Learning and Instruction 49 (2017) 64–80

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc

Instructional influences on English language learners' storytelling

Shufeng Ma [a, *], Richard C. Anderson [a, **], Tzu-Jung Lin [b], Jie Zhang [c], Joshua A. Morris [a], Kim Nguyen-Jahiel [a], Brian W. Miller [d], May Jadallah [e], Theresa Scott [f], Jingjing Sun [g], Kay Grabow [h], Beata M. Latawiec [i], Sherry Yi [a]

[a] *University of Illinois at Urbana-Champaign, United States*
[b] *Ohio State University, United States*
[c] *University of Houston, United States*
[d] *Towson University, United States*
[e] *Illinois State University, United States*
[f] *Booker T. Washington STEM Academy, Champaign, IL, United States*
[g] *University of Montana, United States*
[h] *Thomas Paine Elementary School, Urbana, IL, United States*
[i] *Wichita State University, United States*

ARTICLE INFO

Article history:
Received 9 September 2015
Received in revised form
2 December 2016
Accepted 8 December 2016
Available online 19 December 2016

Keywords:
English language learner
Collaborative group work
Direct instruction
Storytelling
Communicative competence
Multi-link causal reasoning

ABSTRACT

The goal of this study was to evaluate instructional influences on the storytelling of English Language Learners (ELLs). Participants were 210 fifth-grade Spanish-speaking ELLs (mean age = 10.8) from schools serving low-income neighborhoods in the Midwest of the United States. They received a six-week socio-scientific unit involving collaborative group work or direct instruction, or were in control classes that continued regular instruction. In an essay to evaluate mastery of the instructional unit, students from collaborative groups produced significantly longer chains of reasoning (more chains with 5–8 links) than direct instruction students (more chains with 3 or 4 links), while control students were unable to display any extended reasoning. Following the unit, students individually told a story prompted by a wordless picture book to evaluate their oral English proficiency. The stories were coded for several features of basic language production, story completeness, and multi-link causal reasoning. The results indicated that students who received the socio-scientific unit told stories with more complicated syntax than the control students, while no difference in complexity of syntax was observed between students from the two instructional conditions. Stories told by students who had participated in collaborative groups contained significantly more elaboration of essential story elements than the stories produced by direct instruction students or control students. Students who had interacted in collaborative groups also generated significantly longer chains of reasoning (many 5–7 link chains) connecting story events than students in the other two conditions (mostly 1 or 2 link chains). The results suggest collaborative group work may be an effective instructional approach to foster ELLs' communicative competence and causal reasoning.

1. Introduction

In the 2013–2014 academic year, the number of registered English Language Learners (ELLs) in public elementary and secondary schools in the United States was more than 4.9 million, which is 10% of the school population (United States Department of Education, 2016). Nearly 77% of the ELLs were children of Spanish-speaking immigrants from Central and South American countries. According to the 2015 National Assessment of Educational Progress (NAEP), fourth-grade Hispanic ELLs on average scored 49 points lower in reading than non-ELL white students (scale ranges from 0 to 500, standard deviation = 36), and eighth-grade Hispanic ELLs averaged 56 points lower than non-ELL white students. Influential factors that impede Hispanic ELL' literacy growth include low social

* Corresponding author. Center for the Study of Reading, University of Illinois at Urbana-Champaign 51 Gerty Drive Champaign, IL 61820, United States.
** Corresponding author.
E-mail addresses: sma10@illinois.edu (S. Ma), csrrca@illinois.edu (R.C. Anderson).

http://dx.doi.org/10.1016/j.learninstruc.2016.12.004
0959-4752/© 2016 Elsevier Ltd. All rights reserved.

economic status, limited home literacy resources, and poor oral English proficiency (Candelaria & Llorente, 2009; Goldenberg, 1996; Helman, 2009; Snow, Burns, & Griffin, 1998).

Among these factors, oral English proficiency has the most profound relationship with ELLs' academic attainment (Genesee, Lindholm-Leary, Saunders, & Christian, 2005). In a longitudinal study of a nationally-representative sample of Spanish-speaking ELLs in U.S. schools, Kieffer (2012) reported that ELLs' oral English proficiency in kindergarten significantly predicted their English reading in third through eighth grade. Prevoo, Malda, Mesman, and van IJzendoorn (2016) found in a meta-analysis of 61 studies that the association between language proficiency and reading comprehension increases as children enter higher grades where reading instruction focuses more on comprehension than word recognition. The National Literacy Panel on Language-Minority Children and Youth (August & Shanahan, 2008) concluded that half of ELLs with low English proficiency drop out of high school; however, for ELLs who can speak English well, the high school completion rate is 82%.

It is apparent that there is an urgent need for effective intervention programs to improve ELLs' oral English proficiency. In the next sections, we critically review issues in research on programs to accelerate ELLs' oral English and provide the rationale for the study described in this paper.

1.1. *Issues in research on the oral language proficiency of English language learners*

Systematic empirical research with ELLs has focused largely on basic elements of language and literacy (August & Shanahan, 2008). For example, the assessment of reading typically evaluates letter-sound knowledge, speed and accuracy of decoding, vocabulary knowledge, and sentence and passage comprehension. And the assessment of receptive and expressive oral language usually examines the ability to repeat words and phrases that have been orally presented, to name pictured objects to demonstrate expressive vocabulary knowledge, and to supply words missing at the ends of brief orally presented passages to demonstrate listening comprehension.

Relatively neglected in research with ELLs is what Hymes termed *communicative competence* (Hymes, 1972; see also; Cazden, 2011). Hymes explained that, “a normal child acquires knowledge of sentences not only as grammatical but also as appropriate. He or she acquires competence as to when to speak, when not, and as to what to talk about, with whom, when, where, and in what manner” (1972, p. 277). This means children not only need to know the basic elements of language, but also need to understand the genres and forms of speaking appropriate to the context. Therefore, we are among those (e. g. Baker, 2011) who believe that the assessment of ELLs' language should include open-ended production of language, rather than being limited to asking students to fill in blanks or choose the one correct answer from several options.

Available evidence supports the use of explicit, direct instruction with English language learners. With teachers' explicit instruction, students can develop fundamental language and literacy skills such as phonological awareness, decoding, and basic vocabulary (Avila & Sadoski, 1996; Kamps et al., 2007; Kelcey & Carlisle, 2013). Research has documented that teacher-guided practice of words, concepts, and step-by-step comprehension strategies can enhance basic literacy skills such as vocabulary (Silverman, 2007) and elements of reading comprehension (Roberts & Neal, 2004). August, McCardle, and Shanahan (2014) summarized findings from the National Reading Panel report (2000) that explicit instruction improves ELLs' phonological awareness, phonics, vocabulary, oral reading fluency, reading comprehension, and writing, and concluded that, since this influential report, the teaching of ELLs in the United States has increasingly focused on direct instruction of elements of reading and language.

However, direct instruction may restrict opportunities for language use in communicative circumstances. Classroom talk during direct instruction ordinarily takes place in the question-response-evaluation format. The interaction starts with a question asked by the teacher, proceeds with students' response to that question, and ends with an evaluation from the teacher. The question-response-evaluation format makes it difficult for students to produce extended talk (Dillon, 1988; Nystrand, 2006), especially students who are English language learners (Arreaga-Mayer & Perdomo-Rivera, 1996). Students' thinking is constricted because students have to follow the teacher's logic rather than initiate their own thinking. Students have limited control over when they can speak, little or no say about the topic of discussion, and negligible authority to evaluate whether contributions are acceptable (Wells & Arauz, 2006). In an observational study of 145 third-, fourth-, and fifth-grade classrooms in 20 low-income schools with large enrollments of ELLs in the southwest of the United States, McCaslin et al. (2006) reported that direct, teacher-led instruction predominated in virtually every classroom. Nearly 75% of instructional time in these classrooms focused on fundamental facts, basic skills, content learning, along with modest levels of elaboration and related thinking. Only 3% of instructional opportunities were devoted to thinking and reasoning. The few questions asked by students were mainly concerned with task procedures and correctness of answers; only 3% of student questions were reported to involve thinking or knowledge exploration (McCaslin et al., 2006).

Because of the constraints on spontaneous language and extended thinking associated with direct instruction, interactive learning approaches are often proposed as an alternative or supplement to facilitate development of oral language proficiency (cf. Ellis, 2005; Genesee et al., 2005; Vaughn et al., 2006). A family of interactive methods, with names such as Collaborative Reasoning, Shared Inquiry, Thinking Together, and Instructional Conversations, has proven to enhance classroom interaction and produce gains on several types of outcomes (Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009; Resnick & Schantz, 2015). For example, the Thinking Together intervention, in which students constructively and critically develop each other's ideas, has significantly improved the performance of third-through sixth-grade British and Mexican children on the Raven's Progressive Matrices, a test of non-verbal reasoning (Rojas-Drummond & Mercer, 2003). Instructional Conversations, promoted by Tharp and Gallimore (1989) and Goldenberg (1992) to improve teacher-student interaction, was found to increase fifth-grade Spanish-speaking ELLs' story comprehension and conceptual understanding (Saunders & Goldenberg, 2007). Zhang, Anderson, and Nguyen-Jahiel (2013) studied whether Collaborative Reasoning would improve fifth-grade Spanish-speaking ELLs' English reading, listening, speaking, and writing. They found that Collaborative Reasoning discussions of ethical and practical dilemmas raised in stories accelerated ELLs' oral and written English, as well as their motivation, engagement in discussions, and English learning attitudes.

We believe that interactive learning environments are likely to be more effective than direct instruction in promoting ELLs' ability to spontaneously communicate in English. Several socialization mechanisms may enhance children's language and thinking during cooperative and collaborative learning, including observing and emulating other children (Bandura, 1986; Lin et al., 2012), assistance less-competent children receive from others (Vygotsky, 1978; Webb & Mastergeorge, 2000), and the stimulation all of the children experience while resolving sociocognitive conflicts (Johnson & Johnson, 2009; Piaget, 1976/1947). But other scholars, notably

Kirschner, Sweller, and Clark (2006), have argued that direct instruction may work better than inquiry-oriented interactive approaches because novice learners may face challenges in complex learning environments when trying to process information that goes beyond their working memory capacity.

Research in second language acquisition suggests that interaction between language learners and more competent discussion partners may push learners to produce comprehensible, appropriate, and accurate output and use feedback to repair their language (Mackey & Goo, 2012). Swain and Watanabe (2012) concluded that in collaborative dialogue language learners may be able to co-construct meaningful content and produce complex linguistic forms that individuals could not achieve on their own.

Language is not only a tool for sharing thoughts and feelings with others, but can be regarded as the medium for thought itself. The symbolic structure of a language probably facilitates the formation of mental representations in children's minds and may enable other high-level cognitive processes (Boekaerts & Corno, 2005; Schunk & Zimmerman, 2007). And, the thinking process cannot be externalized without using words and sentences to describe it (Aydede, 2010). So, there is a relationship between language and thought that cannot be neglected, especially in bilingual education.

Compared to instruction for native speakers, instruction for ELLs frequently gives less emphasis to development of academic language, conceptual understanding, and thinking and reasoning. Oral language opportunities that enable children to develop the conversational skills for informal social situations are unlikely to be sufficient to impact the children's school performance (Cummins, 1986; Gersten & Baker, 2000). Language for social situations is easier in several respects than 'academic language.' Snow (2014) explains that "features of academic language include sophisticated vocabulary forms, explicit discourse markers (e.g. *nonetheless, therefore*), information packing through the use of nominalizations, embedded relative clauses, and subjectless passives … [These features] constitute an enormous challenge to struggling readers, second-language readers, and to those who have not been inducted into the use of academic language in oral contexts (p. 120)."

Schools with a large enrolment of ELLs tend to narrow the curriculum content to less challenging material (Sunderman, Kim, & Orfield, 2005), and reduce or eliminate exposure to untested subjects such as science and social studies in order to leave more time for basic skills instruction (Barksdale-Ladd & Thomas, 2000; Shepard, 2010). Several scholars have warned that lack of exposure to rich systems of concepts in an intellectually stimulating environment impoverishes children's thinking and reasoning (Chambliss & Calfee, 1998; Moll, 2010).

1.2. *Rationale for the present study*

This study compared the effects of two contrasting instructional approaches, collaborative group work and direct instruction, on the oral English development of fifth-grade Spanish-speaking ELLs. To increase exposure to academic language, enable conceptual understanding and stimulate higher-order thinking, students in both instructional conditions completed a curriculum unit in which they learned science and social science concepts in order to make an informed decision about a controversial public policy issue. Following the curriculum unit, among other tasks, the children told a story prompted by a wordless picture book to provide an authentic evaluation of their oral English proficiency.

Storytelling is a central oral language skill that emerges in preschool and continues to develop through the elementary school years, invoking higher-order cognitive skills and complex syntax (Berman & Slobin, 1994). The choice of storytelling for language assessment is supported by the finding of Francis et al. (2006) in a study enrolling third-grade Spanish-speaking ELLs from transitional bilingual education classrooms. This study found that overall skill in producing an English oral narrative, reflected in measures of its length, sophistication, and fluency, was "uniquely related" (Francis et al., 2006, p. 318) to an innovative measure of English reading comprehension (August, Francis, Hsu, & Snow, 2006) that deemphasizes word level processes and, instead, encompasses higher-level processing of text information, integration of prior knowledge and new information, combination of knowledge from multiple sources, and restructuring of knowledge constructs for longer retention.

Another reason we chose storytelling as the outcome measure to evaluate intervention effects is that the cognitive skills required for telling a coherent story are also essential for causal reasoning. According to Koslowski and Masnick (2010), most psychological research interprets causal reasoning as a person's sensitivity to Humean indices of causality, namely, temporal priority, contiguity, and covariation (Hume, 1748/1999). For example, when people constantly observe that one event happens right after another event, it is natural to think that the former causes the latter. In this sense, causal reasoning is a type of higher-order cognitive skill by which people explain how the change of one event or object may lead to the change of another event or object. Similarly, a coherent story contains a sequence of critical events that are connected in a meaningful way to advance the development of the main theme (Shapiro & Hudson, 1991), which requires the story events to be both temporally and causally linked (Stein & Albro, 1997). Thus, children's ability to connect story events is an indication of thinking and reasoning (Trabasso, Stein, Rodkin, Park Munger, & Baughn, 1992).

The present study takes a multi-dimensional approach to assessing ELLs' oral language development as represented in storytelling. Children's stories were coded for several measures of language production, story elaboration and completeness, and thinking and reasoning. These three facets of oral narrative skill were compared among students who received *collaborative group work* or *direct instruction*, or were in the control condition.

Collaborative Group work (CG) was a combination of Collaborative Reasoning (Anderson, Chinn, Waggoner, & Nguyen-Jahiel, 1998) and other group activities. Collaborative Reasoning is an interactive alternative to direct instruction that has shown promise in changing the quality of classroom talk (Chinn, Anderson, & Waggoner, 2001) and improving educational outcomes (Reznitskaya et al., 2009). Collaborative Reasoning discussions provide an open forum for students and minimize the dominant role of the teacher; this not only creates an interactive learning environment, but also facilitates communication monitoring as students learn to make contextually appropriate contributions to discussions and to ask for clarification of imprecise or ambiguous statements. In comparison to typical forms of classroom discussion, students' rate of talk almost doubles during Collaborative Reasoning and the talk more frequently involves elaborating text propositions, making predictions, using evidence, asking for and providing clarification, expressing and considering alternative perspectives, and drawing analogies between real and imagined situations (Chinn et al., 2001; Lin et al., 2012).

Research summarized by Reznitskaya et al., 2009 suggests that Collaborative Reasoning discussions may help students develop generalized skills of argument needed to reason about complicated problems. In studies in China and Korea (Dong, Anderson, Kim, & Li, 2008; Kim, Anderson, Miller, Jeong, & Swim, 2011), as well as the United States (Kim et al., 2011; Reznitskaya, Anderson, & Kuo, 2007; Zhang et al., 2013), students who participated in Collaborative Reasoning independently wrote essays, about a dilemma they had not previously discussed, that contained more acceptable arguments, counterarguments, and rebuttals than control students. Reznitskaya et al., 2009 explained these findings in terms of 'argument schema theory.' They hypothesized that Collaborative Reasoning enhances students' abstract knowledge of argumentation; consequently, they said, students learn "not *what* to think, but *how* to think" (p. 29).

Direct Instruction (DI) entailed explicit, teacher-led instruction which, according to Kirschner et al. (2006), is more effective than open-ended, inquiry-oriented instruction in situations like the present one in which the curriculum involves interrelationships among difficult or novel concepts. Direct instruction as implemented in the present study was based on the precepts of Stein, Carnine, and Dixon (1998). According to Stein and her colleagues, well-designed direct instruction emphasizes integration of skills and concepts, scaffolded instruction, explicitly taught strategies, a balance of highlights and details, and systematic review. There is ample evidence that students benefit from direct instruction in the basic elements of language and literacy (August et al., 2014) including evidence that students benefit from explicit teaching of comprehension strategies via direct explanation and modeling of strategies (Pressley & Wharton-McDonald, 1997).

The present study is a replication and extension of the study by Zhang et al. (2013), who found that Collaborative Reasoning enhanced the oral English narratives of fifth-grade Spanish-speaking ELLs. Zhang et al. was a small study involving just four classrooms and 75 students, so the findings are in need of replication with a larger sample. In addition to including many more classrooms and students than Zhang et al. the present study incorporated improvements in design and procedure: Collaborative group work was not only compared to a business-as-usual control, but also pitted against direct instruction; more advanced methods for analyzing oral narratives were employed.

To summarize, this study aims to understand whether students make greater progress in oral English after studying a conceptually rich curriculum unit via two instructional approaches, as compared to the uninstructed control students, and if so, whether collaborative group work and direct instruction have differential effects on the three facets of oral narrative skill, namely language production, story completeness, and thinking and reasoning. Based on the previous findings that the Collaborative Reasoning approach substantially increased the quantity and quality of classroom talk (Chinn et al., 2001; Lin et al., 2012), the collaborative group work approach was expected to lead to gains in all three aspects, because in collaborative groups children have more opportunities for high quality interaction. It was anticipated that the direct instruction approach might also lead to gains, as compared to the control condition, because as implemented in this study direct instruction involved richer concepts and greater use of connected academic language than is typical in classrooms containing large numbers of ELLs.

2. Method

2.1. Participants

The intervention was conducted in two waves across two academic years and each wave had 18 classrooms. Pooling over the two waves and setting aside students with other ethnic backgrounds, there were 324 Spanish-speaking ELLs from the 18 classrooms with predominant Hispanic American enrollment in four schools in a city in northern Illinois, nine classrooms in each wave. Classrooms within triples of classrooms matched on demographic characteristics and previous academic performance were randomly assigned to one of three instructional conditions: Collaborative Group work (CG), Direct Instruction (DI), or wait-listed Control that continued regular instruction and received the intervention via collaborative group work in the following semester after the data were collected.

In the first wave, the Hispanic American students in each instructional condition were enrolled in one bilingual classroom with instruction in both English and Spanish and two mainstream classrooms with instruction entirely in English. Students were assigned to classrooms based on their performance on the Illinois Measure of Annual Growth in English test (Illinois State Board of Education, 2000), an alternate statewide achievement test used in Illinois at the time of the study to evaluate English language learners. Those below the cut-score were assigned to sheltered bilingual classrooms; those above the cut-score were instructed in mainstream classrooms. Before the second wave, the participating schools abandoned the bilingual program for middle grade students, so all of the students were in mainstream classrooms, although it was reported that assistance in Spanish was still occasionally provided. Participant observers rarely saw any use of Spanish in the second wave, and not much use in the first wave even though the students least proficient in English were in bilingual classes. The infrequent use of Spanish reflected school district policy which emphasized acquiring proficiency in English rather than proficiency in both languages.

Among the 324 ELLs who participated in the project, the analyses reported in this paper involved the 210 students who received the individual storytelling assessment, 70 students in the CG condition, 68 in the DI condition, and 72 in the control condition. The remainder could not be given the assessment because of limitations on time and resources. Spanish was the first language of these students. More than 80% of them were registered for free or reduced price lunch. The sample was balanced in gender (Girl: $N = 103$; Boy: $N = 107$). The average age was 10.8 ($SD = 0.4$). Four students were registered in an Individualized Education Program (IEP) to receive special educational services, including two CG students, one DI student, and one control student.

Information about family background and language use was obtained from a parent questionnaire (both Spanish and English versions were provided). Table 1 presents the demographic characteristics of the 210 ELLs who took the storytelling task, as well as information about the language children spoke at home, parent education level, the language of instruction in the first grade, and home literacy resources and practices. Table 1 shows that 91% of the children communicated with their parents in Spanish, or a combination of Spanish and English and that 72% had instruction in Spanish, or a combination of Spanish and English, in the first grade. Most families reported limited home literacy resources.

Students who received the individual storytelling assessment were approximately 65% of the Hispanic American students in the participating classrooms. To determine which students would receive the assessment, assistants who conducted the assessment were given a list of students with 'target students' on the top and the remainder of the students randomly ordered below. Target students were so called because the video camera was trained on them throughout the intervention. In Control classrooms, there were nominal target students who were not videotaped during the period of the intervention. Target students were selected at the beginning of the study with the help of the teacher to be a representative cross-section of the class in terms of academic level, talkativeness, gender, and ethnicity. The number of target students in each class ranged from five to eight depending on the class size. To avoid interrupting classes all day long and moving back and forth between different schools, the storytelling assessment was conducted in one class at a time. Target students in each classroom received the assessment first, then according to the randomized list as many additional children as could be accommodated in the

Table 1
Demographic characteristics, language use, and home literacy resources and practices of participants (N = 210).

Class	Condition[a]	Program[b]	Sample size	Speak Spanish with parents	Spanish in first grade[c]	Parents' education[d]	Household child books	Household adult books	Book reading at home	Storytelling at home	Library visits
			>Boy; Girl	%	%	%	Less than 40 books (%)		Less than once a week (%)		
Wave 1: first year of intervention											
1	CG	M	5; 2	71	43	86	86	100	71	71	57
2	DI	M	5; 5	80	60	80	100	100	50	50	80
3	Control	M	5; 4	89	89	89	67	78	67	89	89
4	CG	B	6; 7	100	100	85	92	85	54	62	62
5	DI	B	5; 8	100	100	92	100	92	85	77	92
6	Control	B	8; 2	100	100	90	80	80	70	70	90
7	CG	M	2; 8	90	90	90	70	70	80	80	100
8	DI	M	4; 7	91	82	73	73	73	64	82	73
9	Control	M	4; 5	100	67	89	89	89	67	78	78
Wave 2: second year of intervention											
1	CG	M	8; 7	87	67	53	53	53	67	80	33
2	DI	M	6; 4	80	30	70	80	80	50	60	40
3	Control	M	11; 4	93	60	80	87	80	73	67	40
4	CG	M	5; 4	100	89	89	89	100	67	56	44
5	DI	M	4; 6	100	80	90	80	60	80	80	70
6	Control	M	8; 6	100	57	100	79	57	86	86	64
7	CG	M	8; 8	94	50	100	94	94	75	50	38
8	DI	M	5; 9	86	71	71	93	86	79	86	64
9	Control	M	8; 7	73	67	80	80	93	60	40	67
Average			**12**	**91**	**72**	**84**	**83**	**82**	**69**	**70**	**66**

[a] Condition: CG refers to collaborative group work, DI refers to direct instruction, and Control refers to the control condition.
[b] Program: M stands for mainstream classroom and B stands for sheltered bilingual classroom.
[c] Percentage of children who speak Spanish or English-Spanish mixture in the first grade among those who took the storytelling assessment.
[d] Percent whose parents had a high school education or less.

available time were assessed. Section 2.3 contains an analysis of whether there were differences between students who received the storytelling assessment and students who did not.

2.2. Instructional conditions

Students in the CG and DI conditions studied a six-week-long curriculum unit on wolf reintroduction and management developed by Jadallah et al. (2009). The unit is constructed around a science and public policy issue in an imaginary community. The people of Winona County are concerned about a pack of wolves that has been sighted nearby. Fears for the safety of children and pets, as well as potential threats to ranching and tourism, two of the principal businesses in the country, lead the County Board to write a letter to the Wolf Management Agency asking permission to hire professional hunters to kill the wolves. Students played the role of officials in the agency who must make the decision about eradicating the wolves.

As Jadallah et al. (2009) explained, the Wolf Unit uses a variety of information sources to help students learn about issues surrounding wolf reintroduction and management. Students read texts that incorporate different genres (e.g., expository text, newspaper articles, and formal letters). The unit integrates language arts, math, science, and social science. Three domains of knowledge are covered—ecosystem, economy and public policy. Each knowledge domain represents one thread of thinking on the relationship between wolves and the world around them. Reading materials and activities in each domain have an argumentative structure and cover both sides of the issues. Across the domains there are five overarching themes: interdependence, competition, balance, trade-offs, and the common good.

The intervention encompassed 22 class sessions, including a baseline lesson videotaped before the intervention and 21 Wolf Unit lessons. The sessions in the CG classroom were distributed as follows: introduction to the Wolf Unit (2), first Collaborative Reasoning discussion of the ‘big question’ (1), wolves in the United States (2), wolves and the ecosystem/economy/public policy (8), poster making and poster presentation of major concepts in each knowledge domain (6), the second Collaborative Reasoning discussion of the ‘big question’ (1), and a debriefing session (1). The sessions in the DI classroom covered the following topics: introduction to the Wolf Unit (2), wolves in the United States (2), wolves and the ecosystem (5), wolves and the economy (5), wolves and public policy (5), review of major concepts in the Wolf Unit (1), and a debriefing session (1).

Students studied Wolf Unit from Monday through Thursday, with occasional missed days for such reasons as all-school events, averaging roughly 1 h a day over a period of six weeks. Students in the control classes continued to receive regular language art instruction during the intervention period, but had the opportunity to study the Wolf Unit through collaborative group work in the semester following the study.

Collaborative group work was a combination of Collaborative Reasoning discussions and other group activities. After an introduction to Winona and its problem with wolves, students were broken into groups to discuss the ‘big question’—whether Winona should be permitted to hire professional hunters to the kill the wolves. On a typical day during the unit, small groups discussed a sub-question related to the ‘big question,’ for example, “What effect would killing the wolves have on the elks?” Groups worked independently and spoke freely among themselves, with occasional assistance from the teacher. Each small group was assigned to become ‘experts’ in one of the three domains of knowledge (ecosystem, economy, public policy). After four weeks of group work, the children in each expert group shared what they had learned in a poster presentation to the whole class. Then new discussion groups were created, with members from the three different expert groups, to reconsider the ‘big question’ in a Collaborative Reasoning discussion. As the last activity in the unit, students independently wrote a policy decision letter on whether killing the wolf pack should be permitted.

Direct Instruction entailed teacher-guided whole-class activities and individual seatwork. Students in DI condition sat facing toward the teacher. Students were supposed to raise their hands and wait for the teacher to select them before speaking. The teacher led students through all three domains of knowledge in the Wolf Unit. Activities that were completed in small groups in CG classrooms were completed individually as seatwork in DI classrooms. DI students discussed the policy decision as a whole class. The penultimate lesson in DI classrooms was a teacher-led review of the entire unit. As in the CG condition, students independently wrote a decision letter on whether killing the wolves should be allowed.

A total of 12 teachers participated in the project. Six of the 12 teachers participated in both years of the project, three teachers participated only in the first year, while the other three participated only in the second year. Among the six teachers who participated in both years, five were assigned to different conditions in the two years and one teacher had uninstructed control classes in both years. There were five male teachers and seven female teachers. Four of the teachers were Hispanic American and eight were European American. The average length of teaching experience prior to the intervention was 12 years, with a range of 3–20 years. There were no substantial differences in gender, ethnicity, or years of teaching experience among the teachers assigned to the three conditions.

Teachers who implemented CG or DI interventions attended parallel two-day workshops to receive a detailed introduction to the Wolf Reintroduction and Management Unit, discuss the design and content of the curriculum, and receive training in the method to which they were assigned. Teachers watched videos of Wolf Unit as it had been implemented in other classrooms taught by teachers using the method they were supposed to use. Teachers who implemented collaborative group work learned about the goal of the intervention, the research and theory supporting collaborative group work, how to facilitate Collaborative Reasoning discussions and effective strategies for promoting group work. Teachers who implemented whole-class direct instruction learned about the research and theory supporting explicit teaching of concepts and strategies, and effective methods for direct instruction. The workshop staff included elementary school teachers known for their expertise in collaborative group work or direct instruction.

A staff member was assigned to every classroom as a participant observer to administer tests, video record every lesson, take field notes following a classroom observation protocol developed by the research team, and provide feedback to the teacher. Participant observers were trained in each of the functions they were to perform and were supervised on site by an experienced member of the research team.

The field notes written by the participant observers recorded descriptions of classroom dynamics, including pacing of sessions, special events going on that day, noteworthy aspects of climate of the class, implementation of the unit that day including teacher's comments about implementation, students' interactions and attitudes toward one another and the teacher, students' interactions with the text and the activities such as unusual interpretations or questions. These field notes left no doubt that the Wolf Unit was implemented in every classroom and did not suggest noteworthy departures from the assigned instructional approach, although from time to time partner activities among the students were observed in some DI classrooms and most CG teacher occasionally explained concepts to the whole class.

The participant observer provided feedback to the teacher and exchanged ideas with the teacher regarding implementation issues after every lesson. Once a week the whole research team discussed any problems that the participant observers had encountered in the implementation of collaborative group work or direct instruction. On a rotating basis, the team viewed video clips from all of the classrooms, commented on problems and successes, and provided suggestions for the participant observers to share with teachers.

2.3. Pre-intervention assessments

Before the intervention, the Gates-MacGinitie reading comprehension test was administered (MacGinitie, MacGinitie, Maria, & Dreyer, 2000). The test entails answering multiple-choice questions after reading short passages. The raw score was corrected for guessing, which improved reliability and predictive validity. Also before the intervention, students individually completed a rapid naming task (Snodgrass & Vanderwart, 1980) to provide an assessment of speed of lexical access that is known to be sensitive to bilingualism (Misdraji-Hammond, Lim, Fernandez, & Burke, 2015). Students named common objects, such as bike, rabbit, and bus, in two sets of pictures. Students were asked to name the objects as quickly as they could but were allowed to say ‘skip.’ Both the total time for naming each set and any errors or skipped words were recorded. The final score was the number of words that students correctly named per minute. Regrettably, we did not assess ELLs' storytelling before the intervention, which would probably have allowed a more sensitive evaluation of condition differences. Another measure available from school files was performance on the Illinois Standards Achievement Test (ISAT) of reading which students took six months before the intervention when they were in the fourth grade. However, 50 participants (24%) did not take the pre-intervention ISAT due to high student mobility in the cooperating schools, including 20 CG students, 14 DI students, and 16 control students.

Prior to the intervention, students completed questionnaires to obtain information about social relationships and individual characteristics. Two of the questions asked students to nominate up to five of the quietest students in their class and up to five students who have the most to say during class discussions. Talkativeness was calculated as peer nominations for talkativeness minus nominations for quietness divided by the number of students in the class.

Comparing students who received the storytelling assessment and those who did not, the demographic characteristics, percentage of children who spoke Spanish with parents, and percentage who used English in the first grade were all very similar. Separate ANOVAs were performed (all the statistical analyses reported in this paper were performed using SAS Version 9.3) with Gates-MacGinitie reading score, speed of lexical access, and fourth-grade ISAT reading score as dependent variables, instructional condition and whether students received the individual storytelling assessment or not as fixed effects, and classroom as a random variable. The results showed that students who received the storytelling assessment had significantly higher Gates-MacGinitie reading scores than those who did not, $F(1, 303) = 4.40$, $p = 0.037$; however, there was no significant condition difference, $F(2, 303) = 0.65$, $p = 0.52$, or interaction between condition and whether students received the storytelling assessment or not, $F(2, 303) = 0.02$, $p = 0.98$. This result indicates that selection favored students who had higher reading comprehension, but this applied to all conditions. Students who received the storytelling assessment and students who did not receive the storytelling assessment showed equivalent performance on the rapid object naming task, $F(1, 303) = 1.11$, $p = 0.29$, as well as in the fourth-grade ISAT reading test, $F(1, 219) = 1.66$, $p = 0.20$. Therefore, although selection of students to receive the storytelling task was associated with reading comprehension, the students were alike in other respects and characteristics of selected students were not differentially related to instructional condition.

2.4. Post-intervention assessments

After the Wolf Unit, students completed an assessment battery providing extensive information about learning outcomes. Included in the battery were a 105-item sentence verification test that provided a broad, although not deep, assessment of concepts and information acquired from the Wolf Unit; a 50-min individually-written essay, called the policy decision letter, in which students explained their own decisions about whether the pack of wolves should be eradicated; an individual oral interview about an analogue to the wolf question, whether whaling should be allowed; a 50-min essay writing task about a moral and practical dilemma, whether a boy should tell on a classmate who cheated in a model car race; and the individual storytelling task.

The current paper reports children's performance on the storytelling task and their thinking and reasoning in the policy decision letter. The storytelling task was included in the assessment battery to determine if students had made generalizable improvements in language proficiency and higher-order thinking and reasoning (Larsen-Freeman, 2013). The purpose of analyzing the policy decision letter was to provide a benchmark for gauging the thinking and reasoning the children displayed in the storytelling task.

In the storytelling task, an assistant asked individual students to tell a story prompted by the wordless picture book, *Frog, Where Are You?* (Mayer, 1969). The assistant elicited the story following the procedure described by Berman and Slobin (1994). First, students looked through the book so that they could get a sense of the whole story and prepare for the task. Then, they were asked to tell a story while they turned the pages of the book. Assistants used standardized prompts when needed depending on students' behavior. For instance, assistants were to ask, "Can you tell more," if students stopped in the middle of the story. Students were allowed to code switch to Spanish if they did not know the English expression, although only one child in the current dataset switched to Spanish, when he did not know the English words for *jar, beehive*, and *antler*. The assistants were not blind to the intervention condition of the participants because the assessment was conducted in one class at a time in order to avoid interrupting classes all day long and to avoid moving back and forth between different schools. A limitation of the study was that we did not have any Spanish-speaking assistants, which may have been one reason why there was little code-switching in storytelling.

In the policy decision letter, children expressed their ideas about the 'big question' in the Wolf Unit, whether the community should be permitted to hire professional hunters to kill the wolves. The examiner explained the policy decision letter to the whole class and gave the students 10 min to write an outline. Then the examiner distributed a decision letter template and students began writing. They were given 40 min to complete the letter and were not allowed to use class notes during the task. Three DI students were absent on the day when the policy decision written task was administered. Control students were not asked to write decision letters in the first wave of intervention, so among those who took the oral narrative assessment the dataset includes 70 CG students, 65 DI students, and 44 control students who also wrote the decision letter.

2.5. Coding stories

2.5.1. Features of language production

Students' oral narratives were transcribed following the Systematic Analysis of Language Transcripts (SALT) conventions (Miller & Chapman, 2010). The transcripts were segmented into communication units (C-units). A C-unit is "a proposition or group of words that cannot be further broken down without loss of essential meaning" (Loban, 1976, p. 9). A C-unit represents an independent clause with all its modifiers, that is to say, one main clause plus its subordinate clauses. For example, "While the boy and the dog were sleeping, the frog decided to go out for a walk" is considered a C-unit, in which "the boy and the dog were sleeping" is a subordinate clause and "the frog decided to go out for a walk" is the main clause. Clauses with compound predicates were further segmented; for example, "The gopher popped out and bit the boy on his nose" was parsed into two C-units.

One analyst segmented all the transcripts into C-units. A different analyst independently segmented 20% of the transcripts. The percentage of agreement between the two analysts in C-unit segmentation was 98.4% (Cohen's K = 0.89). Next the segmented transcripts were coded following SALT conventions. The codes included bound morphemes (marked by a slash; e.g., take/3s), mazes (in parenthesis), omissions (denoted by an asterisk), pauses (denoted by a colon), and errors (denoted by word-level error code [EW:word] and utterance-level error code [EU]). A word-level error was marked when a word was used incorrectly; e.g., the dog falled [EW:fell] out of the window. An utterance-level error was marked when the error was not simply associated with a certain word; e. g., then the boy up (uh the) the rock/s [EU]. A maze referred to false starts, repetition and revisions, filled pauses, and part words, e.g., the dog was (barking um) barking at the beehive. Omissions included omitted words (e.g., the boy went *to the forest) and omitted bound morphemes (e.g., the boy look/*ed in the hole). Pauses included within-utterance pauses (denoted by a colon followed by the amount of time) and between-utterance pauses. A semi-colon followed by a colon was used to indicate pauses within the same speaker and two adjacent colons were used to indicate pauses between two speakers. Pauses were only noted when more than 3 s. The subordination index is represented by the SI-number, which refers to the total number of clauses in one C-unit. The following example includes bound morphemes, mazes, between- and within-utterance pauses, omissions and the code for subordination index. C stands for child and E refers to examiner.

> C (Once there):04 once there was a boy (who got a fr*) who had a frog and a dog [SI-2].
>
> C But now we/'re gonna tell you (how the story) how they got the frog [SI-2].
>
> E Ok.
>
> :: 05 between speaker pause
>
> C (O*) one day the frog got lost [SI-1].
>
> ;: 04 between utterance pause
>
> C Then: while: the (k*) boy was sleep/ing (um um) the boy woke up the next day (find*) find/ing out that (the) the frog *was miss/ing [SI-3].

The 13 measures obtained from SALT representing aspects of language production were factor analyzed using maximum likelihood estimation and varimax rotation. The measures loaded on five clearly defined language factors. The five factors were *length* (total number of C-units, total number of complete words, and total number of different words), *syntactic complexity* (mean length of utterance in words and in morphemes, and subordination index), *verbal fluency* (words per minutes, between and within utterance pauses), *mazes* (number of mazes, maze words, percentage of maze words, and utterance with mazes), and *errors and omissions* (omitted words, omitted bound morphemes, word-level errors, and utterance-level errors).

2.5.2. Story element coding

Story elements were defined on the basis of Stein and Glenn's (1979) story schema theory. The process of storytelling is "a product of interaction between incoming information and strategies, mental operations, and structures inherent in the [storyteller]" (Stein & Glenn, 1979, p. 54). The fundamental elements in a story include a *setting* and an episode system. The episode system (i.e., a collection of different episodes) can be further split into several components—initiating event, internal response, external response, and consequence. *Initiating Event* refers to the event that initiates the response of a main character. *Internal Response* refers to "the psychological state of a character after an event" (Stein & Glenn, 1979, p. 65) and takes three forms—*goals, cognitions,* and *affects*. *External Response* refers to a sequence of behavior that reveals the main character's attempts to change the "disequilibrium that was caused by the initiating event" (Stein & Glenn, 1979, p. 67). External response includes external events with or without a purpose. Events with an explicitly specified purpose are *attempts*, whereas events that do not explicitly specify a purpose are *actions*. *Consequences* in the episode system describe whether the main character achieves a goal or not. A consequence could be an *outcome* of an attempt to reach a goal or an *end state*. The episode structure of the frog story is presented in Fig. 1. The first author coded all the transcripts and a different coder coded 20% of the data to check the reliability of coding of each story element. The definition and example of story elements are provided in Table 2, along with coding reliabilities.

Previous studies (e.g., Francis et al., 2006; Zhang et al., 2013) have relied on the Narrative Scoring Scheme (Miller et al., 2006) as an index of children's ability to produce well-formed narratives as defined in story schema theory. In this scheme, judges rate the quality of stories in terms of each story schema category. The overall score is the sum of the ratings in the various categories. However, a rating scale is inherently subjective and different raters may apply a different standard. A given rater's standard may drift over the course of rating a number of stories. Ratings are subject to halo effects from the rater's general impression (Nisbett & Wilson, 1977).

To avoid these problems, we developed what we hoped would be an improved measure called *Essential Story Elements* that entails low-inference rule-governed assignment to categories. The Essential Story Elements measure is also based on story schema theory (Stein & Glenn, 1979), but each utterance was judged in terms of whether it expressed one of the story schema categories. In this study, Essential Story Elements was defined as the sum of non-repetitive statements in six critical story schema categories, namely *setting, initiating event, internal response, attempt, outcome,* and *end state*. The coding scheme for Essential Story Elements in the frog story is described in Fig. 2. When evaluating the stories of children with limited English, Essential Story Elements is probably less vulnerable to negative halo effects than the Narrative Scoring Scheme; the poor general impression created by disfluencies may induce lower ratings of the stories of ELLs.

2.5.3. Coding for multi-link reasoning

Multi-link reasoning refers to the ability to organize information and bridge inferences into causal chains (Lin et al., 2011). Six types of multi-link chain were identified in decision letters, including the ecosystem-economy model, oxygen model, leftover model, competition model, supply model, and tourism model. One coder coded all the letters and a different coder independently coded 20% of the letters. The intercoder percentage of agreement on multi-link reasoning chains in policy decision letters was 93.0% (Cohen's K = 0.88). The number of links in chains was then compared between the different instructional conditions.

The ecosystem-economy model presented for illustration in Fig. 3 consists of an eight-link chain. The first link of the chain contains the simple relation: "People kill wolves" formed from two objects *wolf* and *elk,* along with a relationship *eat.* The consequence of this simple relation is expressed as, "People killing wolves will cause a change to the wolf population." This proposition contains a cause-effect relationship with simple relations serving as basic

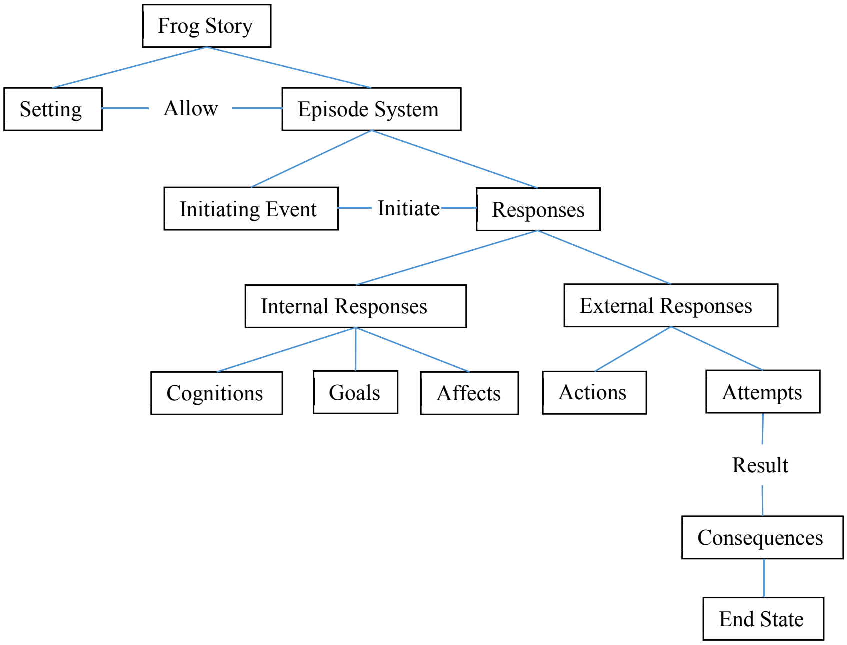

Fig. 1. Story gammar of the *Frog, Where Are You* story.

Table 2
Definitions and examples of story elements.

Categories	Definitions		Examples in the frog story	Coding reliability
Setting	The introduction of main characters and the description of the temporal and physical context in which the story takes place.		There was a boy who had a frog and his and had a dog. It was a beautiful night.	99.6% (Kappa = 0.96)
Initiating event	The event that initiates the response of a main character.		Then the kid went to sleep while the frog um was bored and got out of the yard. Then when the kid woke up, he saw that the frog wasn't there.	98.7% (Kappa = 0.87)
Internal response	*Goals*	The purpose for taking an action.	The boy searched in the hole to look for the frog.	86.8% (Kappa = 0.74)
	Cognitions	Characters' thoughts or beliefs.	He thought the dog would die.	
	Affects	Characters' emotional reactions such as happiness, sadness, worry, or fear.	They were worried where the frog was at.	
External response	*Attempts*	Events with an explicitly specified purpose.	The boy was screaming or yelling at b* tree holes or ground to find if the frog was there.	97.5% (Kappa = 0.86)
	Actions	Events that do not explicitly specify a purpose.	Then the boy opened the window.	94.0% (Kappa = 0.82)
Consequences	*Outcome*	An outcome of an attempt to reach a goal.	He kept looking for the frog everywhere in his shoes, his clothes all around his room. But he couldn't find her.	86.3% (Kappa = 0.77)
	End state	The main character stops making further attempts either by achieving the goal or completely giving up.	Then the boy found his frog. And he took the frog to his home and was saying goodbye to the other frog family.	98.5% (Kappa = 0.85)
Minor events	The activities of supporting characters (i.e., bees, gopher, owl, and deer), which are not critical to the plot development but are mentioned in response to the picture book.		The dog had knocked down the beehive and broke the beehive. And a bunch of bees came out.	93.5% (Kappa = 0.83)

units. In the second link, the simple relation is "Wolves eat elk." The consequence is expressed as, "The elk population will change." The two links are connected by considering the two consequences to be a causal relation. Successive links are connected in the same way. An example of the ecosystem-economy model in a student's decision letter is as follows:

I think that [wolves should not be killed] because wolves eat elk and elk eat trees. Now think of this if the elk eat most of the trees it could possibly cause the Garcia Timber Company to shut down. If that happened, the Happy Burger would make less money because loggers come to eat there every day. But if there were no loggers then Happy Burger would lose money and would not be able to buy Ringo's Beef Ranch or Bubbly Colds products. That alone would cause Winona's economy to lose millions of dollars.

Parallel to the multi-link reasoning structures in policy decision letters, we further examined construction of causal chains in storytelling. Trabasso and his associates (Trabasso & van den Broek, 1985; Trabasso, van den Broek, & Suh, 1989) created an influential model of story coherence in which events in the narrative are organized as a causal network. Each story event serves as consequence for the previous event and acts as antecedent for the following event until the character's goal is fulfilled or no further attempts are made. The importance of an event is determined by its relationship with other events and its position in the hierarchical structure. According to Magliano (1999), the causal network model explains how story elements are bound together based on the different types of causal relationships between episodic categories. *Goals* can activate *attempts* and *attempts* can result in *outcomes*.

In storytelling, a multi-link chain is a sequence of story events connected together based on causal relationships. Narratives consist of interconnected sequential events. Each event describes an attempt to achieve the goal of the main character. The failure of one attempt causes the next attempt. In other words, the occurrence of subsequent events is based on the outcome of previous events. By linking sequential events together, a causal reasoning chain is formed, which makes the story more organized and coherent. Trabasso and colleagues (Trabasso & van den Broek, 1985; Trabasso et al., 1989) used the causal model to analyze the stories of adult authors, not the stories told by children. So, it is a step forward to analyze child-created stories in terms of causal connectedness.

The coding of multi-link reasoning chains is based on the identification of attempts and outcomes. In the frog narrative, an outcome to an attempt is either the failure (fail to find the frog) or the success (find the frog) of an attempt. The main character keeps making attempts until the fulfillment of the goal. Every event with an explicit outcome in this causal chain is considered as a link in multi-link reasoning process. The intercoder percentage of agreement for multi-link reasoning chains in the oral narratives was 93.3% (Cohen's K = 0.83). An example of a seven-link causal chain is as follows.

They went outside and looked. *Nothing happened.* They went to the trees. *And nothing happened*. They went to a beaver hole. (Th*) *She wasn't there*. They went to look on the (bee) bee hole. *And there was nothing there*. They look*ed everywhere, even on trees, even where the owl lives. (They lo*) they look*ed under rocks. They looked over rocks. They called him his name. They even looked (on) on (um) deer, *but the deer didn't want them to check on there*. They looked on the water. *And nothing was there*. They look*ed over the log. *Something was there*. (Th*) finally they found the frog.

3. Results

3.1. Aspects of home background, initial reading and language proficiency

Two-level logistic regression models were created to test if there were condition differences in the eight aspects of demographics, language use, and home literacy resources and practices summarized in Table 1, with condition as the fixed effect and classroom as the random effect. The results indicated no condition difference in age of English acquisition, $F(2, 192) = 0.08$, $p = 0.92$; use of Spanish with parents at home $F(2, 192) = 0.10$, $p = 0.91$; parents' education level, $F(2, 192) = 0.59$, $p = 0.56$; household child books, $F(2, 192) = 0.72$, $p = 0.50$; household adult books, $F(2, 192) = 0.09$, $p = 0.91$; book reading at home, $F(2, 192) = 0.05$, $p = 0.96$; storytelling at home, $F(2, 192) = 0.31$, $p = 0.74$; or library visits, $F(2, 192) = 0.99$, $p = 0.40$.

Table 3 summarizes the descriptive statistics on the pretest measures of reading comprehension and speed of lexical access, and the fourth-grade ISAT reading test. Separate ANOVAs were conducted in which instructional condition was a fixed effect, and classroom was a random effect to account for variance due to the teacher or the student cohort. The results indicated no condition difference for pretest reading comprehension, $F(2, 192) = 0.64$, $p = 0.53$, pretest speed of lexical access, $F(2, 192) = 0.58$, $p = 0.56$, or fourth-grade ISAT reading, $F(2, 142) = 0.69$, $p = 0.51$.

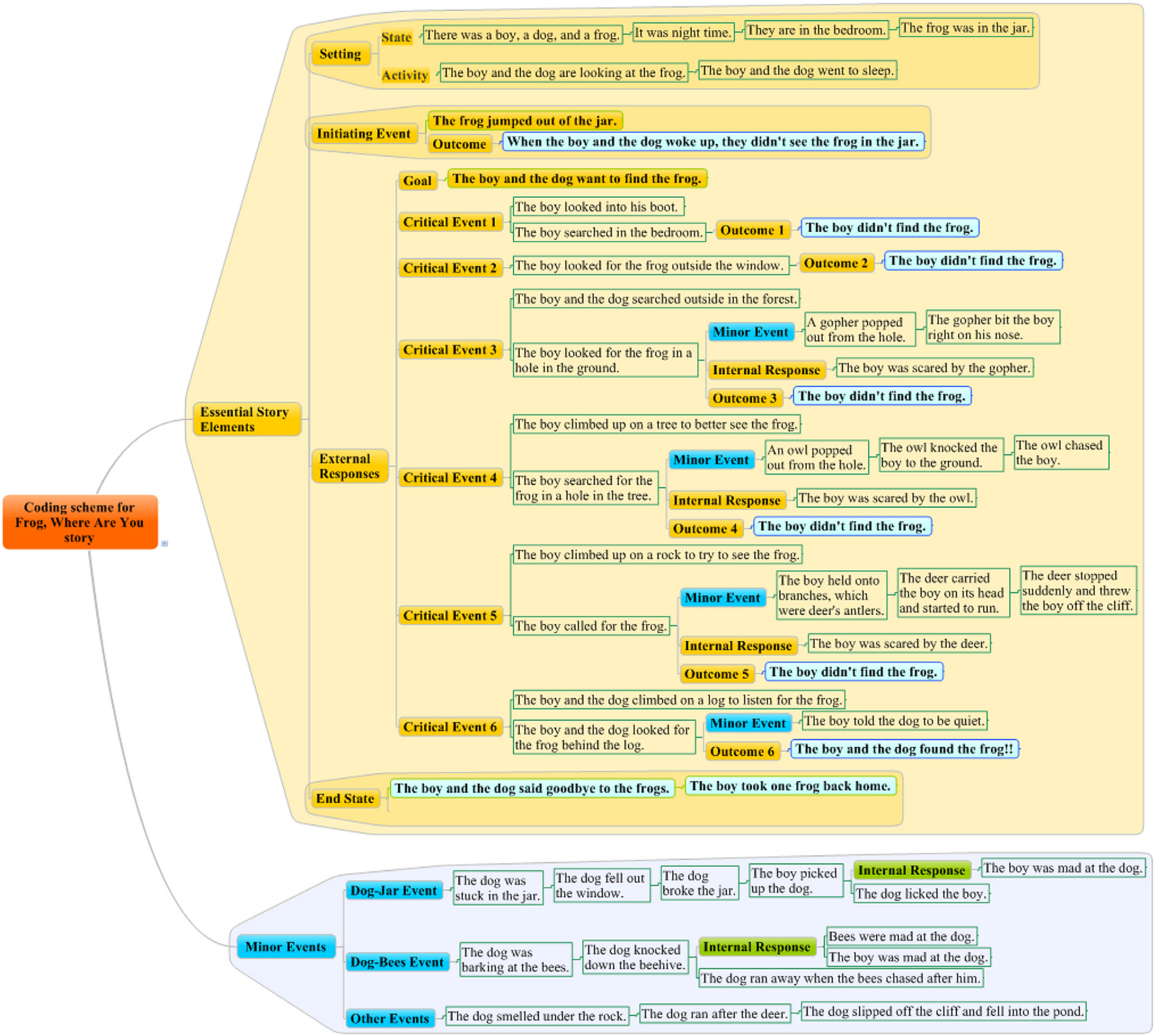
Coding scheme for Frog, Where Are You story
Essential Story Elements
Setting
State
There was a boy, a dog, and a frog.
It was night time.
They are in the bedroom.
The frog was in the jar.
Activity
The boy and the dog are looking at the frog.
The boy and the dog went to sleep.
Initiating Event
The frog jumped out of the jar.
Outcome
When the boy and the dog woke up, they didn't see the frog in the jar.
External Responses
Goal
The boy and the dog want to find the frog.
Critical Event 1
The boy looked into his boot.
The boy searched in the bedroom.
Outcome 1
The boy didn't find the frog.
Critical Event 2
The boy looked for the frog outside the window.
Outcome 2
The boy didn't find the frog.
Critical Event 3
The boy and the dog searched outside in the forest.
The boy looked for the frog in a hole in the ground.
Minor Event
A gopher popped out from the hole.
The gopher bit the boy right on his nose.
Internal Response
The boy was scared by the gopher.
Outcome 3
The boy didn't find the frog.
Critical Event 4
The boy climbed up on a tree to better see the frog.
The boy searched for the frog in a hole in the tree.
Minor Event
An owl popped out from the hole.
The owl knocked the boy to the ground.
The owl chased the boy.
Internal Response
The boy was scared by the owl.
Outcome 4
The boy didn't find the frog.
Critical Event 5
The boy climbed up on a rock to try to see the frog.
The boy called for the frog.
Minor Event
The boy held onto branches, which were deer's antlers.
The deer carried the boy on its head and started to run.
The deer stopped suddenly and threw the boy off the cliff.
Internal Response
The boy was scared by the deer.
Outcome 5
The boy didn't find the frog.
Critical Event 6
The boy and the dog climbed on a log to listen for the frog.
The boy and the dog looked for the frog behind the log.
Minor Event
The boy told the dog to be quiet.
Outcome 6
The boy and the dog found the frog!!
End State
The boy and the dog said goodbye to the frogs.
The boy took one frog back home.
Minor Events
Dog-Jar Event
The dog was stuck in the jar.
The dog fell out the window.
The dog broke the jar.
The boy picked up the dog.
Internal Response
The boy was mad at the dog.
The dog licked the boy.
Dog-Bees Event
The dog was barking at the bees.
The dog knocked down the beehive.
Internal Response
Bees were mad at the dog.
The boy was mad at the dog.
The dog ran away when the bees chased after him.
Other Events
The dog smelled under the rock.
The dog ran after the deer.
The dog slipped off the cliff and fell into the pond.

Fig. 2. Coding scheme for the *Frog, Where Are You* story.

3.2. Features of language production

Table 4 presents standardized factor scores (percentaged z-scores, Straus, 1980; with $M = 50$, $SD = 20$) by instructional condition for each of the five language production factors (see the Pearson product-moment correlations of pretests and outcome measures in the Appendix). The overall difference among the three conditions was examined in a MANCOVA with the five language factor scores as dependent variables. Instructional condition (CG, DI, Control), wave of intervention (1 = first year, 2 = second year), age of English acquisition (1 = English in the first grade, 2 = Spanish or a mixture of English and Spanish in the first grade), parents' education level (1 = up to and including high school education, 2 = two-year college education and higher), and whether or not students were registered in an Individualized Education Program (IEP) were entered as fixed effects. Reading comprehension and speed of lexical access (class-mean-centered scores) and talkativeness were covariates. Class-mean reading comprehension and speed of lexical access were entered to account for between-classroom variance.

There was a significant overall effect of instructional condition on language production, Wilks' Lambda = 0.90, $F(10, 388) = 2.16$, $p = 0.019$. There was also a significant difference between the first and the second wave, Wilks' Lambda = 0.83, $F(5, 194) = 7.98$, $p < 0.001$; and a significant difference between students who used and did not use English in the first grade, Wilks' Lambda = 0.94, $F(5, 194) = 2.47$, $p = 0.034$. Individual-level reading comprehension, individual-level speed of lexical access, class-level reading comprehension, and class-level speed of lexical access were all significant, $ps < 0.01$. Parent education level, whether or not students had an IEP, and student talkativeness did not significantly predict overall language production.

Two-level models were constructed for follow up univariate analysis of each of the five language production factors in order to

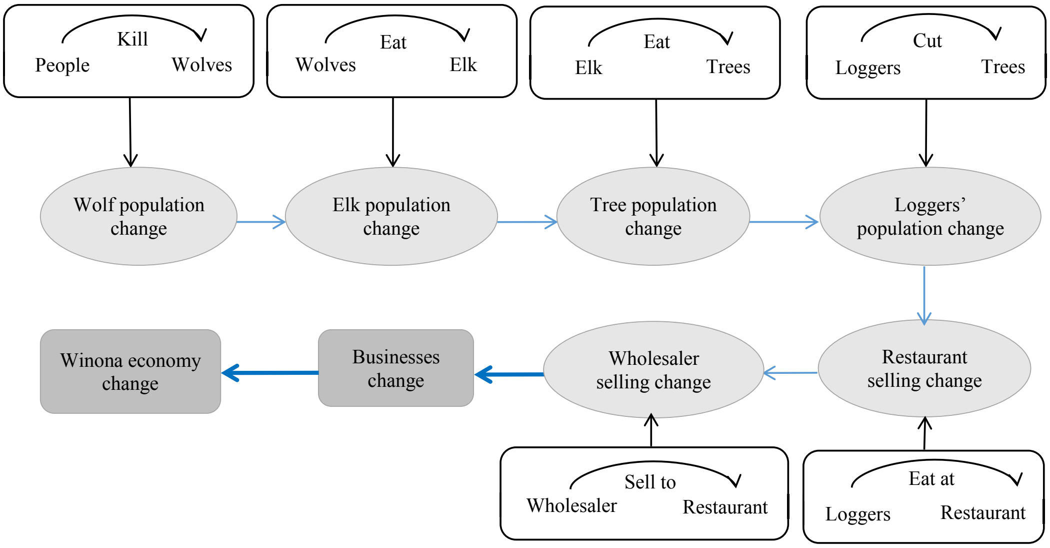

Fig. 3. Illustration of ecosystem-economy model in the wolf policy decision letter.

Table 3
ELLs' performance on the Gates-MacGinitie reading comprehension test, speed of lexical access, and fourth-grade ISAT reading test in three intervention conditions.

Pretest	Intervention condition		
	Collaborative group work (N = 70)	Direct instruction (N = 68)	Wait-listed control (N = 72)
	M (SD)	M (SD)	M (SD)
Gates-MacGinitie reading[a]	16.24 (8.45)	15.38 (8.92)	13.46 (9.32)
Speed of lexical access[b]	53.71 (12.03)	54.86 (12.20)	52.32 (11.32)
Fourth-grade ISAT reading[c]	211.34 (21.00)	209.33 (21.77)	203.50 (20.15)

Note. M = mean, SD = standard deviation.

[a] Gates-MacGinitie reading score was corrected for guessing (final score = right − wrong/3).

[b] Number of words that students correctly named per minute in the object naming task.

[c] ISAT is the Illinois Standards Achievement Test. Fifty participants (24%) did not take the ISAT due to high student mobility in the cooperating schools, including 20 CG students, 14 DI students, and 16 control students.

Table 4
Performance on language production and story elements in post-test oral narrative task.

	Intervention condition		
	Collaborative group work (N = 70)	Direct instruction (N = 68)	Wait-listed control (N = 72)
	M (SD)	M (SD)	M (SD)
Language production[a]			
Length	49.35 (20.17)	48.91 (23.21)	51.66 (16.41)
Syntactic complexity	53.73 (21.03)	50.75 (18.67)	45.67 (19.63)
Fluency	49.34 (17.76)	51.97 (18.48)	49.80 (20.70)
Mazes	51.42 (23.15)	49.90 (18.44)	48.71 (18.22)
Errors & Omissions	46.83 (12.94)	45.81 (13.97)	54.58 (23.17)
Story elements[b]			
Setting	2.56 (0.93)	2.51 (0.92)	2.38 (0.76)
Initiating events	2.50 (0.74)	2.13 (0.83)	2.42 (0.80)
Internal responses	5.47 (3.20)	5.09 (3.19)	4.93 (2.85)
Attempts	3.93 (2.45)	3.72 (2.34)	3.43 (2.03)
Outcomes	3.44 (2.10)	2.57 (1.76)	2.88 (1.47)
End State	3.10 (0.97)	3.22 (0.94)	2.86 (0.91)
Essential story elements[c]	21.00 (6.64)	19.25 (6.98)	18.89 (5.62)
Actions	5.16 (2.71)	4.54 (2.67)	6.08 (2.48)
Minor events	11.70 (6.64)	12.15 (6.95)	12.60 (4.98)

Note. M = mean, SD = standard deviation.

[a] Means and standard deviations of percentaged language factor scores.

[b] Means and standard deviations of number of non-repetitive communication units.

[c] Essential Story Elements = sum of descriptions in six story schema categories: setting, initiating event, internal response, attempt, outcome, and end state.

identify the aspects of oral language production that were influenced by instructional condition. Predictors that were significant in the MANOVA were included in the models whereas non-significant predictors were dropped. Provided along with tests of significance are standardized effect sizes, δ, estimated as the mean difference between treatment and control group divided by the square root of the between-subjects variance of the outcome variable (Spybrook et al., 2011). Two significant effects of instructional condition were observed.

First, instructional condition had an effect on *syntactic complexity*, $F(2, 188) = 3.19$, $p = 0.043$. CG students' narratives had more complicated syntax than control student' narratives, mean difference $M_{diff} = 8.06$, $t(188) = 2.31$, $p = 0.022$, $\delta = 0.37$. DI students also displayed more complicated syntax than control students, $M_{diff} = 5.08$, $t(188) = 2.03$, $p = 0.044$, $\delta = 0.33$. However, there was no difference between CG and DI, $M_{diff} = 2.98$, $t(188) = 0.27$, $p = 0.79$.

Second, instructional condition affected *errors and omissions*, $F(2, 188) = 3.45$, $p = 0.034$. DI students produced fewer errors and omissions than the control students, $M_{diff} = -8.77$, $t(188) = -2.50$, $p = 0.013$, $\delta = 0.41$, while there was no significant difference between CG and control, $M_{diff} = -7.75$, $t(188) = -1.89$, $p = 0.061$, $\delta = 0.30$, or between CG and DI, $M_{diff} = 1.32$, $t(188) = 0.65$, $p = 0.52$.

3.3. Essential story elements

The aggregate Essential Story Elements score and number of non-repetitive C-units in each story schema category are presented in Table 4. Factors influencing Essential Story Elements were evaluated in a two-level regression analysis with classroom as the second-level factor. Individual-level predictors included class-mean centered reading comprehension, class-mean centered speed of lexical access, and talkativeness. Classroom-level predictors were class-averaged reading and speed of lexical access. Instructional condition, wave, age of acquisition of English, parent education level, and whether or not students had an IEP were entered as fixed effects. The results indicated a significant condition effect on Essential Story Elements, $F(2, 185) = 3.40$, $p = 0.035$. CG students outperformed control students, $M_{diff} = 2.11$, $t(185) = 2.44$, $p = 0.016$, $\delta = 0.41$. CG students also outperformed DI students, mean difference $M_{diff} = 1.75$, $t(185) = 2.02$, $p = 0.045$, $\delta = 0.34$. However, there was no difference between DI and control students, $M_{diff} = 0.36$, $t(185) = 0.40$, $p = 0.69$. IEP student' stories contained fewer Essential Story Elements than non-IEP students, $F(1, 185) = 4.25$, $p = 0.041$. No difference was observed between the two waves of intervention, $F(1, 185) = 3.03$, $p = 0.083$, $\delta = 0.24$. The between-classrooms effect was not significant. Neither class-level nor individual-level reading nor speed of lexical access were significant.

Follow up tests of individual story schema categories, involving either two-level Poisson regression analysis or two-level negative binomial regression analysis, depending on the distribution of the counts of story elements, indicated that CG students significantly exceeded students in the other two conditions in elaborations of *outcomes*, M_{diff} (CG *vs. DI)* $= 0.87$, $t(188) = 2.98$, $p = 0.003$ and M_{diff} (CG *vs. Control)* $= 0.56$, $t(188) = 2.10$, $p = 0.037$.

Students who told longer stories may have had a greater chance to produce Essential Story Elements, but it is also possible that long stories contained extensive descriptions of minor events. We reanalyzed Essential Story Elements including the length factor from the language production analysis as a covariate. After controlling for length, the condition difference remained significant and the effect was larger, $F(2, 184) = 8.51$, $p < 0.001$. CG students outperformed control students, $t(184) = 4.12$, $p < 0.001$, $\delta = 0.52$, and DI students, $t(184) = 2.28$, $p = 0.024$, $\delta = 0.29$. There was no significant difference between DI students and control students after controlling for length, $t(184) = 1.80$, $p = 0.073$, $\delta = 0.23$. Students from the second wave produced more Essential Story Elements than students from the first wave, $F(1, 184) = 6.77$, $p = 0.010$, $\delta = 0.27$. No difference between students who did or did not use English in the first grade was observed, $F(1, 185) = 0.27$, $p = 0.61$, but non-IEP students performed better than IEP students, $F(1, 184) = 5.56$, $p = 0.019$. No other covariates were significant.

No condition difference was found in description of minor events, $F(2, 199) = 0.31$, $p = 0.73$, and there was no effect of age of acquisition of English, $F(1, 199) = 0.66$, $p = 0.42$. Students with higher speed of lexical access generated more minor events, $F(1, 199) = 5.19$, $p = 0.024$.

3.4. Multi-link causal chains

3.4.1. Multi-link causal chains in policy decision letters

A total number of 127 multi-link reasoning chains were identified among 179 ELLs' decision letters. Eighty-four (47%) students produced at least one causal chain, including 42 CG students, 40 DI students, and two control students. Among students who generated causal chains, the mean length of causal chains was 3.90 links in the CG condition and 3.41 links in the DI condition.

A multinomial logistic regression was conducted to examine instructional effects on multi-link reasoning in the policy decision letters. The covariates included age of acquisition of English, reading comprehension, speed of lexical access, talkativeness, and length of decision letter (total number of words). The length of the longest causal chain was treated as the outcome variable, which consisted of five categories: 1–2 links ($n = 95$), 3 links ($n = 46$), 4 links ($n = 17$), 5 links ($n = 12$) and 6–8 links ($n = 9$). Only two control students produced a chain in their letters, so we report only the differences between the CG and DI conditions and used the DI condition as the reference category. We used 1–2 links as the reference category for length of causal chains.

The results indicated a significant condition effect, $\chi^2 (8) = 20.95$, $p = 0.007$, and an effect of age of acquisition of English favoring those who spoke English in the first grade, $\chi^2 (4) = 10.13$, $p = 0.038$. DI students were more likely to generate 3-link chains (*odds ratio* DI/CG = 1.40) and 4-link chains (*odds ratio* DI/CG = 1.53), as compared to the CG students. However, CG students had a higher probability of generating chains with 5 links (*odds ratio* CG/DI = 2.25) and 6–8 links (*odds ratio* CG/DI = 3.81) than the DI students. Reading comprehension significantly predicted the production of causal chains in decision letters, $\chi^2 (4) = 9.59$, $p = 0.048$, while speed of lexical access did not, $\chi^2 (4) = 9.16$, $p = 0.057$. Neither talkativeness nor length of essay predicted chain length, *ps > 0.30*.

3.4.2. Multi-link causal chains in oral narratives

A multinomial logistic regression analysis was also conducted to examine whether there was an intervention effect on the length of causal chains in the children's frog stories. The covariates in this analysis were age of acquisition of English, reading comprehension, speed of lexical access, talkativeness, and length of narrative. Number of links in the longest causal chain in a student narrative was treated as the outcome variable. There were four categories: 1 link ($n = 117$), 2 links ($n = 48$), 3–4 links ($n = 34$), and 5–7 links ($n = 11$). We used 2 links as the reference category for number of links in chains and the control group as the reference category for instructional condition.

The results indicated a significant condition effect, χ^2 (6, $N = 210) = 14.39$, $p = 0.026$. CG students had a higher probability of generating chains with 5–7 links (*odds ratio* CG/control = 13.69 vs. DI/control = 6.18), whereas DI students had a higher probability of generating 1–2 link chains (*odds ratio* CG/control = 1.17 vs. DI/control = 2.56). The two conditions were comparable in generating 3–4 link chains (*odds ratio* CG/control = 0.97 vs. DI/control = 0.85). The predicted probability of generating multi-link chains of different lengths as a function of condition are depicted in Fig. 4. The figure shows that CG students were more likely to connect several story events into causal chains while DI and control students tended to describe each story event separately.

3.5. Classroom talk during the Wolf Unit

To document an important feature of the classroom talk during collaborative group work and direct instruction, as well as gauge implementation of the Wolf Unit, we compared students' and teachers' use of academic vocabulary in 146 4-min episodes that were systematically sampled from the roughly 500 h of videos of Wolf Unit lessons recorded in the CG and DI classrooms (lessons in Control classrooms were not recorded). One excerpt was sampled from each of six important lessons (seven in one case) in each CG classroom (N = 12) drawn from the following: introduction to the Wolf Unit, first Collaborative Reasoning discussion of the ‘big question,’ wolves in the United States, wolves and the ecosystem, a poster presentation of major concepts in the ecosystem domain, and the second Collaborative Reasoning discussion of the ‘big question.’. Likewise, one excerpt was sampled from each of six important lessons (seven in one case) in each DI classroom (N = 12) to cover the introduction to the Wolf Unit, wolves in the United States, wolves and the ecosystem, wolves and the economy, wolves and public policy, and review of major concepts in the Wolf Unit.

Rate of use of academic vocabulary was calculated based on a search of the 146 4-min episodes for uses by teachers and students of any of the 76 academic vocabulary words listed in the Wolf Unit glossary – such as *omnivore, extinction, economy, balance*. If participants were on-task and actually studying the Wolf Unit, uses of academic vocabulary should be evident. If participants played the roles assigned to them, teacher use of the terms should be higher in DI classrooms and student use of the terms higher in CG classrooms. Fig. 5 shows the rate per minute of any of the academic vocabulary words by teachers and students in the twelve CG and twelve DI classrooms. The classrooms are ordered from the lowest rate to the highest rate of use within condition. The expected differences between CG and DI classrooms are readily apparent. Among students, the rate per minute of academic vocabulary was over twice as high in CG classrooms (M_{CG} = 2.34, *SD* = 1.28) as compared to DI classrooms (M_{DI} = 1.01, *SD* = 0.72). Among teachers, the rate of use was four times higher among DI teachers (M_{DI} = 1.44, *SD* = 0.68) than CG teachers (M_{CG} = 0.33, *SD* = 0.25). We built a two-level Poisson model to compare condition differences in the use of academic vocabulary, with classroom as the second level factor, 4-min episodes within classrooms as first-level factor, and duration of talk as a first-level covariate. The results indicated that, after controlling for duration of talk, students in the CG condition produced academic vocabulary words at a higher rate than students in the DI condition, *F*(1, 122) = 16.56, $p < 0.001$, and teachers in DI condition produced the words at a higher rate than teachers in the CG condition, *F*(1, 122) = 33.36, $p < 0.001$. Overall, in CG classrooms, students in aggregate produced 86% of the academic vocabulary while 14% was produced by teachers. In contrast, in DI classrooms teachers produced 60% of the academic vocabulary whereas the students in aggregate produced 40%. The foregoing analysis of the rate of use of academic vocabulary was completed with all CG and DI classrooms [*N* = 24] in the larger project. When the sample is restricted to classrooms with a high proportion of ELLs [*N* = 12], the pattern and magnitude of the significant effects stayed the same.

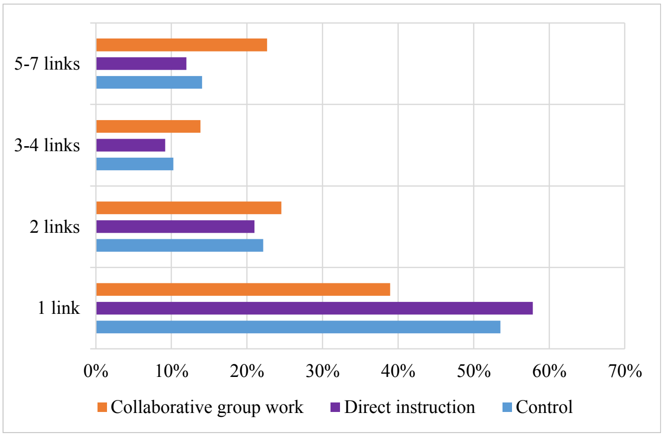

Fig. 4. Probability of generating multi-link reasoning chains of different lengths in storytelling task by instructional condition.

To give an impression of discourse features of classroom talk during the Wolf Unit, we selected two 4-min excerpts, one from a DI classroom and one from a CG classroom, that we judge to be representative of the discourse in the DI and CG classrooms. The excerpts were from classes that represent mid-level performance in terms of rate of use of academic vocabulary.

In the excerpt from a direct instruction classroom, the teacher comments on a student's idea and asks everyone to evaluate its plausibility. Some students support the idea, although in face of the strong implication from the teacher suggesting the opposite, they abandon their turns before they express themselves completely. The teacher explains her thinking and guides students to follow her. The contrary idea is never mentioned again by the students or the teacher. This is an example of thoughtful dialogue inasmuch as the teacher picks up on and responds to student ideas about a fundamental issue in the unit.

> Teacher Okay. That is a possibility but based on the facts uh the facts and the research that we've been doing, is that likely?
>
> Student 1 Yes.
>
> Student 2 Unless you do something, it's not gonna …
>
> Teacher Is it?
>
> Student 3 It's outrageous.
>
> Student 4 Yeah because they said (in the um, the um) one of those things about (wolves) wolves …
>
> Teacher The people are afraid of the wolves. That is a fact. They (they) do think they're dangerous, but based on the research and the numbers that we looked at, are they really, truly a threat? There've been instances but are those instances, do they occur frequently or not frequently?
>
> Students No. Uh-uh. Not frequently.

The following excerpt is from a collaborative group work classroom. Student 1 states a position and is immediately challenged by Student 2. Student 1 then provides a reason to support his position. His idea is picked up by Student 3 and further developed into a chain of reasoning. Student 4 supports the emerging idea. Student 5 starts to express the counter-position, but before she can finish, Student 4 asks a challenging, “Why?” Student 5 provides a reason to support her viewpoint. The excerpt illustrates a frequent pattern in CG classrooms in which students challenge

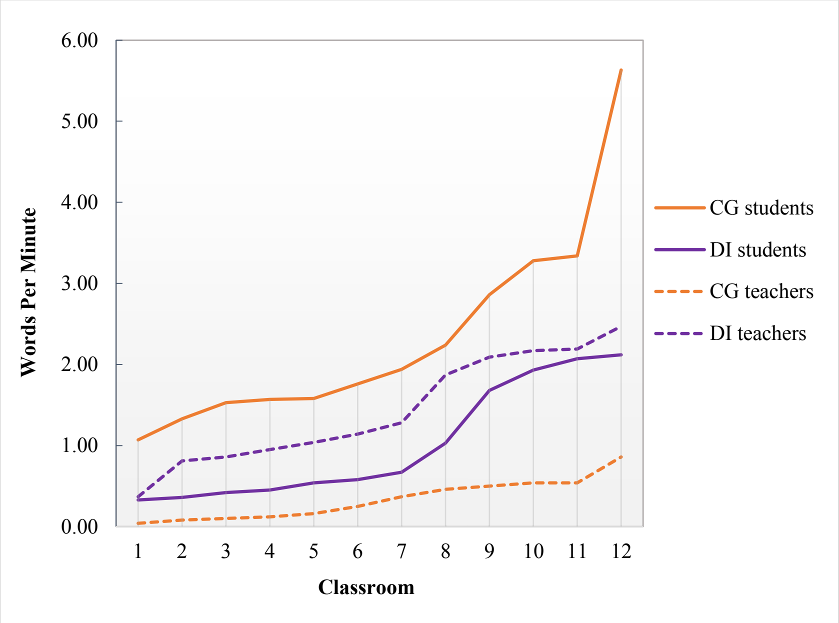

Fig. 5. Students' and teachers' academic vocabulary words per minute by classroom within condition. The classrooms are ordered from the lowest rate to the highest rate of use within condition. The figure shows all 12 collaborative group work (CG) and 12 direct instruction (DI) classrooms in the larger project.

each other, ask for elaboration and explanation, express complicated ideas, build on each other's thinking, and co-construct ideas.

Student 1 I think that we shouldn't.

Student 2 Why?

Student 1 Because you just hurting the, population of (of) the wolves.

Student 3 I say they shouldn't because, um, the wolves kill the elk, and when the elks are dead, there are more trees, so there's more oxygen for us.

Student 4 I was going to say that.

Student 3 So the wolves make it better for us.

Student 5 Hm. I- I think they should hire people to …

Student 4 Why?

Student 5 Becaaause, the wolves are just killing all the animals.

4. Discussion

The goal of this study was to determine whether contrasting instructional approaches have differential effects on English language learners' oral language development. The major finding was that fifth grade Spanish-speaking ELLs who participated in collaborative groups during a six-week unit on wolf reintroduction and management told more elaborated and coherent stories than comparable students who received direct instruction or were wait-listed controls. As compared to students in the other two conditions, students who had interacted in collaborative groups constructed stories that contained significantly more explanation of essential story elements and generated causal chains connecting story elements that contained significantly more links. The concepts and vocabulary needed to tell a story from the wordless picture book, *Frog, Where Are You,* which served as the story prompt, bear little relationship to the specific concepts and vocabulary taught in the Wolf Unit. Thus, the superior performance of students who participated in collaborative groups implies that they acquired, or further developed, some generalized competencies in language and thought.

With respect to basic features of language production, the overall difference between instructional conditions was significant. As compared with control students, both CG students and DI students told stories with more complicated syntax. DI students also made fewer omissions and word- and utterance-level errors than control students. There were no significant differences between CG and DI students on any of the basic language production measures.

With respect to story elaboration and completeness, CG students outperformed DI and control students in the production of Essential Story Elements with or without a control for story length. The better performance of CG students in Essential Story Elements indicates that students who had six-weeks of collaborative group work told stories with more description and explanation of the critical events that constitute the main theme of the story, which implies that CG students had begun to learn that to communicate successfully one has to focus on the central ideas and explain them fully, which is at the heart of communicative competence.

Regarding multi-link causal reasoning, CG students made more connections between events than DI or control students. In the policy decision letter about whether a pack of wolves should be killed, only two control students produced a causal chain with as many as three links. Control students simply did not have the background for extended reasoning about the wolf policy question. Comparable percentages of CG students (60.0%) and DI students (61.5%) wrote wolf decision letters containing causal chains with three or more links, but the chains in the letters of CG students were significantly longer. CG students had a greater likelihood of generating chains with 5–8 links whereas DI students had a greater likelihood of generating chains containing 3 or 4 links. When telling a story about a boy, a dog, and a frog, CG students had a higher probability of generating chains with 5–7 links whereas DI students had a higher probability of generating 1–2 link chains.

These findings support the conclusion that students who participated in collaborative groups developed a generalized competence in causal reasoning, because they displayed longer causal chains in both the frog story and the wolf decision letter. In contrast, the findings indicate that the multi-link reasoning of direct instruction students was restricted to the wolf decision letter, which was a recapitulation of the unit that they had studied for six weeks; in the frog story, a task involving a different set of facts than the Wolf Unit, they produced no more causal connections than control students. Our hypothesis to explain these findings is that while direct instruction students worked with the facts during the Wolf Unit, they were not fully responsible for identifying and expressing the connections among these facts. Students participating in collaborative groups constantly had to explain and justify their reasoning about relationships among facts to their classmates, which evidently was the key to developing a generalized ability and disposition to engage in explicit causal reasoning.

In collaborative discussions, students act as providers as well as receivers of information, as was illustrated in the dialogue excerpted from a lesson in a CG classroom (see p. 29). One person's statement is likely to be extended or evaluated by other students. If the statement is not clear, clarification may be requested. If students agree on the same idea, supplementary evidence may be offered to support the agreed upon point of view. If students disagree with one another, counterarguments and rebuttals will be made to support different opinions (Reznitskaya et al., 2009). One student's talk is usually extended by other students through making connections between their own opinions and the other persons' ideas (Lin et al., 2015). In comparison, DI students are less likely to extend the talk of the teacher or peers. Usually students in DI classrooms are only receivers of information. They have few opportunities to initiate ideas. Teachers do much of the talking and students just answer questions, usually with answers that are brief and unelaborated. The thinking required for extended talk may be suppressed in teacher dominated lessons (Nystrand & Gamoran, 1991).

Probably, the major reason that students in the CG condition generated more coherent stories, as indicated by longer multi-link chains, than students in the other two conditions is that collaborative group work provided more opportunities to use language to make connections. Morris et al. (2013) examined the frequency of use of the coordinating conjunctions *because, so, if, then, and*, and *but*, which are low-inference indicators of connected talk and relational thinking (Lin et al., 2015). Students' rate of use of coordinating conjunctions was over four times higher in CG classrooms (5.52 per minute) than DI classrooms (1.15 per minute) enrolled in the present study. Thus, it is highly plausible that CG students generated more elaborated and connected wolf letters and frog stories because of the experience of expressing elaborated and connected ideas during collaborative group work. DI students, in contrast, depended on teachers to initiate ideas and make connections and the students were left with only small pieces to add to a narrative largely told by the teacher.

CG students are encouraged to elaborate not only *what* and *how* but also *why* (Clark et al., 2003). CG students are allowed to freely express ideas and are expected to provide supporting reasons and evidence during discussions. Such experience stimulates students to generate more convincing arguments and fuels the development of multi-link reasoning and other forms of relational thinking (Lin et al., 2012; Reznitskaya et al., 2009). A well-structured story has logically arranged event sequences based on explicit statements of the relationships between events. With the elaboration of goals and outcomes, students are more capable of creating causal links between events. “Understanding why and under what circumstances people act on beliefs, true or false, may well be part of a more general scheme of understanding what causes events, states, states changes, and actions” (Trabasso et al., 1992, p. 164). The appreciation that one must strive to explain the causal relationships between events, rather than simply say what happened, is probably foundational for the development of causal reasoning ability.

Turning to the elemental language production measures, our theory was that CG students would perform better than DI students, because of more extensive opportunities for student talk during the Wolf Unit in CG classrooms. Contrary to expectation, CG students did not significantly exceed DI students on any of the language production measures, although both groups told stories with greater syntactic complexity than control students and the direct instruction group also had fewer omissions and errors. Perhaps the explanation is that DI teachers, who spoke more often and at greater length than CG teachers, provided better models of language use than ELL peers whose imperfect English CG students were listening to most of the time. In other words, superior input from DI teachers may have compensated for the fact that there were fewer opportunities for student output in DI classrooms (Swain, 2005). Another possibility is that DI teachers were positioned to provide scaffolding and feedback that helped students improve their English, again compensating for less student talk in DI classrooms. Further analysis of classroom dialogue during the Wolf Unit may help sort out these possibilities.

Of the previous studies of instructional interventions to accelerate ELLs' language development, the present study was consistent with the similar study by Zhang et al. (2013) with regard to the superior performance of fifth-grade ELLs who participated in collaborative group work in telling more complete and coherent stories. The difference is that Zhang and associates employed the Narrative Scoring Scheme (Miller & Heilmann, 2009) whereas the present study employed Essential Story Elements as the measure of story completeness, supplemented with the analysis of multi-link causal chains to more fully represent story coherence. Findings of the two studies diverged with respect to elemental features of language production. The present study found that ELLs who had participated in collaborative interaction told stories with greater syntactic complexity, but there was no hint of an effect on syntactic complexity in Zhang et al.

Several characteristics of students affected features of their stories. Students who acquired English early [English only in the first grade versus Spanish only or a mixture of Spanish and English in the first grade] produced longer stories, showed higher verbal fluency, and had fewer repetitions and revisions. Birdsong (2005) concluded that the chances for native-like attainment of a second language decrease with age of acquisition. In this study, age of acquisition of English had more effect on the language production measures than on the story completeness measure or multi-link reasoning. Thus, it seems that age of acquisition may have a stronger effect on basic linguistic proficiency than on communicative competence or reasoning.

A limitation of this study is that we did not give a pretest measure of storytelling, or any other pretest measure of oral language production, beyond the level of discrete words assessed by asking children to rapidly name common objects. A pretest measure of oral discourse production no doubt would have explained additional variance in storytelling and enabled more sensitive tests of intervention effects. Another limitation is that, although the study was fairly large compared to most previous intervention studies with ELLs, 18 classrooms is at the lower margin for fitting multi-level models that account for teacher and cohort effects. Classroom-level predictors explained some variance in language production measures but effects at the classroom level were weakly estimated.

Overall, collaborative group work improved Spanish-speaking ELLs' oral narrative skills. They expressed more complicated ideas, as suggested by the greater syntactic complexity of their utterances, although the utterances of students who experienced direct instruction also had greater syntactic complexity than the utterances of control students. Stories produced by students who had interacted in collaborative groups were more complete, coherent, and causally structured. They were more likely to elaborate essential story elements and organize events into causal chains than either control or direct instruction students. All these gains suggest that collaborative group work may be a promising approach to promote ELLs' language proficiency and, at the same time, aspects of their cognitive development.

The likely reason for the generally superior performance of CG students is that collaborative discussion provides more opportunities for high quality student talk than the teacher-dominated discourse prevalent during direct instruction. An indicator of quality talk is use of academic language. During the Wolf Unit CG students had twice as high a rate as DI students in use of academic vocabulary words.

For children who are second language learners and cannot always express themselves well in their new language, it is natural to conclude that they are facing language difficulties rather than having thinking problems. However, we cannot assume that children's natural ability to think will come out as soon as they know enough words and have control of the grammar of the second language. Abundant time is spent trying to improve basic language skills of second language learners while much less attention is given to fostering their thinking or enabling them to acquire science, social science, art, and humanities concepts (Moll, 2010). Meanwhile, native speakers are getting the chance to improve their thinking and conceptual understanding as well as their language. Bilingual educators should recognize the synergy that comes from the co-evolution of elemental language skills, communicative competence, thinking and reasoning, and conceptual understanding.

Acknowledgement

The research reported in this paper was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A080347 to the University of Illinois at Urbana-Champaign, Richard C. Anderson, Principal Investigator. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The authors are pleased to acknowledge the contributions of Nikisha Blackmon, Aini Marina Ma'rof, and Xiaoying Wu.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.learninstruc.2016.12.004.

References

Anderson, R. C., Chinn, C., Waggoner, M., & Nguyen-Jahiel, K. (1998). Intellectually stimulating story discussions. In J. Osborn, & F. Lehr (Eds.), *Literacy for all: Issues in teaching and learning* (pp. 170–186). New York: Guilford.

Arreaga-Mayer, C., & Perdomo-Rivera, C. (1996). Ecobehavioral analysis of instruction for at-risk language minority students. *Elementary School Journal, 96*(3), 245–258.

August, D., Francis, D., Hsu, H.-Y. A., & Snow, C. (2006). Assessing reading comprehension in bilinguals. *Elementary School Journal, 107*(2), 221–238.

August, D., McCardle, P., & Shanahan, T. (2014). Developing literacy in English language learners: Findings from a review of the experimental research. *School Psychology Review, 43*(4), 490–498.

August, D., & Shanahan, T. (2008). *Developing reading and writing in second-language learners: Lessons from the report of the national literacy Panel on language-minority children and Youth*. Mahwah, NJ: Lawrence Erlbaum Associates.

Avila, E., & Sadoski, M. (1996). Exploring new applications of the keyword method to acquire English vocabulary. *Language Learning, 46*(3), 379–395.

Aydede, M. (Fall 2010). Edition. In E. N. Zalta (Ed.), *The language of thought hypothesis, the stanford encyclopedia of philosophy*. retrieved from http://plato.stanford.edu/archives/fall2010/entries/language-thought.

Baker, C. (2011). *Foundations of bilingual education and bilingualism (* (5th ed.). Clevedon, UK: Multilingual Matters.

Bandura, A. (1986). *Social foundations of thought and action: A social-cognitive view*. Englewood Cliffs, NJ: Prentice-Hall.

Barksdale-Ladd, M. A., & Thomas, K. F. (2000). What's at stake in high-stakes testing – teachers and parents speak out. *Journal of Teacher Education, 51*(5), 384–397.

Berman, R. A., & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Birdsong, D. (2005). Interpreting age effects in second language acquisition. In J. F. Knoll, & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 109–127). New York: Oxford University Press.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology, 54*(2), 199–231.

Candelaria, M. A., & Llorente, A. M. (2009). The assessment of the Hispanic child. In C. Reynolds, & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 401–424). US: Springer. http://dx.doi.org/10.1007/978-0-387-78867-8_18.

Cazden, C. (2011). Dell Hymes's construct of "communicative competence." *Anthropology & Education Quarterly, 42*(4), 364–369.

Chambliss, M. J., & Calfee, R. C. (1998). *Textbooks for learning: Nurturing children's minds*. Malden, Massachusetts: Blackwell Publishers.

Chinn, C. A., Anderson, R. C., & Waggoner, M. A. (2001). Patterns of discourse in two kinds of literature discussion. *Reading Research Quarterly, 36*(4), 378–411.

Clark, A., Anderson, R. C., Kuo, L., Kim, I. H., Archodidou, A., & Nguyen-Jahiel, K. (2003). Collaborative reasoning: Expanding ways for children to talk and think in school. *Educational Psychology Review, 15*(2), 181–198.

Cummins, J. (1986). Empowering minority students: A framework for intervention. *Harvard Educational Review, 56*(1), 18–36.

Dillon, J. T. (1988). *Questioning and teaching: A manual of practice*. New York: Teachers College Press.

Dong, T., Anderson, R. C., Kim, I. H., & Li, Y. (2008). Collaborative reasoning in China and Korea. *Reading Research Quarterly, 43*(4), 400–424.

Ellis, R. (2005). *Instructed second language acquisition: A literature review*. Wellington, New Zealand: Ministry of Education.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading, 10*(3), 301–322.

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in US schools: An overview of research findings. *Journal of Education for Students Placed at Risk, 10*(4), 363–385.

Gersten, R., & Baker, S. (2000). What we know about effective instructional practices for English-language learners. *Exceptional Children, 66*(4), 454–470.

Goldenberg, C. (1992). Instructional conversations: Promoting comprehension through discussion. *The Reading Teacher, 46*(4), 316–326.

Goldenberg, C. (1996). Latin American immigration and U.S. schools. *Social Policy Reports: Society for Research in Child Development, 10*(1).

Helman, L. (2009). Factors influencing second-language literacy development: A road map for teachers. In L. Helman (Ed.), *Literacy development with English learners: Research based instruction in grades K-6* (pp. 1–17). New York, NY: Guilford Press.

Hume, D. (1748/1999). In T. L. Beauchamp (Ed.), *An enquiry concerning human understanding*. Oxford, UK: Oxford University Press, Oxford Philosophical Texts.

Hymes, D. (1972). On communicative competence. In J. B. Pride, & J. Homes (Eds.), *Sociolinguistics: Selected readings* (pp. 269–293). Harmondsworth, UK: Penguin Books.

Illinois State Board of Education. (2000). *Work study session II: Illinois measure of annual growth in English*. Retrieved from http://www.isbe.state.il.us/board/meetings/2000-2002/dec00meeting/122000FINALIMAGE.doc.

Jadallah, M., Miller, B., Anderson, R. C., Nguyen-Jahiel, K., Archodidou, A., Zhang, J., et al. (2009). Collaborative Reasoning about a science and public policy issue. In M. McKeown, & L. Kucan (Eds.), *Bringing reading research to life: Essays in honor of Isabel L. Beck* (pp. 170–193). New York: Guilford Press.

Johnson, D. W., & Johnson, R. T. (2009). Energizing learning: The instructional power of conflict. *Educational Researcher, 38*(1), 37–51.

Kamps, D., Abbott, M., Greenwood, C., Arreaga-Mayer, C., Wills, H., Longstaff, J., … Walton, C. (2007). Use of evidence-based, small-group reading instruction for English language learners in elementary grades: Secondary-tier intervention. *Learning Disability Quarterly, 30*(3), 153–168.

Kelcey, B., & Carlisle, J. F. (2013). Learning about teachers' literacy instruction from classroom observation. *Reading Research Quarterly, 48*(3), 301–317.

Kieffer, M. J. (2012). Early oral language and later reading development in Spanish-speaking English language learners: Evidence from a nine-year longitudinal study. *Journal of Applied Developmental Psychology, 33*(3), 146–157.

Kim, I. H., Anderson, R. C., Miller, B., Jeong, J., & Swim, T. (2011). Influence of cultural norms and collaborative discussions on children's reflective essays. *Discourse Processes, 48*(7), 501–528.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational psychologist, 41*(2), 75–86.

Koslowski, B., & Masnick, A. (2010). Causal reasoning and explanation. In U. Goswami (Ed.), *The Wiley-Blackwell handbook of childhood cognitive development* (2nd ed., pp. 377—398). Oxford: Wiley-Blackwell.

Larsen-Freeman, D. (2013). Transfer of learning transformed. *Language Learning, 63*(S1), 107—129.

Lin, T.-J., Anderson, R. C., Hummel, J. E., Jadallah, M., Miller, B. W., Nguyen-Jahiel, K., et al. (2012). Children's use of analogy during collaborative reasoning. *Child Development, 83*(4), 1429—1443.

Lin, T.-J., Anderson, R. C., Jadallah, M., Nguyen-Jahiel, K., Kim, I.-H., Kuo, L.-J., et al. (2015). Social influences on the development of relational thinking during small-group discussions. *Contemporary Educational Psychology, 41*, 83—97.

Lin, T.-J., Ma, S., Zhang, J., Nguyen-Jahiel, K., Anderson, R. C., Morris, J. A., … Jadallah, M. (April, 2011). Nurturing conceptual understanding and systems thinking. In *Paper presented at the annual meeting of the american educational research association* (New Orleans, LA).

Loban, W. (1976). *Language development: Kindergarten through grade twelve*. Urbana, IL: National Council of Teachers of English.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie reading tests* (4th ed.). Itasca, IL: Riverside Publishing. Level 4, Form S.

Mackey, A., & Goo, J. (2012). Interaction approach in second language acquisition. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Chichester, UK: Wiley-Blackwell.

Magliano, J. P. (1999). Revealing inference process during text comprehension. In S. R. Goldman, A. C. Graesser, & P. van den Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 55—76). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mayer, M. (1969). *Frog, where are you?* New York: Dial Books for Young Readers.

McCaslin, M., Good, T. L., Nichols, S., Zhang, J., Wiley, C. R., Bozack, A. R., … Cuizon-Garcia, R. (2006). Comprehensive school reform: An observational study of teaching in grades 3 through 5. *Elementary School Journal, 106*(4), 313—331.

Miller, J. F., & Chapman, R. S. (2010). *Systematic analysis of language transcripts [computer software]*. Madison, WI: Language Analysis Laboratory, Waisman Center, University of Wisconsin.

Miller, J., & Heilmann, J. (2009). New tool assesses narrative structure. *Advance, 19*(21), 10—11.

Miller, J., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice, 21*(1), 30—43.

Misdraji-Hammond, E., Lim, N. K., Fernandez, M., & Burke, M. E. (2015). Object familiarity and acculturation do not explain performance difference between Spanish-English bilinguals and English monolinguals on the Boston Naming Test. *Archives of Clinical Neuropsychology, 30*(1), 59—67.

Moll, L. C. (2010). Mobilizing culture, language, and educational practices: Fulfilling the promises of Mendez and Brown. *Educational Researcher, 39*(6), 451—460.

Morris, J., Miller, B., Anderson, R. C., Lin, T.-J., Nguyen-Jahiel, K., Sun, J., … Wu, X. (2013). *Instructional discourse and argumentative writing*. Champaign, IL: Center for the Study of Reading, University of Illinois.

Murphy, P. K., Wilkinson, I. A. G., Soter, A. O., Hennessey, M. N., & Alexander, J. F. (2009). Examining the effects of classroom discussion on students' high-level comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*(3), 740—764.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington DC: National Institute of Child Health and Human Development.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology, 35*(4), 250—256.

Nystrand, M. (2006). Research on the role of classroom discourse as it affects reading comprehension. *Research in the Teaching of English, 40*(4), 392—412.

Nystrand, M., & Gamoran, A. (1991). Instructional discourse, students' engagement, and literature achievement. *Research in the Teaching of English, 25*(3), 261—290.

Piaget, J. (1976/1947). *The psychology of intelligence*. New York: Littlefield.

Pressley, M., & Wharton-McDonald, R. (1997). Skilled comprehension and its development through instruction. *School Psychology Review., 26*(3), 448—466.

Prevoo, M. J., Malda, M., Mesman, J., & van IJzendoorn, M. H. (2016). Within-and cross-language relations between oral language proficiency and school outcomes in bilingual children with an immigrant background: A meta-analytical study. *Review of Educational Research, 86*(1), 237—276.

Resnick, L. B., & Schantz, F. (2015). Talking to learn: The promise and challenge of dialogic teaching. In L. B. Resnick, C. S. C. Asterhan, & S. N. Clarke (Eds.), *Socializing intelligence through academic talk and dialogue* (pp. 441—450). Washington, DC: American Educational Research Association.

Reznitskaya, A., Anderson, R. C., & Kuo, L. J. (2007). Teaching and learning argumentation. *Elementary School Journal, 107*(5), 449—472.

Reznitskaya, A., Kuo, L., Clark, A., Miller, B., Jadallah, M., Anderson, R. C., et al. (2009). Collaborative reasoning: A dialogic approach to group discussions. *Cambridge Journal of Education, 39*(1), 29—48. http://dx.doi.org/10.1080/03057640802701952.

Roberts, T., & Neal, H. (2004). Relationships among preschool English language learner's oral proficiency in English, instructional experience and literacy development. *Contemporary Educational Psychology, 29*(3), 283—311.

Rojas-Drummond, S., & Mercer, N. (2003). Scaffolding the development of effective collaboration and learning. *International Journal of Educational Research, 39*(1), 99—111.

Saunders, W. M., & Goldenberg, C. (2007). The effects of an instructional conversation on English Language Learners' concepts of friendship and story comprehension. In R. Horowitz (Ed.), *Talking texts: How speech and writing interact in school learning* (pp. 221—252). Mahwah, NJ: Erlbaum.

Schunk, D. H., & Zimmerman, B. J. (2007). Influencing children's self-efficacy and self-regulation of reading and writing through modeling. *Reading & Writing Quarterly, 23*(1), 7—25.

Shapiro, L. R., & Hudson, J. A. (1991). Tell me a make-believe story: Coherence and cohesion in young children's picture-elicited narratives. *Developmental Psychology, 27*(6), 960—974.

Shepard, L. A. (2010). Next-generation assessments. *Science, 330*(6006), 890.

Silverman, R. D. (2007). Vocabulary development of English-language and English-only learners in kindergarten. *Elementary School Journal, 107*(4), 365—383.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, familiarity and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory, 6*(2), 174—215.

Snow, C. E. (2014). Input to interaction to instruction: Three key shifts in the history of child language research. *Journal of Child Language, 41*(S1), 117—123.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design software version 3.0*. Available from: www.wtgrantfoundation.org.

Stein, N. L., & Albro, E. R. (1997). Building complexity and coherence: Children's use of goal-structured knowledge in telling stories. In M. Bamberg (Ed.), *Narrative development: Six approaches* (pp. 5—44). Mahwah, NJ: Erlbaum.

Stein, M., Carnine, D., & Dixon, R. (1998). Direct instruction integrating curriculum design and effective teaching practice. *Intervention in School and Clinic, 33*(4), 227—233.

Stein, N. L., & Glenn, C. (1979). An analysis of story comprehension in elementary school children. In R. Freedle (Ed.), *New directions in discourse processing* (pp. 53—120). Norwood, NJ: Ablex.

Straus, M. A. (1980). *The ZP scale: A percentaged Z score*. Durham, NH: Family Research Laboratory, University of New Hampshire.

Sunderman, G. L., Kim, J. S., & Orfield, G. (2005). *NCLB meets school Realities: Lessons from the field*. Thousand Oaks, CA: Corwin Press.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471—483). Mahwah, NJ: Erlbaum Associates.

Swain, M., & Watanabe, Y. (2012). Languaging: Collaborative dialogue as a source of second language learning. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Oxford: Wiley-Blackwell.

Tharp, R., & Gallimore, R. (1989). *Rousing minds to life*. New York: Cambridge University Press.

Trabasso, T., Stein, N. L., Rodkin, P. C., Park Munger, M., & Baughn, C. R. (1992). Knowledge of goals and plans in the on-line narration of events. *Cognitive Development, 7*(2), 133—170.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory and Language, 24*(5), 612—630.

Trabasso, T., van den Broek, P., & Suh, S. Y. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes, 12*(1), 1—25.

United States Department of Education. (2016). *Digest of education statistics. National center for education statistics*. Retrieved from https://nces.ed.gov/programs/digest/d15/tables/dt15 204.27.asp.

Vaughn, S., Cirino, P. T., Linan-Thompson, S., Mathes, P. G., Carlson, C. D., Hagan, E. C., et al. (2006). Effectiveness of a Spanish intervention and an English intervention for English language learners at risk for reading problems. *American Educational Research Journal, 43*(3), 449—487.

Vygotsky, L. S. (1978). *Mind in society*. Cambridge, MA: Harvard.

Webb, N. M., & Mastergeorge, A. M. (2000). The development of students' helping behavior and learning in peer-directed groups. *Cognition and Instruction, 21*(4), 361—428.

Wells, G., & Arauz, R. M. (2006). Dialogue in the classroom. *Journal of the Learning Sciences, 15*(3), 379—428.

Zhang, J., Anderson, R. C., & Nguyen-Jahiel, K. (2013). Language-rich discussions for English language learners. *International Journal of Educational Research, 58*, 44—60.