The Effects of Team-Based Learning on Social Studies Knowledge Acquisition in High School

Jeanne Wanzek, Sharon Vaughn, Shawn Kent, Elizabeth Swanson, Greg Roberts, Martha Haynes, Anna-Maria Fall, Stephanie Spisak, and Michael Solis
The Effects of Team-Based Learning on Social Studies Knowledge Acquisition in High School

Jeanne Wanzek
Florida State University

Sharon Vaughn
The University of Texas at Austin

Shawn C. Kent
Florida State University

Elizabeth A. Swanson, Greg Roberts
The University of Texas at Austin

Martha Haynes
Florida State University

Anna-Mária Fall, Stephanie J. Stillman-Spisak, Michael Solis
The University of Texas at Austin

Jeanne Wanzek, Florida Center for Reading Research and School of Teacher Education, Florida State University; Sharon Vaughn, The Meadows Center for Preventing Educational Risk, The University of Texas at Austin; Shawn C. Kent, Florida Center for Reading Research and School of Teacher Education, Florida State University; Elizabeth A. Swanson, Greg Roberts, The Meadows Center for Preventing Educational Risk, The University of Texas at Austin; Martha Haynes, Florida Center for Reading Research, Florida State University; Anna-Mari Fall, Stephanie J. Stillman-Spisak, Michael Solis, The Meadows Center for Preventing Educational Risk, The University of Texas at Austin.
The research reported here was supported by Grant R305F100013 from the Institute of Education Sciences, U. S. Department of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Correspondence concerning this article should be addressed to Jeanne Wanzek, FCRR, Florida State University, 1107 W. Call St., P.O. Box 306-4304, Tallahassee, FL 32306.

Email: jwanzek@fcrr.org
Abstract

This randomized control trial examined the efficacy of team-based learning implemented within 11th grade social studies classes. A randomized blocked design was implemented with 26 classes randomly assigned to treatment or comparison. In the treatment classes teachers implemented team-based learning practices to support students in engaging in dialogue about course content, application of content to solve problems, and use of evidence to support responses. Significant differences in favor of the treatment group on content acquisition were noted (Hedges’ $g = .19$). Examination of differences in response to the treatment indicated groups of students classified with high or moderate pretest scores benefitted from the treatment, whereas a group of students classified with low pretest scores did not benefit from the treatment.

Keywords: team-based learning, social studies instruction, discourse, collaborative learning, secondary instruction
The Effects of Team-Based Learning on Social Studies Knowledge Acquisition in High School

Although there is often debate among educators as to the best curricula and instructional methods, there is ready agreement around the goal of improving students’ knowledge and critical thinking about content. The social studies provide a primary opportunity for achieving this goal. Curriculum and instruction that foster development of critical thinking and reasoning skills have been advocated by the National Council for Social Studies. Their 2008 position statement (National Council for the Social Studies, 2008) promotes several key instructional ideas including, (a) connecting content to students’ lives, (b) incorporating a broad range of resources for learning involving critical and creative thinking, (c) promoting reflective development/appreciation of democratic values (e.g., supporting differing points of view), (d) discussing content-based issues, and (e) engaging actively in a learning community. The importance of critical thinking is reflected in the number of times (5) it is recommended in the position statement. Such ideas are not new. Brophy (1990) recommended a pedagogical focus in social studies on higher-order thinking, moving beyond simple rote memorization of factual information and focusing on deep and meaningful connection of content within and across concepts to fully acquire content knowledge and understanding.

Achievement of these goals may be challenging considering only 17% of eighth graders and 12% of 12th graders demonstrated proficiency in the area of United States History according the most recent National Assessment of Educational Progress (National Center for Education Statistics, 2011). For high school seniors, results were unchanged from 1994, almost two decades prior. These findings suggest the need for continued investigation in refining current pedagogical practices in social studies classrooms. With this in mind, we conducted a
randomized control trial to examine the impact of implementing team-based learning (TBL; Michaelsen & Sweet, 2011), a pedagogical practice focused on generating critical thinking and engagement with content, on the social studies content acquisition of 11th grade students. TBL provides an instructional framework grounded in active engagement through collaborative group discourse about critical content, which in many ways represents a departure from more traditional social studies instruction often seen in practice.

**Instruction in the Social Studies**

In practice, the acquisition of social studies concepts has often occurred through reading informational text sources such as course textbooks (Bulgren, Deshler, & Lenz, 2007; Paxton, 1999) or through instruction guided by content-driven approaches for learning. These content approaches are often marked by lecture and worksheet-based activities that include recall of vocabulary and factual information (Scruggs & Mastropieri, 2003). Paxton (1999) explained that in the typical social studies classroom, passive reading and searching for facts in text is the norm rather than engaging in activities requiring the synthesis and interpretation of information in order to explain to and convince others. Bolinger and Warren (2007) surveyed secondary teachers in the social studies and found that almost two-thirds of teachers viewed lecture as the most effective teaching strategy in their classrooms followed by individual projects, and individual worksheets.

In a review of the literature on social studies instruction, Beck and Eno (2012) identified two distinct pedagogical models. The current mainstream pedagogy in practice tends to focus on the transmission of knowledge and often involves the short-term memorization of key facts, while student-centered instruction, an emerging pedagogy in practice, attempts to foster a deeper and more authentic understanding of content through historical inquiry.
Active student engagement with content supported through discourse and inquiry has been linked to the development of critical thinking skills as well as a high level of academic rigor in the classroom (Brophy, 1990; Wolf, Crussia, & Resnick, 2004). Discourse-based instruction focuses on collaborative and reflective student discussions regarding academic content that may involve arguments, explanations, or other conversational activities (Nussbaum, 2008). Although research on discourse suggests the relative paucity of such practices in the classroom (Applebee, Langer, Nystrand, & Gamoran, 2003; Nystrand & Gamoran, 1991), several studies have demonstrated positive outcomes when collaborative discourse approaches are part of instruction, including improvements in argumentation (Goldberg, Schwarz, & Porat, 2011; Kuhn & Crowell, 2011; Reznitskaya et al., 2001) and critical reasoning skills (Chinn, Anderson, & Waggoner, 2001; Gillies & Khan, 2009).

However, the extant literature on discussion-based approaches in the social studies has been limited. Further, as noted by Nussbaum (2008) in his review of the discourse literature, there exist few experimental studies establishing causal links between collaborative discourse and content learning in general. Two studies reviewed reported significant gains in content knowledge and conceptual understanding for students participating in discussion activities (Asterhan & Schwarz, 2007; Zohar & Nemet, 2002). Discussion in both studies was marked by critical, collaborative peer dialogue between peers focused on examining multiple viewpoints on a topic or issue. As collaborative discourse is a key component of TBL, it holds promise for increasing students' content-area learning.

**Team-Based Learning Framework**

whereby socially mediated problem-solving and discourse plays a key role in the development of meaning and reasoning. Essentially, instruction is structured to require students not just to learn and gain knowledge about the content but to engage in the scholarly practices of the discipline, allowing students to learn to think about the topic, express their thoughts and perspectives, provide opportunity for ideas to be challenged, consider perspectives of others, and socially construct knowledge via interaction and discourse with others. In this way, collaborative learning and discourse-based instruction represent important aspects of the TBL framework allowing for collective problem-solving, the assumption of multiple roles, confrontation of misconceptions, and collaborative work skills.

The TBL framework for instruction involves four foundational components or “pillars” (Michaelsen & Sweet, 2011). The first component is the creation of strategically formed, heterogeneous, permanent teams of students in the class. This process diverges from some uses of cooperative learning in the respect that such teams are considered permanent rather than temporary for a specific activity, allowing for increased communication and trust. The second pillar of TBL is known as a readiness assurance process, and provides a check on student understanding of the content and their readiness for higher level application exercises. In this component, individual students are administered a short, basic assessment (typically multiple choice) about previously taught content. Following the individual assessment, students work in teams to take the same quiz, but the teams are required to achieve consensus on the answers through dialogue and are provided immediate feedback on responses. The readiness assurance process is followed by re-teaching and clarification of content to address areas of confusion as evidenced on the individual and team assessments. The third pillar of TBL involves asking teams to engage in an application activity that requires students to use the content to address a
problem, make a specific choice, and report team decisions among classmates. The final pillar of TBL is a peer evaluation process allowing students to reflect on the contributions of team members and the team’s success.

Thus, the TBL framework incorporates several evidence-based instructional practices, incorporating individual and group accountability, feedback cycles including reteaching, motivational aspects, and structured discussions (Applebee et al., 2003; Guthrie et al., 2000; Hattie & Timperley, 2007). The framework of TBL allows it to be utilized across a variety of content-area domains. As stated by Michaelsen and Sweet (2011), “this sequence of activities shifts the focus of class time away from nearly 100% content ‘delivery’ to students actively helping each other learn how to apply the content” (p. 42).

**Research on Team-Based Learning**

Despite the promise of TBL to enhance critical thinking, student engagement, and content acquisition, we were unable to locate research on the use of TBL practices and its distinctive impact for school-age students who are in the college and career-readiness stage. To date, the extant research on TBL implementation has been conducted at the university level and almost exclusively in the medical and nursing fields, with generally positive outcomes on student satisfaction, student engagement, and content learning (Sisk, 2011). One limitation of the existing research on TBL is the absence of rigorous experimental studies; the corpus of studies in TBL primarily utilizes descriptive studies, comparison to historical control groups, and, in a few cases, quasi-experimental designs.

One quasi-experimental study utilized a crossover design to compare instruction using TBL to a more traditional case-based discussion approach for pharmacology students (Koles, Stolfi, Borges, Nelson, & Parmelee, 2010). Students performed significantly better on
comprehensive examination questions related to content learned via TBL than on questions related to the case-based learning. Similarly, Thomas and Bowen (2011) used a crossover design to contrast TBL instructional methods and small-group lecture on the content knowledge of medical students. Significantly higher performance was noted for learning during the TBL condition in five of six topic areas with an average effect size of .51.

Vasan and colleagues (Vasan, DeFouw, & Compton, 2011; Vasan, DeFouw, & Holland, 2008) have also examined the implementation of TBL practices, albeit somewhat modified, on the performance outcomes for medical students in courses such as anatomy and embryology. In these studies, the content application phase of TBL was not implemented. Performance of students on content-area examinations and national certification tests when taught using the modified TBL practices was compared to performance of a historical control sample who received traditional lecture-based instruction. In general, students receiving instruction utilizing TBL practices performed significantly better than the historical control on both outcomes. In addition, there was a reduction in the percentage of students failing the examinations and an increase in course grades. Other researchers have found similar results for both content learning and course performance when comparing TBL to more traditional methods of instruction in non-experimental studies (e.g., Carmichael, 2009; Wiener, Plass, & Marz, 2009; Zgheib, Simaan, & Sabra, 2010; Zingone et al., 2010).

Most recently, a multi-component instructional approach incorporating TBL, as well as the teaching of key vocabulary and student engagement with text, was examined in a randomized study in eighth grade history courses (Vaughn et al., 2013). Vaughn et al. represents the first published study of TBL use with school-age students. In comparison to typical instruction, the students in treatment classes demonstrated significantly higher content knowledge as well as
higher reading comprehension following instruction. Although this study provides encouraging findings for the use of TBL with school age students, the unique effect of TBL practices on either content knowledge or reading comprehension could not be separated from the other instructional strategies. Thus, in the current study we sought to examine the effects of TBL practices on these same constructs (content knowledge and reading comprehension) in order to further examine the distinct effects of TBL implementation with school-age students. We hypothesized that TBL practices would result in improved content knowledge through extended discourse and critical thinking about the content, but would not produce significant effects for general reading comprehension.

**Study Purpose**

In sum, TBL has demonstrated potential effectiveness for increasing knowledge acquisition of important content at the university level as well as initial evidence of efficacy as part of an instructional routine in middle school. Given the limited experimental research – only Vaughn et al. (2013) used a randomized trial and the findings reflect a multi-component intervention in which TBL was embedded – and the absence of studies at the secondary level, further research is needed to determine the impact of TBL practices on student learning. The purpose of this study was to examine the efficacy of TBL implemented within 11th grade social studies classes. Our primary research question addressed the effects of TBL (treatment) on student knowledge acquisition and reading comprehension compared to typical content instruction (comparison). Our secondary research question examined differences in student response to the TBL treatment based on initial content knowledge.

**Method**

**Participants**
The study was conducted in three schools in two school districts in the Southeastern part of the United States. A total of seven, 11th grade social studies teachers and their 26 history classes participated. Each teacher taught two to five sections of 11th grade history. Class sections were randomly assigned to treatment and comparison conditions blocking on teacher. Thus, each teacher had at least one treatment and one comparison class, allowing for systematic control of teacher effects. When a teacher had an odd number of class sections, the additional class was assigned to the treatment condition (5 sections = 3 treatment sections and 2 comparison sections). Fifteen classes were randomly assigned to treatment and 11 classes were randomly assigned to comparison.

Teachers. All seven teachers (three female and four male) were certified and held bachelor’s degrees; six of the teachers also held graduate degrees. Teaching experience ranged from three to 34 years ($M = 14.71$ years, median $= 14$ years) with all or most of that experience at the 11th grade level.

Students. A total of 463 students in the teachers’ 26 class sections consented for participation in the study. The total sample included 193 (42%) and 270 (58%) females with approximately 47% White, 45% African American, 4% Hispanic, 2% Asian, and 3% multiracial. Thirty-eight percent of the sample was enrolled in free or reduced lunch programs. Six percent of the sample was identified with a disability (i.e. speech impaired, language impaired, specific learning disability, deaf/hard of hearing). The 11th grade sample was representative of the demographics for district high schools. Over the course of the study, 31 students moved out of the participating schools. Thus, 432 (93%) students completed all three units of instruction. There were no significant differences on any of the demographic variables for participants in the treatment or comparison condition.
Procedures

All 11th grade social studies teachers at each school participated in the project. Thus, all of the general 11th grade social studies classes in the participating schools were included in the study. Advanced placement classes consisting only of students identified by the school as significantly above average in academic achievement were not included. The classes were randomly assigned within each teacher to either treatment (TBL) or comparison (typical practice). All consented students in each class were administered a battery of assessments by the research team prior to unit instruction (within five days prior to day one of the first unit). Teachers taught three units (Gilded Age, Imperialism and World War I, The Twenties) to both the treatment and comparison students during their regularly scheduled 11th grade classes. In the comparison condition, the teachers taught the three units as they usually would, using their typical practices. In the treatment classes, the teachers implemented TBL during each of the units. Teachers taught the same content and topics to both the treatment and comparison conditions. Only the implementation of TBL, not the content or curriculum, differed between the class sections. Class periods were 50-55 min and instruction occurred daily. Following the completion of the units, all students were assessed with the complete set of posttest measures within two weeks of the last unit day.

Description of Treatment Instruction

Professional development and support. All teachers attended a one-day professional development session (6 hrs) prior to implementation of the first unit. Teachers learned about the underlying principles of TBL, examined and worked with the TBL materials developed by the research team for each unit, identified implementation strategies for effective use of TBL in the classroom, and practiced implementing each of the TBL strategies. Teachers were also taught
the design of the study including the importance of maintaining a firewall between treatment and comparison classes related to the TBL strategies. All teachers continued to introduce and teach the content topics for each unit as they normally did in both treatment and comparison classes. Both the treatment and comparison classes covered the same curriculum and content aligned with the state standards over the same period of time. The only difference in the treatment classes was the integration of the TBL practices to apply and discuss the content that was taught. Thus, the contrast of interest was the delivery of content (with or without TBL) and not the content/curriculum provided.

In addition to the initial professional development session, each teacher was provided in-class support by one research team member. The researcher met with teachers to plan the TBL within each unit and discuss any implementation issues. The researcher also attended several class sessions to provide support for implementation, model facilitation of team discussions or other implementation procedures, and provide feedback on strengths and weaknesses of implementation. Teacher support in the form of planning or meetings with the teacher or in-class assistance was provided a minimum of four times for unit one, two times for unit two, and one time for unit three. All support sessions were related to TBL implementation; no support was provided related to the introduction or instruction on topics or content for the unit, which was provided as usual to both treatment and comparison classes.

**Team-based learning.** Teachers were asked to implement the intervention in three 15-day cycles corresponding to each unit. During each 15-day unit/cycle, two short comprehension checks (Days 4 and 8) and one cumulative comprehension check (Day 13) were conducted and each was followed by targeted instruction (Days 5, 9, and 14). Finally, one knowledge application activity occurred in each unit as a culminating activity and was spread across part of
two class periods (Days 14-15). Teachers implemented the three units for approximately 45 days (range 43-47 days).

**Student teams.** During comprehension checks and knowledge application activities, students worked in teams of three to five students. Teachers strategically assigned students to teams prior to the first unit, and students remained in the same team throughout all three units in order to build team cohesion. Minor changes in teams were made only if significant behavior issues occurred. Teachers formed teams strategically to distribute student skills evenly across teams taking into account such aspects as background knowledge, general academic achievement, temperament, participation disposition, and motivation or ability to focus on tasks. Teachers were asked to include evaluation of teamwork into the comprehension check and knowledge application activities to focus students on obtaining high-level, scholarly discussions and encourage team members to hold each other accountable for effective discussions. In this way, students were provided opportunities to self-monitor and self-evaluate key principles of the team discussions (e.g., active listening, use of evidence, critical thinking, extending thinking, active contributing, teamwork) with teacher feedback. When needed, some teachers also assigned rotating student roles (e.g., evidence facilitator, summarizer/synthesizer) within teams to help students work effectively as a team.

**Comprehension checks.** Comprehension checks included individual and team checks of content understanding through multiple choice questions. Quiz questions were designed to require students to integrate and evaluate information or perspectives, with the goal of eliciting discussion during the team checks. First, each student individually took a multiple choice quiz covering the unit content taught to date. Students then turned in their answer sheets. These individual checks provided teachers with information on student understanding and retention of
the content taught, including misunderstandings. Following the individual check, students moved into their teams and took the same quiz again. The teams were asked to discuss each question, identify evidence from text or other class materials to support their answers, and arrive at a consensus as to the correct answer. Each team received immediate feedback about their answer through the use of a scratch-off card keyed to the comprehension check answers. Once the students in the team reached consensus, they would scratch off their answer. If the answer was correct, a star would appear. If the incorrect answer was selected, the team was asked to further discuss the question and evidence in order to come to a consensus on a different answer. Teams discussed each question until they arrived at the correct answer. The scratch-off cards provided a permanent record of how many attempts it took each team to arrive at the correct answer, supplying the teacher with additional data regarding areas of confusion to target in instruction. Teachers facilitated discussion and use of evidence during the team comprehension checks. Teachers provided quiz points for both the individual and team checks, with full credit if the team obtained the correct answer on the first try, and the points reduced for each successive try. This point system increased motivation for carefully discussing questions with the team prior to scratching off the answer. Teams were also able to challenge a question or answer if they could provide evidence for one of their answers that was marked as incorrect. The teacher then discussed the evidence with the team and determined whether to accept the answer.

The two short comprehension checks (15 min each) consisted of five multiple choice questions covering the content taught prior to the administration of the check. The cumulative comprehension check occurred during a full class period and covered all material from the unit with 10 multiple-choice questions and one open-ended short answer question.
**Targeted instruction.** Teachers used the individual and team comprehension checks to identify misunderstandings or content requiring reteaching or clarification. Targeted instruction occurred for approximately 5 min at the beginning of each class period following a short comprehension check and for approximately 10 min at the beginning of the next class period following the cumulative comprehension check. During this time, teachers were expected to reteach content, clarify points of confusion, or discuss specific questions from the comprehension check with students.

**Knowledge application.** Knowledge application took place at the end of each unit and was designed for students to integrate content learned through discussion of authentic, multilayer, overarching questions or activities incorporating key content from the unit (e.g., *Establish the top three priorities for the nation as it moves from isolationism to expansionism. Then, consider the most pivotal events in each president’s term and make recommendations as to whether the same or a different course of action is recommended.*). During knowledge application each team discussed the various aspects of the question, identified evidence from handouts, presentations, class notes, or readings to support their points, and recorded key information. A final written activity was completed to address the question and integrate the key information from the discussion. Following completion of the written activity, each team reported their conclusions and rationale to the group, and teachers were encouraged to facilitate an evaluation of the team process.

**Fidelity**

Two independent researchers were trained to conduct observations in each teacher’s treatment and comparison classes to determine the extent to which the treatment was implemented and to identify any evidence of TBL elements in the comparison condition. For
each element of the intervention, the observer indicated the extent to which the teacher implemented that aspect of the intervention on a five point Likert-type scale (e.g., 1 = not implemented, 3 = approximately half of the required elements and procedures completed, 5 = all or nearly all of the required elements and procedures completed). In addition, the quality of implementation was also rated on a scale of 1 (low) to 5 (high), addressing how well the element was implemented exploring such issues as pacing, connecting information to prior knowledge, facilitating active engagement of students, monitoring student performance, and providing feedback. Because not all elements were expected on a given day of the unit, a score of NA was recorded in the treatment classes if the component was not a designated part of the intervention for the day.

Prior to beginning observations, the coders were trained on the fidelity form. The project coordinator served as the *gold standard* of reliability. The two coders established reliability of 85% or higher with the coordinator by independently observing videotapes of teacher implementation of the components from a previous pilot study.

**Treatment classes.** Table 1 provides the number of observations at each rating for each of the treatment components. The average rating for targeted instruction across teachers was 3.76 ($SD = 1.69$) with a quality of implementation mean at 3.44 ($SD = .86$). Targeted instruction is the only intervention component that was occasionally not observed in implementation. Targeted instruction was not implemented by teachers in 24% of the observations. The comprehension checks were implemented consistently with all or most of the required elements and procedures noted in each observation. The mean implementation rating for the comprehension checks was 4.94 ($SD = .24$) with a mean quality of implementation at 3.41 ($SD = 1.03$). The knowledge application was also implemented consistently, but proved to be the most
challenging for teachers, with some required elements and procedures missing from most of the observations. Knowledge application also had the greatest variation in implementation score. The mean implementation rating for knowledge application was 3.22 ($SD = 1.33$) suggesting that, overall, a little more than half of the required elements and procedures were implemented during the three knowledge application implementations (one implementation per unit). The mean quality of implementation for the knowledge application activities was 2.71 ($SD = 1.25$). The most common missing elements of implementation in knowledge application were: (a) highlighting key contributions and extending student thinking during the group discussions, and (b) highlighting key discussions as well as facilitating teams’ self-critique during the whole class summary/debriefing.

**Typical instruction classes.** We also observed classes assigned to the comparison condition in order to examine any possible overlap in instructional practices between the TBL treatment and the comparison condition. No TBL elements were observed during any teacher’s comparison classes. The average implementation score was 1.00 for targeted instruction, comprehension checks, and knowledge application indicating no elements of these practices were observed in the comparison classes. Teachers did implement whole class lectures, oral and written question/answer activities, text reading, and independent work during class periods. Teachers were observed activating background knowledge, teaching vocabulary, asking students questions about the content, giving quizzes, and asking students to take notes. All of these same activities were observed in the treatment classes as part of the content instruction as well. Thus, the content instruction was similar in treatment and comparison groups and TBL elements occurred only in the treatment classes.

**Measures**
In the Vaughn et al. (2013) study, the authors implemented TBL as well as reading comprehension practices as a means of positively influencing both content learning and reading comprehension. In the current study, we implemented only TBL but administered measures addressing the same constructs of content knowledge (Assessment of Social Studies Knowledge) and reading comprehension (Gates-MacGinitie Reading Comprehension subtest) in order to further isolate the effects of TBL. As noted earlier, we hypothesized that TBL practices would result in improved content knowledge, but would not produce significant effects for general reading comprehension.

**Assessment of Social Studies Knowledge (ASK; Vaughn et al., 2013).** The ASK assessment is a researcher-developed measure of content knowledge in the three units that comprised the intervention (Gilded Age, Imperialism and World War I, and The Twenties). The measure is untimed and consists of 44, four-option, multiple-choice items. A large pool of content-relevant items with known difficulty parameters from released state tests used in Texas and Massachusetts, and from released advanced placement tests, was assembled into an item bank. Permission to use selected items and relevant psychometric information were secured from the sponsoring entities. We ran a series of pilot tests in samples of students from the target population to validate the provided difficulty parameters, to refine the instructions for test administrators, and to estimate the amount of time necessary for administration (though untimed, we were concerned with test fatigue). We selected a final sample of items that performed comparably well across multiple occasions, our own pilot tests as well as the large-scale administrations reported by each item’s developer, was manageable within the larger assessment battery, and was maximally sensitive to change in the content areas of interest.
We used a series of item-level confirmatory factor analysis to evaluate model fit, to estimate item parameters, and to refine and finalize the group of selected items. Factor indicators were binary (correct /incorrect item response), variance was fixed as 1.0, 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. This specification is analogous to a 2-parameter logistic (2PL) item response model and yields indices of overall fit as well as estimates of difficulty and discrimination for each item. Observed data (n = 715) the 44-item assessment fit the model very well, $\chi^2 = 934.492$, $df = 902$, $p = .22$, CFI = .97, RMSEA = .009, suggesting that the selected items measured the targeted constructs with the greatest information available in the average ability distribution ($\theta$). The peak of the test information function (information = .91) occurred at −0.25 on the ability distribution, about one quarter of a standard deviation below average $\theta$, fixed at 0 by default. The information function was above .80 for ability levels ranging from -2.0 to 2.0, indicating reasonable coverage across the range of $\theta$. Test information indices indicate the precision of measurement for persons at different levels of the underlying latent construct, with higher information denoting more precision (Guion, 2011). It represents an IRT-based analogue for estimates of reliability used in classical test models. We also estimated Cronbach’s alpha. Across all sample members, the coefficient was .95.

**Gates-MacGinitie Reading Tests (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2006).** The Gates-MacGinitie is a group-administered, norm-referenced reading test for grade K-adult. We administered the Reading Comprehension subtest. Students are provided with expository and narrative reading passages followed by multiple choice questions.
Questions address facts, inferencing, and drawing conclusions. Internal consistency reliability ranges from .91 to .93 and alternate form reliability is reported as .80 to .87.

**Data Analysis**

**Hierarchical linear modeling.** To estimate the main effect of treatment on student outcomes, we fit hierarchical linear models (MLwIN 2.23; Rasbash, Steele, Browne, & Goldstein, 2004) with students nested in classes and classes nested in teachers. We grand-mean centered pretest scores on level-1 of the model (Enders & Tofghi, 2007). The effects of treatment (0 = control, 1 = intervention) were modeled at level 2. The corresponding model in MLwIN is:

\[ Y_{ijk} = \beta_{0ijk} + \beta_1 (\text{PRETEST} - \text{gm})_{ijk} + \beta_2 (\text{Treatment})_{jk} \]

\[ \beta_{0ijk} = \beta_0 + v_{0k} + u_{0jk} + e_{0ijk} \]

\[ v_{0k} \sim N(0, \sigma_v^2) \]

\[ u_{0jk} \sim N(0, \sigma_u^2) \]

\[ e_{0ijk} \sim N(0, \sigma_e^2) \]

where \( \beta_{0ijk} \) is the intercept (adjusted mean outcome after controlling for the differences in pretest status), which is random at all three levels. The random effect \( v_{0k} \) at level 3 represents variation in the adjusted mean outcome between teachers; \( u_{0jk} \), the random effect at level 2, represents random variation in the adjusted mean outcome between classes within teachers; and \( e_{0ijk} \) represents variation between students within classes within schools in the adjusted mean outcome. The level-1 covariate (pretest scores) is indicated by \( \beta_1 (\text{PREASK} - \text{gm})_{ijk} \) and \( \beta_2 (\text{Treatment})_{jk} \) represents the effect of treatment, which we represent as fixed for purposes of estimating main effects. We include the covariate at level 1 (versus level 2) because the study was sufficiently powered. Note that the subscripting routines used in MLwIN differ from those
used to specify models in the HLM (Raudenbush & Bryk, 2012) software. Statistical significance in this scheme is represented by a non-zero coefficient for the dummy-coded treatment variable (i.e., the .95 confidence interval does not include 0). 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.

**Growth mixture modeling.** To examine patterns of response to the TBL treatment, we used growth mixture modeling (Muthén, et al., 2002; Nylund, Asparouhov, & Muthén, 2007) to establish subgroups based on pretest scores and change over time. Growth mixture modeling (GMM) identifies unobserved subpopulations within a given sample based on differences in latent trajectories (and/or other latent attributes such as model-estimated scores at time 1). This contrasts with conventional latent growth models which assume that a given sample is from a single population. While GMM is not exploratory in the usual sense, it represents a more formative approach to data modeling and may be particularly useful when the extant research base is inadequate for supporting formal hypotheses.

We compared five models, differing in number of classes (2-class model through 6-class model), using multiple indices to identify the most informative and parsimonious model. Indices included the Akaike Information Criterion (AIC; Akaike, 1987), the Adjusted Bayesian Information Criterion (ABIC; Nylund et al., 2007), Lo–Mendell–Rubin Likelihood Ratio test (LMR; Lo, Mendell, & Rubin, 2001), and the index for entropy (Celeux & Soromenho, 1996). Simulation studies investigating the relative performance of these tests suggest that the adjusted BIC (Sclove, 1987) and the LMR are superior to AIC (Tofghi & Enders, 2007; Yang, 2006) in identifying the optimal number of classes. The LMR test is particularly useful because it compares the improvement in fit between $k$-1 and $k$ class models and provides a $p$ value for the
difference between nested models. A significant LMR value (p < .05) suggests that a model with $k$ classes represents the data better than the $k-1$ class model. The entropy index indexes the percentage of individuals in the sample that were correctly classified given the specified class model. Entropy values range from 0 to 1, with values of .70 or higher indicating good classification accuracy (Reinecke, 2006). For models with the same or similar levels of fit, we selected the more parsimonious model. We used Mplus 7.0 to fit growth mixture models (Muthén & Muthén, 1998-2012). Mplus 7.0 uses all available cases and a full information maximum likelihood estimator (Enders, 2010; Little & Rubin, 1987) to estimate model parameters. Standard errors and degrees of freedom were adjusted to account for clustering due to the data’s nestedness (Stapleton, 2006).

**Results**

**Main Effects of Team-Based Learning on Student Achievement**

Table 2 presents descriptive statistics for all available cases at pretest and at posttest. Strong, positive correlations were noted between the ASK and GATES measures (pretest $r = .601$; posttest $r = .787$). There were no statistically significant pretest differences between students in intervention and comparison group, suggesting pretreatment equivalence on the ASK ($\beta = -.39, SE = 2.13, p = .86$) and on the Gates-MacGinitie Reading Comprehension ($\beta = .038, SE = 1.44, p = .98$). Table 3 summarizes fixed and random effects for the models predicting reading outcomes. There was a significant main treatment effect on students’ content area knowledge as measured by the ASK content measure ($\beta_2 = 1.92, SE = .78, p = .01$). Participants in the intervention outperformed comparison students by an average of 1.92 raw score points, all else being equal. The Hedges’ $g$ was .19. Average posttest differences on the Gates MacGinitie
Reading Comprehension subtest did not differ across condition ($\beta_2 = .36, SE = 1.84, p = .84$). The Hedges’ $g$ was .03.

The random effects for the ASK content measure were significantly different from 0 only at Level 1 (i.e., student level) of the model, indicating that the variance in this variable was explained by differences between students rather than differences between classes, between teachers, or between some combination of classes, teachers, and students. On the Gates MacGinitie, variances at both the class and student level differed significantly from 0, with 16.8\% of the total variance at the class level and 71.0\% at the student level. Variance at the teacher level (12.18\%) did not significantly differ from 0.

**Growth Mixture Models**

The fit indices in Table 4 suggested a three-group model. The LMR was not significant for any of the five tested models. Values in the scree plot for the ABIC, were the lowest for the four-group solution; however, because this value was only 1 point lower (ABIC = 5343) than the ABIC value for the three-group model (ABIC = 5344), we decided that the additional class (i.e., adding a fourth group) did not meaningfully improve model fit. Accordingly, we selected the more parsimonious three-group model. Entropy was at an acceptable level for all five models, ranging from ,.71 and .80. The three-group solution correctly classified 71\% of students.

This solution indicated three classes of students: (1) high baseline in content knowledge and strong growth over time (group 1), (2) moderate baseline and no statistically significant growth (group 2), and (3) low baseline and no statistically significant or observed growth over time (group 3). Note that classification was independent of treatment condition. Group 1 consisted of 215 students, 137 (63.7\%) in the intervention condition and 78 (36.3\%) in the comparison condition. The average pretest score for content knowledge was 27.2 items correct
across all students in this group (i.e., independent of treatment condition), and the average growth from pretest to posttest was 6.83 ($p = .00$) items correct. Students in this high-performing group had high pretest scores on the ASK and made significant gains over time.

Group 2 had 159 students, 98 (61.6%) in intervention and 61 (38.4%) in the comparison condition. The model-estimated mean on the pretest ASK was 18.3 and the predicted growth rate was -.59 ($p = .61$), on average, for all students in this groups. Students in this medium-performing group scored lower at pretest than students in Group 1, and did not improve significantly from pretest to posttest, on average and independent of treatment condition.

Lastly, Group 3 had 76 students, 47 (61.8%) in intervention and 29 (38.2%) in the comparison condition. The estimated ASK pretest mean for this low-performing group was 12.49 and the average change was -.13 ($p = .87$). Students in this group scored very low at pretest and at posttest.

The intervention effect for the high-performing group (see Figure 1) was statistically significant. The average growth was greater by 1.585 ($\beta = 1.59; p = .04$) raw score points in the treatment group relative to the comparison (Hedges’ $g = .42$). We also found a positive and potentially meaningful difference in the medium-performing group; the intervention group outperformed the comparison at posttest by an average of 3.88 ($\beta = 3.88; p = .11$) raw score items, all else being equal. The difference was not statistically different, although the effect was large (Hedges’ $g = .79$). The lack of significance may be due to low power given the smaller number of students in this group. There were no significant treatment effects in the low-performing group ($\beta = -.54; p = .62$; Hedges’ $g = -.30$).

Discussion
In this study we examined the effects of integrating TBL in 11th grade social studies classes. Social studies teachers implemented the TBL components in the treatment classes for three instructional units. We first examined the general impact of TBL on content knowledge and reading comprehension. Overall, students participating in the classes with TBL significantly outperformed students in the typical instruction classes in content knowledge. Students in the treatment classes learned and retained more content from the instructed units than students in the typical instruction classes. These effects were noted only on the targeted construct of content knowledge. As hypothesized, no differences between groups were noted on a standardized measure of general reading comprehension. The previous randomized trial that included TBL as part of a multi-component intervention reported improved outcomes in both social studies knowledge ($ES = .17$) and standardized reading comprehension ($ES = .20$ [Vaughn et al., 2013]). Our findings support and extend the findings of Vaughn et al. indicating that TBL can be an active ingredient for improving social studies content knowledge demonstrating similar effects to those found in the Vaughn et al. study; however TBL does not appear to be an active ingredient in increasing reading comprehension. Thus, additional instructional components, such as those provided in the Vaughn et al. implementation are likely needed to address both knowledge acquisition and reading comprehension during social studies instruction.

Practically speaking, the effect for TBL implementation was small ($ES = .19$). However, this small effect is notable for two reasons: (a) the intervention length was relatively short, with teachers implementing TBL in only three instructional units; and (b) some aspects of TBL were challenging for teachers to learn and implement with their students at high fidelity in the short implementation period. Thus, increased content knowledge in the treatment classes was noted in a short instructional period with a novice implementation of TBL.
This study represents the first examination of TBL effects at the high school level. All of the previous work specifically on TBL practices has been conducted at the post-secondary level. In addition, most of the previous studies examining the effects of TBL were single group studies reporting differences in content knowledge from pretest to posttest (Sisk, 2011). Only a few studies have exclusively examined TBL relative to comparison conditions in quasi-experimental within-subject designs. Positive effects on content knowledge in various disciplines at the postsecondary level have been noted in the previous research but, as Sisk (2011) highlights, the non-experimental nature of the previous studies led to failures to control for variables such as content difficulty, content knowledge assessment equivalence, and/or differences in class size for the various conditions. The current study allowed us to examine the effects of TBL with a randomly assigned and equivalent comparison group where content and content difficulty remained the same across study groups, teacher effects were controlled for, and the content knowledge examination (ASK) was the same for both groups. Our findings provide support for TBL as an instructional method that can result in small increases in student knowledge and retention of content. In addition, our findings are the first evidence suggesting this improved learning can occur when TBL is applied in college-readiness, high school level courses.

This study also provides one of the first examinations of the fidelity of implementation of TBL. None of the aforementioned studies specifically examining TBL in postsecondary settings have addressed the teacher implementation of the independent variable. Only one previous study examined some aspects of TBL implementation as part of a multi-component intervention (Vaughn et al., 2013). In the current study with 11th grade social studies classes, the fidelity data suggested teachers had difficulty consistently implementing some elements of the knowledge application with their students. One area of difficulty occurred in the facilitation of
team discussions during knowledge application. Vaughn et al. examined implementation of the group discussion aspect of TBL and also reported only 50% of the observations included all of the discussion behaviors for knowledge application. During our observations, we noted that although teachers provided the general procedural guidance to students for working in teams to complete the discussions and activities, including facilitating students in using evidence to support their ideas, few teachers provided feedback during and after discussions to extend student thinking, highlight key discussions, or assist students in evaluating/critiquing their discussions. These facilitating behaviors are designed to provide specific feedback to students to assist them in improving their discussions and use of the content. We hypothesize that improved student effects would result from more consistent implementation of this instructional element. Our observations suggest that teachers and students might benefit from more opportunities to engage in these types of discussions before implementing a TBL approach. Previous studies have noted that student talk, particularly in key discussion behaviors such as asking questions and responding to group member statements, increased after six or more discussion opportunities (Hall, 2012; Murphy et al., 2009). In the present study, the knowledge application discussion, which focused on more open-ended content questions, occurred once each unit and, thus, teachers and students had three opportunities to engage in these types of discussions.

Further differentiation of the effects of TBL yielded findings indicating that some students differentially benefitted from treatment. We identified three groups of students differing in their response to the instruction. Students entering the instructional units with the highest pretest scores (e.g., knowledge of the designated social studies units) grew the most in knowledge. This growth was significantly greater for students participating in the TBL than for students who did not participate in TBL. TBL significantly benefitted these students who came
in with the most background knowledge. These students also had the highest mean reading comprehension scores, suggesting this group may have been an overall high academic achievement group. Students in the medium-performing group also appeared to benefit from the TBL implementation in a practically meaningful way, demonstrating the largest effect size among the three subgroups. In this medium-performing group, observed means for the TBL students were almost four points higher than for students in the typical instruction, with a large effect size of .79; however this difference was not statistically significant. Therefore, for both the medium- and high-performing groups of students, participating in TBL resulted in practically meaningful higher content-knowledge scores compared to similar students who participated in typical practice, but only the high-performing group differences were statistically significant.

There are multiple elements in the TBL framework that could account for the improved performance of treatment students in the medium- and high-performing groups. It may be that TBL engaged these students more fully in the acquisition and use of social studies. It may also be that the socially-mediated instruction not only engaged these students more, but provided an avenue to enhance critical thinking and reasoning. The students had opportunities to consider their own thinking along with the thinking of their team peers and reconcile differences in ways that may have promoted greater understanding of the content. In addition, it is possible that the required use of evidence during the TBL discussion led to more student-directed opportunities to review the content and make connections between key ideas furthering student understanding. The medium- and high-performing groups also entered the units with pretest scores that indicated some knowledge of the content prior to instruction. Their pretest knowledge on the upcoming content may have interacted with the TBL practices by improving their ability to make connections with the new content and engage in and maintain discussions of the topic, including
identifying evidence and using critical reasoning skills, which ultimately deepened their knowledge and retention of the content.

Although TBL implementation seemed to have some benefit for the students in the medium- and high-performing groups, students with the lowest pretest scores did not show growth in content knowledge whether they were in the typical classes or the TBL classes. The mean pretest score for the low-performing group as a whole was basically no higher than chance, suggesting this group of students had very little background knowledge for the upcoming content. Despite the content instruction they received in both the treatment and comparison classes, mean scores at posttest were essentially the same as pretest scores, indicating these students did not gain content knowledge. In addition, students in the low-performing group who participated in TBL classes did not perform better at posttest than the low students who were in typical practice classes. Students in the low-performing group also scored the lowest in reading comprehension prior to the intervention, though the confidence interval for the reading comprehension mean overlaps with the medium-performing group mean. Thus, the low-performing group may have characteristics of overall low academic achievement that go beyond the specific social studies content knowledge examined in this study. As we noted above, the brief nature of the intervention, particularly the knowledge application discussions, may have prevented students with the lowest achievement from developing the skills to participate fully in the discussions; whereas students with some pre-unit background knowledge may have been more equipped to make connections in the content instruction that allowed them to participate in TBL at higher levels.

This study does not inform us as to why the students in the lowest group made no gains in content knowledge whether in the treatment (TBL) condition or the comparison condition.
However, we present possible explanations as a mechanism for future research. Students with the lowest content knowledge may benefit from even more carefully scaffolded instruction for engaging in and applying the content in the TBL activities. Several interventions incorporating collaborative work or student-led extended discussions with students with learning difficulties have implemented highly structured, teacher-directed instruction of the processes for several days or weeks prior to full student implementation (Klingner et al., 2012; Rosenshine & Meister, 1994). Thus, a more structured, expert scaffolding process may be needed when TBL is implemented at the secondary level in order to assist the lowest students in benefiting from the instruction. Further, it is reasonable to think that students with low knowledge of the topic may not be fully engaged by their classmates in the group work. Students with low knowledge would have the least to offer their classmates in answering questions and this may be evident in their group work. It is conceivable then that students with low knowledge play a more passive role and rely on better students to respond and answer questions (Hall, 2012). Importantly, our findings for the comparison group reveal that there were students in our social studies classrooms who required additional supports to benefit from even the currently implemented typical instruction.

**Implications and Future Research**

This study provides initial evidence that TBL implementation may be a useful instructional practice for many adolescent students in the college and career-readiness stage. We note the increased content knowledge for many students when they participated in the TBL classes. However, we also note the three units of instruction occurred over a relatively brief period of time, suggesting additional research with longer implementations may provide additional evidence regarding the practical implementation of TBL within high school classes.
and the ability to increase knowledge application implementation to more expert levels. In addition, longer implementations may provide students with higher levels of content knowledge across a large variety of topic areas. Increased content knowledge, and thus background knowledge, does have the potential of increasing student’s reading comprehension abilities, particularly when reading within the content area. The strong correlations noted between the knowledge measures and reading comprehension measures suggest this possibility. Thus, although we hypothesized TBL was not directly related to improving reading comprehension, and our findings suggest this is the case, future research on longer implementations could provide additional evidence regarding the potential effects of TBL. Moreover, future research on the effects of individual elements (e.g., comprehension checks) of TBL could provide teachers with important information regarding the active ingredients of the instruction. This additional research, including examining other grade levels and content areas, would provide a better indication of the effects of TBL as well as further indication for whom and under what conditions TBL can be effective.

The differential results for students with the lowest pretest scores suggest that in these classrooms neither the TBL nor the typical instruction assisted these students in gaining knowledge of the content. Thus, our findings indicate further research is needed to identify the classroom and instructional supports required for students with lower initial achievement to understand and learn the intended social studies content.
References


http://dx.doi.org/10.1007/BF01246098

http://dx.doi.org/10.1598/RRQ.36.4.3


http://dx.doi.org/10.1080/03057640802701945

http://dx.doi.org/10.1080/07370008.2011.556832


http://dx.doi.org/10.1037/0022-0663.92.2.331
http://dx.doi.org/10.1080/10573569.2012.702037


doi:10.3102/10769986006002107


http://dx.doi.org/10.1177/0956797611402512


http://dx.doi.org/10.3102/00346543069003315


Table 1

*Frequency for Fidelity Observations in the Treatment Classrooms*

<table>
<thead>
<tr>
<th>Fidelity Element</th>
<th>Targeted Instruction</th>
<th>Comprehension Check</th>
<th>Knowledge Application</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$n$</td>
<td>%</td>
<td>$n$</td>
</tr>
<tr>
<td>Implementation</td>
<td>21</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>5 – All or nearly all of the required elements and procedures</td>
<td>13</td>
<td>61.9%</td>
<td>16</td>
</tr>
<tr>
<td>4 – More than half of the required elements and procedures</td>
<td>0</td>
<td>0.0%</td>
<td>1</td>
</tr>
<tr>
<td>3 – Approximately half of the required elements and procedures</td>
<td>3</td>
<td>14.3%</td>
<td>0</td>
</tr>
<tr>
<td>2 – Less than half of the required elements and procedures</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
</tr>
<tr>
<td>1 – Expected required elements were not completed</td>
<td>5</td>
<td>23.8%</td>
<td>0</td>
</tr>
<tr>
<td>Quality</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 – High</td>
<td>1</td>
<td>4.8%</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>38.1%</td>
<td>5</td>
</tr>
<tr>
<td>3 – Average</td>
<td>4</td>
<td>19.0%</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>14.3%</td>
<td>4</td>
</tr>
<tr>
<td>1 – Low</td>
<td>0</td>
<td>0.0%</td>
<td>0</td>
</tr>
<tr>
<td>NA – Component not completed</td>
<td>5</td>
<td>23.8%</td>
<td>0</td>
</tr>
</tbody>
</table>
Table 2

*Pretest and Posttest Means and Standard Deviations for Study Groups*

<table>
<thead>
<tr>
<th>Measures</th>
<th>Pretest</th>
<th>Posttest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>ASK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention</td>
<td>21.6</td>
<td>7.88</td>
</tr>
<tr>
<td>Comparison</td>
<td>21.8</td>
<td>7.58</td>
</tr>
<tr>
<td>Gates-MacGinitie</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intervention</td>
<td>102.4</td>
<td>11.85</td>
</tr>
<tr>
<td>Comparison</td>
<td>101.9</td>
<td>11.69</td>
</tr>
</tbody>
</table>

*Note. ASK = Assessment of Social Studies Knowledge.*
Table 3

Fixed and Random Effects for Models for Outcomes

<table>
<thead>
<tr>
<th>Measures</th>
<th>Fixed Effects</th>
<th>Predictor</th>
<th>Estimate (SE)</th>
<th>T-ratio</th>
<th>p-value</th>
<th>Hedges’ g</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASK</td>
<td></td>
<td>Intercept</td>
<td>25.09 (1.49)</td>
<td>16.83</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pretest</td>
<td>.83 (.05)</td>
<td>17.72</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td>Gates-MacGinitie Reading</td>
<td></td>
<td>Intervention</td>
<td>1.92 (.78)</td>
<td>2.45</td>
<td>.01</td>
<td>.19</td>
</tr>
<tr>
<td>Comprehension</td>
<td></td>
<td>Intercept</td>
<td>97.98 (1.92)</td>
<td>51.14</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pretest</td>
<td>.65 (.05)</td>
<td>14.51</td>
<td>.00</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Intervention</td>
<td>.36 (1.84)</td>
<td>.20</td>
<td>.84</td>
<td>.03</td>
</tr>
</tbody>
</table>

Random effects

<table>
<thead>
<tr>
<th></th>
<th>Variance</th>
<th>T-ratio</th>
<th>p-value</th>
<th>Percent of total variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASK</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1 (individual)</td>
<td>25.63 (1.99)</td>
<td>12.85</td>
<td>.00</td>
<td>63.61%</td>
</tr>
<tr>
<td>Level 2 (class)</td>
<td>1.79 (1.23)</td>
<td>1.45</td>
<td>.15</td>
<td>.44%</td>
</tr>
<tr>
<td>Level 3 (teacher)</td>
<td>12.88 (7.49)</td>
<td>1.72</td>
<td>.09</td>
<td>31.96%</td>
</tr>
<tr>
<td>Gates-MacGinitie Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comprehension</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 1 (individual)</td>
<td>66.06 (5.14)</td>
<td>12.85</td>
<td>.00</td>
<td>71.02%</td>
</tr>
<tr>
<td>Level 2 (class)</td>
<td>15.62 (6.80)</td>
<td>2.30</td>
<td>.02</td>
<td>16.80%</td>
</tr>
<tr>
<td>Level 3 (teacher)</td>
<td>11.33 (9.41)</td>
<td>1.20</td>
<td>.23</td>
<td>12.18%</td>
</tr>
</tbody>
</table>

Note. ASK = Assessment of Social Studies Knowledge. Hedges’ g effect sizes were estimated as the ratio between the model-derived intervention coefficients and the unadjusted pooled within-group standard deviation across conditions at posttest.

\(^a\) The T-ratio was determined by dividing the estimate by its standard error. \(^b\) Reference group is comparison.
Table 4

*Fit Indices from Model Testing*

<table>
<thead>
<tr>
<th>Classes</th>
<th>AIC</th>
<th>ABIC</th>
<th>Entropy</th>
<th>LMR</th>
<th>Class composition</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>5339</td>
<td>5353</td>
<td>.799</td>
<td>.07</td>
<td>(n_1 = 52%, n_2 = 48%)</td>
</tr>
<tr>
<td>3</td>
<td>5322</td>
<td>5344</td>
<td>.708</td>
<td>.56</td>
<td>(n_1 = 39%, n_2 = 46%, n_3 = 15%)</td>
</tr>
<tr>
<td>4</td>
<td>5314</td>
<td>5343</td>
<td>.763</td>
<td>.34</td>
<td>(n_1 = 18.3%, n_2 = 29.4%, n_3 = 47.9%, n_4 = 4.4%)</td>
</tr>
<tr>
<td>5</td>
<td>5310</td>
<td>5346</td>
<td>.795</td>
<td>.21</td>
<td>(n_1 = 35.9%, n_2 = 1.8%, n_3 = 46.4%, n_4 = 1.4%, n_5 = 14.5%)</td>
</tr>
<tr>
<td>6</td>
<td>5306</td>
<td>5350</td>
<td>.783</td>
<td>.23</td>
<td>(n_1 = 3.6%, n_2 = 48.6%, n_3 = 13.1%, n_4 = 1.6%, n_5 = 10.8%, n_6 = 22.3%)</td>
</tr>
</tbody>
</table>

*Note.* AIC = Akaike Information Criteria, ABIC = Sample size adjusted Bayesian Information Criterion. The percentages reported in the far right column represent the distribution of the participants across the classes for that particular model.
Figure 1. Estimated pretest and posttest mean scores on Assessment of Social Studies Knowledge for the treatment and comparison groups from the three class model.