinions about effective instruction (Bolinger and Warren 2007), 68 % of the secondary teachers thought that lecture was the most effective instructional practice, and identified it as the most frequently implemented. Students also report experiencing largely lecture and related notetaking in their middle- and high-school social studies classes (Chiodo and Byford 2004; Swanson et al. 2014). These more passive approaches to instruction contrast with the Gamoran and Nystrand (1991) research that indicates the combination of student participation, activities, and discourse in social studies classes are associated with significantly improved outcomes in content acquisition. Many secondary social studies teachers struggle with more active approaches to engaging students around text and learning. However, our replication study provides converging evidence of the promise of the PACT practices, including student engagement in discussion, critical thinking, and reading of text in the content, for increasing content knowledge in the social studies. The findings suggest the effort teachers must apply to incorporate more emerging pedagogies in their instruction can have practically important benefits for their students. Given that the same teachers provided both the PACT and traditional approach (randomization at the class level), it could be that additional experience and support for less traditional approaches to instruction (e.g., PACT) are needed to realize even greater impact. However, it is impressive that with respect to vocabulary and knowledge acquisition (not reading comprehension), students in the treatment group outperformed comparisons at all three time points: posttest, 4-week follow-up, and 8-week follow-up.

Limitations Both studies, the original and the replication, were conducted in school district located in the Southeast and Southwest areas of the USA limiting generalization to the sample of teachers and students from these regions. Teachers participating in this study were compensated for the additional time for professional development and working with coaches. Awareness that they were participating in a research study as well as compensation for their efforts may have influenced their dispositions and motivations.

Acknowledgments The research reported here was supported by the Institute of Education Sciences, US Department of Education, through Grant R305F100013 to The University of Texas at Austin as part of the Reading for Understanding Research Initiative. The opinions expressed are those of the authors and do not represent views of the Institute or the US Department of Education.

References

Beck, D., & Eno, J. (2012). Signature pedagogy: a literature review of social studies and technology research. *Computers in the Schools, 29*(1–2), 70–94. doi:10.1080/07380569.2012.658347.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. doi:10.1146/annurev-psych-113011-143823.

Bolinger, K., & Warren, W. J. (2007). Methods practiced in social studies instruction: a review of public school teachers' strategies. *International Journal of Social Education, 22*(1), 68–84.

Burman, L. E., Reed, W. R., & Alm, J. (2010). A call for replication studies. *Public Finance Review, 38*(6), 787–793. doi:10.1177/1091142110385210.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychological Bulletin, 132*(3), 354.

Chiodo, J. J., & Byford, J. (2004). Do they really dislike social studies? A study of middle school and high school students. *Journal of Social Studies Research, 28*(1), 16–26.

Choi, J., Fan, W., & Hancock, G. R. (2009). A note on confidence intervals for two-group latent mean effect size measures. *Multivariate Behavioral Research, 44*(3), 396–406. doi:10.1080/00273170902938902.

Dunlosky, J. (2013). Strengthening the student toolbox: study strategies to boost learning. *American Educator, 37*, 12–21.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods, 12*(2), 121. doi:10.1037/1082-989X.12.2.121.

Gamoran, A., & Nystrand, M. (1991). Background and instructional effects on achievement in eighth-grade English and social studies. *Journal of Research on Adolescence, 1*(3), 277–300. doi:10.1207/s15327795jra0103_5.

Gwet, K. (2001). *Handbook of inter-rate reliability: how to estimate the level of agreement between two or multiple raters*. Gaithersburg: STATAXIS Publishing Company.

Hulleman, C. S., & Cordray, D. S. (2009). Moving from the lab to the field: the role of fidelity and achieved relative intervention strength. *Journal of Research on Educational Effectiveness, 2*(1), 88–110. doi:10.1080/19345740802539325.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2006). *Gates–MacGinitie reading test* (4th ed.). Rolling Meadows: Riverside Publishing.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: how often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542. doi:10.1177/1745691612460688.

Michaelsen, L. K., & Sweet, M. (2011). Team-based learning. *New Direction for Teaching and Learning, 128*, 41–51. doi:10.1002/tl.467.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus: statistical analysis with latent variables; user's guide*. Los Angeles: Muthén & Muthén.

Muthen, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431–462. doi:10.1007/BF02294365.

Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MLwiN*. London: Center of Multilevel Modeling.

Selig, J. P., & Preacher, K. J. (2008). Monte Carlo method for assessing mediation: an interactive tool for creating confidence intervals for indirect effects [Computer software on CD-ROM].

Swanson, E. A., Wanzek, J., Haring, C., Ciullo, S., & McCulley, L. (2013). Intervention fidelity in special and general education research journals. *The Journal of Special Education, 47*(1), 3–13.

Swanson, E. A., Wanzek. J., McCulley, L. V., Stillman-Spisak, S. J., Vaughn, S., Simmons, D., Fogarty, M., & Hairrell, A. (2014). Literacy and text reading in middle and high school social studies and English language arts classrooms. *Reading and Writing Quarterly*. doi:10.1080/10573569.2014.910718.

Vaughn, S., Swanson, E. A., Roberts, G., Wanzek, J., Stillman-Spisak, S. J., Solis, M., & Simmons, D. (2013). Improving reading comprehension and social studies knowledge in middle school. *Reading Research Quarterly, 48*(1), 77–93.

What Works Clearinghouse. (2011). *Procedures and standards handbook* (21st ed.). Washington: U.S. Department of Education, Institute of Education Sciences.

Yong, E. (2012). Replication studies: bad copy. *Nature*, 298–300. doi:10.1038/485298a.