Resubmitting to ERIC to acknowledge funding.

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

Educ Psychol Rev (2016) 28:1–22
DOI 10.1007/s10648-014-9295-x

REVIEW ARTICLE

Inference Instruction for Struggling Readers: a Synthesis of Intervention Research

Colby S. Hall

Published online: 7 January 2015

Abstract Skill in generating inferences predicts reading comprehension for students in the elementary and intermediate grades even after taking into account word reading, vocabulary knowledge, and cognitive ability (Cain et al., *Journal of Educational Psychology, 96*, 671–81, 2004; Kendeou et al., *Journal of Research in Reading, 31*, 259–72, 2008; Oakhill and Cain, *Scientific Studies of Reading, 16*(2), 91–121, 2012; Oakhill et al., *Language and Cognitive Processes, 18*, 443–468, 2003). While research shows that struggling readers are less likely than proficient readers to make inferences when reading text (Cain et al., *Memory and Cognition, 29,* 850–859, 2001; *Oakhill, British Journal of Educational Psychology, 54*, 31–39, 1984), struggling readers may also benefit more from inference instruction than do proficient readers (Hansen and Pearson, *Journal of Educational Psychology, 75*(6), 821–829, 1983; McGee and Johnson, *Educational Psychology, 23*(1), 49–59, 2003; Raphael and Pearson, *American Educational Research Journal*, *22*(2), 217–235, 1985; Yuill and Oakhill, *Applied Cognitive Psychology, 2,* 33–45, 1988). This synthesis assessed (a) the effectiveness of inference instruction in improving reading outcomes for struggling readers and (b) the features of instructional interventions (e.g., duration, type of instruction) that were associated with improved outcomes. One single-case design and eight experimental group design studies were synthesized. Mean effect sizes for group design studies ranged from g=0.72* to g=1.85* for researcher-developed measures of inferential reading comprehension and from g=−.03 to g=1.96* for standardized measures of reading comprehension. The percentage of non-overlapping data for the study that employed a single-case design was 100 % for all measures.

Keywords Reading comprehension · Inferences · Background knowledge · Learning disabilities · Reading difficulties

Reading with comprehension involves constructing a coherent representation, or *situation model*, of a text in memory (Graesser et al. 1994; Kintsch 1998). Rapp et al. (2007) describe a situation model as a "network, with nodes that depict individual facts and events, and connections that depict meaningful relations between them" (p. 292). Coherence reflects the degree to which appropriate, meaningful connections are established between information in

C. S. Hall (✉)
The Meadows Center for Preventing Educational Risk, The University of Texas at Austin, 1 University Station D4900, Austin, TX 78712, USA
e-mail: colbyhall@gmail.com

the text, and between information in the text and the reader's prior knowledge. These connections are known as *inferences*. To read with understanding, the reader not only has to remember literal details but also to generate inferences in order to discover implicit meanings and create a coherent situation model.

A number of correlational studies provide evidence that a student's skill in generating inferences is highly predictive of his or her reading comprehension. In a study of 4-, 6-, and 8-year-olds, inference skill explained variance in reading comprehension even after taking into account word reading and vocabulary knowledge (Kendeou et al. 2008). In another study of 8-, 9-, and 10-year-olds, inference and integration skill predicted reading comprehension after controlling for word reading, vocabulary, and cognitive ability (Cain et al. 2004).

The importance of inference generation skill for reading comprehension is reflected in the new Common Core State Standards (Common Core State Standards Initiative 2010). The standards not only require students to "read and comprehend literary and informational texts proficiently and independently" but also to determine central ideas or themes, to analyze how and why individuals, events, and ideas develop and interact, and to assess the ways in which point of view or purpose shapes the content and style of a text. Students are expected to analyze the implicit "how" and "why" of texts, not just to identify the explicit "who" and "what." For more information on the English Language Arts Common Core State Standards, see Fig. 1.

There is increasing evidence to suggest that struggling readers have particular difficulty generating inferences and that inference-making difficulty is actually a cause of comprehension failure. In a study reported by Oakhill (1982), young readers who were good comprehenders were reported to routinely integrate the meanings of successive sentences, while poor comprehenders were less likely to do so. In their research among 7- and 8-year-old readers, Cain and Oakhill (1999) determined that poor comprehenders made fewer inferences than a younger, comprehension-age match group. Because groups were matched on comprehension skill, this difference in inference generation between groups indicated that inference skill was not dependent on comprehension ability, but rather something that preceded comprehension gains.

There is also evidence that less proficient readers may benefit more from inference instruction than their proficient reader peers. McGee and Johnson (2003) found that less skilled readers between the ages of 6 and 9 years who were given explicit inference instruction showed a

CCSS.ELA-Literacy.CCRA.R.2: Determine central ideas or themes of a text and analyze their development; summarize the key supporting details and ideas.

CCSS.ELA-Literacy.CCRA.R.3: Analyze how and why individuals, events, or ideas develop and interact over the course of a text.

CCSS.ELA-Literacy.CCRA.R.6: Assess how point of view or purpose shapes the content and style of a text.

CCSS.ELA-Literacy.CCRA.R.10: Read and comprehend complex literary and informational texts independently and proficiently.

Fig. 1 English language arts standards: college and career readiness anchor standards for reading (Common Core State Standards Initiative 2010)

significantly greater improvement than did skilled readers who received the same instruction. Similarly, Raphael and Pearson (1985) determined that sixth-grade students at low and average reading ability levels benefitted more than readers at high reading ability levels from a Question–Answer Relationship inference instruction intervention. Hansen and Pearson (1983) and Yuill and Oakhill (1988) reported similar results among fourth and second graders, respectively.

Researchers have proposed a number of inference taxonomies (e.g., Graesser et al. 1994; Johnson and Pearson 1978; Kintsch 1993; Warren et al. 1979), and consensus as to a definitive taxonomy has not emerged. However, many have found it useful to distinguish between *text-connecting inferences* and *gap-filling inferences* (Cain and Oakhill 1999; Kispal 2008). Text-connecting inferences, sometimes called cohesive inferences, rely on linguistic cues present in the text. An example is anaphor resolution: In order to form a coherent situation model of the sentence, "Omar gave Veronica his jacket," the reader must infer that the "his*"* refers to Omar. Gap-filling inferences, on the other hand, require the reader to go beyond the text and draw on prior experience or background knowledge.

Many researchers distinguish between gap-filling inferences that are necessary for a basic understanding of the text and gap-filling inferences that are *not* strictly necessary. Bowyer-Crane and Snowling (2005) provide the following example of a necessary gap-filling inference: "The campfire started to burn uncontrollably. Tom grabbed a bucket of water" (p. 192). In order to understand why Tom grabbed a bucket of water, it is necessary for the reader to activate the background knowledge that water puts out fire, and relate the second sentence to the first by generating the inference that Tom grabbed the bucket of water because he was trying to put out the fire. In contrast, elaborative gap-filling inferences are not strictly necessary for comprehension; instead, they serve to enrich the reader's mental representation of a text. It is not always necessary for a reader to generate elaborative inferences about a character's personality based on that character's actions. Likewise, it is not always strictly necessary to infer about the character's motivations, goals, or fears. Nevertheless, to engage in this kind of elaboration may make reading a richer experience, and result in a more complete situation model.

There is some evidence that struggling readers have more difficulty with gap-filling inferences than with text-connecting inferences. Cain and Oakhill (1999) found that, for 7- and 8-year-olds, the ability of poor comprehenders to generate text-connecting inferences was parallel to that of skilled comprehenders when the students were given the opportunity to look back at the text. Similarly, Bowyer-Crane and Snowling (2005) determined that elementary-aged, less-skilled comprehenders as old as 11 answered questions requiring a text-connecting inference at a level comparable to their skilled peers. It was with questions that required students to make gap-filling inferences that the performances of the less-skilled and skilled groups diverged.

In 2007, the UK's National Foundation for Educational Research (NFER) was commissioned to conduct a review of research "on inference skills for reading, including the skills that constitute inferencing and how to teach them" (Kispal 2008). The review produced by NFER described one intervention study in detail and briefly mentioned six others. However, the results of these studies were not compared in any meaningful way, and none of the studies examined the impact of inference instruction intervention on struggling comprehenders. We were unable to locate any other synthesis that sought to gather, compare, and synthesize evidence gleaned from research on inference instruction. The purpose of this synthesis, then, was to examine all inference instruction intervention studies that targeted struggling readers. Research questions follow:

Research question 1	How effective is inference generation instruction in improving reading outcomes for struggling readers?
Research question 2	What features of instructional interventions (e.g., type of instruction, duration, grade level) are associated with improved outcomes?

In the context of this synthesis, a *struggling reader* refers to a reader who scored statistically significantly lower than his or her age-level peers on one or more tests of reading ability, with a $p<0.05$. Struggling readers also included students with a school-identified learning disability (LD). *Inference-making skill* was defined as skill in generating meaningful links between different parts of a text and/or using prior knowledge to fill in missing details (Cain 2010). *Inference generation instruction* refers to an instructional intervention designed explicitly to improve students' inference-making skill. When identifying the type of instruction, attention was paid to the type of inferences participants were asked to make (e.g., text-connecting or gap-filling), as well as the method of instruction used.

Method

Search Procedures and Inclusion Criteria

A two-step process was employed in order to identify studies for review. First, searches were conducted in the PsychINFO and ERIC databases using the key term *reading comprehension* or *reading* alongside of the key term *inferen** and at least one of the following ability terms: *learning disabilities*, *reading difficulties*, *below average readers*, *below average comprehenders*, *less skilled readers*, *less skilled comprehenders*, *poor readers*, *poor comprehenders*, *struggling readers*, *struggling comprehenders*, and *comprehension ability*. This search yielded 212 citations. Each of the abstracts was reviewed, and articles were excluded that did not meet inclusion criteria on the basis of information provided therein. A total of 41 articles that demonstrated the potential to meet inclusion criteria on the basis of information provided in the abstract were obtained, read, and evaluated. This resulted in eight studies that met the inclusion criteria. Next, references from studies that met criteria for inclusion and relevant chapters and articles addressing the topic of reading comprehension instruction were checked to identify additional studies that might meet criteria. Searched articles included eight meta-analyses and syntheses of reading comprehension instructional interventions for struggling readers (Berkeley et al. 2010; Edmonds et al. 2009; Flynn et al. 2012; Gersten et al. 2001; Scammacca et al. 2013; Solis et al. 2011; Wanzek and Vaughn 2007). Thirty more articles were evaluated based on this additional search, of which one met inclusion criteria. In all, nine articles were synthesized.

Seven inclusion criteria were used for selecting studies:

1. The study was reported in a peer-reviewed journal, in English, between the earliest indexed year of each database and 2013.
2. Participants were students who had been identified as struggling readers. Studies with additional participants were included if disaggregated data were provided for struggling readers.
3. Participants were enrolled in grades 1–12.
4. The primary purpose of the intervention was to improve inferential comprehension.
5. Dependent variables addressed inferential reading comprehension outcomes and/or global reading comprehension outcomes.
6. The intervention did not include multiple components (i.e., it did not include another type of instruction in addition to inference instruction), so that effects could be attributed to inference instruction rather than to some other cause.
7. The research design was experimental, quasi-experimental, or single case.

In order to comply with the second criterion, a study with additional, non-struggling reader participants needed to provide sufficient information for the purpose of calculating the significance of differences between groups or conditions for the subgroup of struggling readers. It was not enough to provide information about the significance of comparisons between larger groups that included participants who were not struggling readers and then to report that there was no significant interaction between reading ability and group membership. Several studies could not be included because they failed to meet this one criterion (Carr and Thompson 1996; Carr et al. 1983; Dewitz et al. 1987; McGee and Johnson 2003; Raphael and Pearson 1985; Raphael and Wonnacott 1985).

In order to satisfy the fourth criterion, a study needed to state, as its primary purpose, an aim to improve inference-making skill. In several studies considered for review, improved inference-making skill was a collateral benefit, but not the primary purpose, of an intervention (e.g., Gurney et al. 1990; Mastropieri et al. 1996); these studies were not included in the synthesis. Other studies were excluded because inference generation instruction was one of multiple intervention components, making it difficult to evaluate the isolated effects of inference instruction (e.g., Ezell et al. 1992; Mason 2004).

Coding

The code sheet used in the present synthesis included elements specified in the What Works Clearinghouse Design and Implementation Assessment Device (IES 2008) and was used in previous syntheses of research (Edmonds et al. 2009; Scammacca et al. 2013; Wanzek et al. 2006). Coded information included research design, criteria for identifying participants as struggling readers, age or grade level of participants, socioeconomic status of participants, type of instructional intervention, type of text used during the intervention, duration of intervention (i.e., total hours of instruction per student), setting, implementer of the intervention, information about fidelity of intervention implementation, outcome measures used, reliability and validity data related to the outcome measures, findings, and information about the clarity of causal inferences that were drawn.

Calculation of Effect Sizes

Studies Employing Group Designs Effect sizes were calculated for all studies that provided adequate statistical information, including means, standard deviations, group sizes, F test scores, t test scores, and/or exact p values. Effect sizes were calculated using what is now the most commonly used effect size index, the standardized mean difference known as Hedges's g. It is defined as the difference between the mean outcome for the intervention group and the mean outcome for the comparison group, divided by the pooled within-group standard deviation of the outcome measure. Procedures and formulas described in the *What Works Clearinghouse Procedures and Standards Handbook, Version 3.0* were employed (IES 2013).

In reporting on the magnitude of the effects achieved by individual studies, we cited effect sizes for comparisons between the treatment condition hypothesized to have the greatest potential to improve inference generation skill and the "business-as-usual" or typical practice comparison condition. If a typical practice comparison condition did not exist, we cited the effect size for the comparison between the treatment condition hypothesized to have the greatest potential to improve inference generation skill and the treatment condition that we judged to be closest to a typical practice comparison condition. We cited negative effects when a comparison condition outperformed a treatment condition.

If a study had more than one outcome in the domain of inferential reading comprehension, the effect sizes for all of those outcomes were combined into a study average effect using the simple, unweighted average of individual effect sizes. If a study employed a researcher-developed measure of inferential reading comprehension, then the mean effect size for the study reflects performance on this researcher-developed measure or the simple average of outcomes on multiple researcher-developed measures. If a study used only a standardized measure, then the mean effect size for the study reflected performance on this standardized measure. In accordance with the WWC (2013) standards, interventions with multiple outcomes within the domain of inferential reading comprehension were determined to have a "statistically significant positive effect" when at least half of the effects achieved were positive and statistically significant, and no effects were negative and statistically significant (p. 24).

Studies Employing Single Case Designs Effect sizes were calculated by determining the percentage of non-overlapping data (PND) between baseline and successive intervention phases (Scruggs et al. 1987). In order to calculate PND, it was necessary to first identify the highest data point in baseline, and then to determine the percentage of data points during intervention that exceeded this level. The interpretation of PND scores was as follows: (a) more than 90 % of PND reflected a highly effective treatment, (b) 70–90 % of PND reflected a fairly effective treatment, (c) 50–70 % of PND reflected questionable effectiveness, and (d) <50 % of PND reflected an unreliable treatment (Wendt 2009).

Results

One single case and eight experimental, group design inference instruction intervention studies were included in this synthesis. Tables 1 and 2 summarize study features and findings, including effect sizes. The first column of each table includes author names, publication date, design information, age of participants, and intervention dosage (i.e., the total number of hours of instruction and/or practice received by each student) for each study. Column 2 describes treatments administered to participants and reports the sample size for treatment groups. Column 3 describes comparison conditions and reports the sample size for comparison groups. Column 4 lists the dependent measures employed in each study. The final column summarizes study results, including effect sizes and indicating statistically significant differences between samples when they were reported. For example, S>CO: g=0.93* signifies that the S group outperformed the CO group on the dependent measure, the size of the effect was g=0.93, and the difference between groups was statistically significant. Statistical significance is denoted by an asterisk following the effect size.

Four studies (Holmes 1985; Winne et al. 1993; Yuill and Joscelyne 1988; Yuill and Oakhill 1988) focused primarily on teaching students to identify and integrate key words or phrases in text, with a much smaller focus, if any, on prior knowledge or experiences outside the text. One study (McMaster et al. 2012) investigated the effect of causal or general question prompts on integration of text. Three studies (Hansen and Pearson 1983; McCormick and Hill 1984; Ouellette et al. 1999) focused on teaching students to activate prior knowledge and integrate this knowledge with information in text in order to generate inferences. In one study, students were taught to activate prior knowledge as well as to identify and integrate key words or phrases in text (Fritschmann et al. 2007).

In four of the nine studies (Fritschmann et al. 2007; Hansen and Pearson 1983; McCormick and Hill 1984; Ouellette et al. 1999), students were taught to make gap-filling inferences that

Table 1 Summary of group intervention study characteristics and findings

Study • Design • Participant age • Dosage (total hours per student)	Treatment(s) (*n*=sample size)	Comparison (*n*=sample size)	Dependent measures	Results/ES (*g*)
Hansen and Pearson (1983) • Experimental (treatment comparison) • 4th graders • 24 sessions, length not reported	Prediction using prior knowledge instruction and practice answering post-reading questions in 7:3 ratio of inferential to literal questions (PP+Q; *n*=10)	English language arts instruction as usual; post-reading questions in 3:7 ratio of inferential to literal questions (CO; *n*=10)	Post-reading worksheets, inferential comprehension Posttest 1, inferential comprehension Posttest 2, inferential comprehension	PP+Q>CO: *g*=1.46* PP+Q>CO: *g*=0.98* PP+Q>CO: *g*=1.15*
Holmes (1985) • Experimental (multiple group comparison) • 4th and 5th graders • 2.67 h	Difficulty-level sequenced materials only (M; *n*=6) Directive inferencing strategy only (S; *n*=6) Strategy Plus Materials (S+M; *n*=6) English language arts instruction as usual and practice answering all inferential questions post-reading (QO; *n*=6)	No "business as usual" comparison condition	Post-reading inferential questions Nelson Reading Test, Form A	S+M>S: *g*=2.18* S+M>QO: *g*=1.85* S+M>M: *g*=1.69* M>QO: *g*=0.67 M>S: *g*=0.50 S>QO: *g*=0.30 S+M>QO: *g*=1.07* M>QO: *g*=1.05* M>S: *g*=0.99* S>QO: 0.74* S+M>S: g=0.58* S+M>M: *g*=0.02
McCormick and Hill (1984) • Experimental (multiple group comparison) • 5th graders • 26.67 h	Prediction using prior knowledge instruction with practice answering a 5:1 ratio of literal to inferential questions (PP+Q; *n*=23) Practice answering 6 inferential questions after reading (QO; *n*=25)	English language arts instruction as usual; practice answering 5:1 ratio of literal to inferential questions after reading (CO; *n*=26)	Post-reading questions second 5 weeks, inferential Post-reading questions third 5 weeks, inferential Post-reading questions fourth 5 weeks, inferential	PP+Q>CO: *g*=0.91* PP+Q>QO: *g*=0.48 QO>CO: *g*=0.42 QO>CO: *g*=1.49* PP+Q>CO: *g*=1.07* QO>PP+Q: *g*=0.16 QO>CO: *g*=1.49* PP+Q>CO: *g*=1.07* QO>PP+Q: *g*=0.37

Table 1 (continued)

Study • Design • Participant age • Dosage (total hours per student)	Treatment(s) (n=sample size)	Comparison (n=sample size)	Dependent measures	Results/ES (g)
			Posttest, inferential questions	QO>PP+Q: g=0.83* QO>CO: g=0.56 CO>PP+Q: g=0.17
			Metropolitan Achievement Test (Metro), Reading Achievement subtest	CO>PP+Q: g=0.03 PP+Q>QO: g=0.27 CO>QO: g=0.31
McMaster et al. (2012) • Experimental (multiple group comparison) • 4th graders • 15 h	Causal questioning (CQ; n=26) General questioning (GQ; n=16)	W-questioning (WQ; n=14)	Liberal items recalled	WQ>CQ: g=0.27 GQ>CQ: g=0.05 WQ>GQ: g=0.25
			Highly connected items recalled	WQ>CQ: g=0.46 GQ>CQ: g=0.17 WQ>GQ: g=0.30
			No-match, consistent with text items recalled	WQ>CQ: g=0.30 CQ>GQ: g=0.20 WQ>GQ: g=0.51
Ouellette et al. (1999) • Experimental (treatment-comparison) • 5th graders • 12 sessions, length not reported	During-reading discussions based on Hansen and Pearson's (1983) pre-reading prediction using prior knowledge instruc tion (PP; n=34)	English language arts instruction as usual (CO; n=33)	Metropolitan Achievement Test, 6th edition (MAT6)	PP>CO: g=0.41
			Story summary	PP>CO: g=0.36
Winne et al. (1993) • Experimental (treatment-comparison) • Outgoing 3rd, 4th, or 5th graders • 9 sessions, length not reported	Explicit, explanatory feedback given to students after they answered post-reading questions (EF; n=11) Inductive feedback given to students after they	No "business as usual" comparison condition	Inference question	EF>IF: g=0.82*

Table 1 (continued)

Study • Design • Participant age • Dosage (total hours per student)	Treatment(s) (n=sample size)	Comparison (n=sample size)	Dependent measures	Results/ES (g)
	answered post-reading questions (IF; n=13)			
Yuill and Joscelyne (1988) • Experimental (treatment-comparison) • 2nd graders • One session, length not reported	Explicit instruction in how to find and use "clue words" to discover implicit information (S; n=5)	Students read stories with no instruction and no post-reading feedback (CO; n=5)	Post-reading comprehension questions Recall	S>CO: g=0.93* CO>S: g=0.22
Yuill and Oakhill (1988) • Experimental (multiple group comparison) • 2nd graders • 13.5 h	Lexical inference skills instruction, question generation, and macro-cloze prediction exercises (S; n=13) Comprehension questions only, a mix of inferential and literal questions (QO; n=7)	Rapid decoding practice only (D; n=6)	Neale Analysis of Reading Ability (NARA), comprehension subtest	S>D: g=1.96*[a] S>QO: g=NR QO>D: g=NR

[a] This is the minimum effect that could have been achieved, given p<.001 (as cited in Yuill and Oakhill 1988, p. 40)

Table 2 Summary of single-case intervention study characteristics and findings

Study • Design • Participant age • Dosage (total hours per student)	Treatment(s) (n=sample size)	Comparison (n=sample size)	Dependent measures	Results/ES (g)
Fritschmann et al. (2007) • Single case (8 cases) • 9th graders • 15 h	Multi-step INFER strategy (S; n=1)	No instruction (CO; n=1)	Reading comprehension questions (1:4 ratio of literal to inferential questions) Reading comprehension questions, maintenance (8 months post-intervention) GRADE standardized reading comprehension test	S was highly effective in comparison to CO: PND=100 % S was highly effective in comparison to CO: PND=100 % S>CO (single group pretest-posttest gain): g=4.18*

were mostly elaborative in nature. In three studies (Holmes 1985; Yuill and Joscelyne 1988; Yuill and Oakhill 1988), students were provided instruction aimed at increasing necessary gap-filling inferences. In two studies (Winne et al. 1993; McMaster et al. 2012), students were prompted to make text-connecting inferences.

Mean effect sizes for group design studies ranged from g=0.72* to g=1.85* for researcher-developed measures of inferential reading comprehension. Every study that employed a researcher-developed measure of inferential reading comprehension achieved a statistically significant mean effect. Mean effect sizes ranged from g=−0.27 to g=0.36 for researcher-developed measures of story recall and from g=−0.03 to g=1.96* for standardized measures of reading comprehension. The percentage of non-overlapping data for the study that employed a single-case design was 100 % for all measures. It was not fruitful to average effect sizes across studies in order to assess mean effect sizes associated with particular study features because comparison groups and dependent measures varied so widely across studies; however, mean effect sizes for individual studies are cited below.

Participants

Participants in the majority of the studies included in this synthesis were fourth and fifth graders (Hansen and Pearson 1983 (g=1.20*); Holmes 1985 (g=1.85*); McCormick and Hill 1984 (g=0.72*); McMaster et al. 2012 (g=−0.27 on measure of liberal recall); Ouellette et al. 1999 (g=0.36 on measure of story recall); Winne et al. 1993 (g=0.82*)). However, participants in two studies (Yuill and Joscelyne 1988 (g=0.93*); Yuill and Oakhill 1988 (g=1.96*)) were second-graders, and in one study (Fritschmann et al. 2007 (PND=100 %)), participants were ninth graders.

Participants in all studies were identified as struggling readers, but methods of identifying struggling readers differed across studies. Identification was based on significantly lower scores than peers on standardized tests of reading ability, including (a) the *Stanford Achievement Test* comprehension subtest, on which "poor readers" had a mean grade-equivalent score of 3.2 and good readers had a mean score equivalent to 6.3 (Hansen and Pearson 1983, p. 822 (g=1.20*)); (b) the *Gates–MacGinitie Reading Test* comprehension subtest, on which struggling readers scored at or below the 23 percentile (McMaster et al. 2012 (g=−0.27 on measure of liberal recall]); (c) the *Metropolitan Achievement Test,* 6th edition, on which struggling readers scored at or below the 31st percentile (Ouellette et al. 1999 (g=0.36 on measure of story recall)); or (d) the *Neale Analysis of Reading Ability*, on which "less skilled comprehenders" had a significantly lower mean comprehension score than "skilled comprehender" peers (Yuill and Joscelyne 1988, p. 156 (g=0.93*)) or comprehension age scores below their chronological age scores and at least 6 months below their accuracy age scores (Yuill and Oakhill 1988, pp. 36–37 (g=1.96*)). It is worth noting that, in these last two studies, participants were identified as having a specific comprehension deficit: their decoding ability, as measured by their reading accuracy age score on the *Neale Analysis of Reading Ability*, was above or equal to their chronological age. In all other synthesized studies, decoding skill was not measured separately from comprehension skill; readers may have struggled with decoding, comprehension, or both.

School identification of LD for study participants that were identified as having LD was based on (a) discrepancies between IQ scores and achievement scores as well as low scores (five grade levels below grade placement) on a standardized reading achievement test (Fritschmann et al. 2007 (PND=100 %)) or (b) low scores alone (two or more years below grade placement and below the 33rd percentile, respectively) on a standardized reading achievement test (Holmes 1985 (g=1.85*); McCormick and Hill 1984 (g=0.72*)). One study

noted only that students were identified as having a learning disability according to "local criteria" (Winne et al. 1993, p. 55 (g=0.82*)).

Information was provided about the socioeconomic status of participants in only five of the nine synthesized studies. One of these five studies took place in a "low socioeconomic area of a large metropolitan city" (McCormick and Hill 1984, p. 220 (g=0.72*)), another in "a low-socioeconomic urban school" (Holmes 1985, p. 543 (g=1.85*)), and a third in a "small town in Maine that included diverse socioeconomic levels" (Hansen and Pearson 1983, p. 822 (g=1.20*)). A fourth study took place in school districts in which 27–29 % of students qualified for the free lunch program (McMaster et al. 2012 (g=−0.27 on measure of liberal recall)), whereas in the final study, more than half of participants qualified for free or reduced-cost lunch (Fritschmann et al. 2007 (PND=100 %)).

Treatment and Comparison Conditions

Treatment Three studies engaged students in prior knowledge activation and prediction work as part of an inference training intervention (Hansen and Pearson 1983 (g=1.20*); McCormick and Hill 1984 (g=1.85*); Ouellette et al. 1999 (g=0.36 on measure of story recall)). In these studies, students were asked a question about their previous experiences with an important idea in a story prior to reading the story. Then, they were encouraged to hypothesize about what might happen under similar circumstances in a story they were about to read. For example, prior to reading a story, a teacher might ask students to "tell us about a time when you were embarrassed about the way you looked" (Hansen and Pearson 1983, p. 823). Then, after listening to students' responses, the teacher would let students know that, "in our next story there is an old man who is embarrassed about the way that he looks," and ask them, "What do you think is the thing that embarrasses him?" (Hansen and Pearson 1983, p. 823). The purpose of these questions was to activate the process of text-to-self connection, in order to "set the stage for interaction with the text during reading" (Hansen and Pearson 1983, p. 823) and thus facilitate the generation of gap-filling inferences—as well as to encourage students to share ideas so that each would have a "larger bank of knowledge to bring to bear upon the story than if they had to use only their own prior knowledge" (Hansen and Pearson 1983, p. 824).

Four interventions focused primarily on encouraging students to identify key or "clue" words, in order to integrate important elements in the text or fill textual gaps by integrating key words with prior knowledge. In the study reported by Holmes (1985) (g=1.85*), students were taught to find and use relevant "key words" in the text, as well as to use self-questioning to confirm tentative answers (p. 544). In the study of Winne et al. (1993) (g=0.82*), students learned to identify a "rule" and a "critical fact" presented in a passage; a teacher explained explicitly, and then modeled, how to apply the rule and make use of the critical fact to answer an inferential question.

A third study (Yuill and Oakhill 1988 (g=1.96*)) used a three-part inference instruction intervention that included, first, "lexical inference training," in which children were taught to identify clue words in passages and then link clue words in order to generate inferences (p. 37). For example, students read the following passage:

> Billy was crying. His whole day was spoilt. All his work had been broken by the wave. His mother came to stop him crying. But she accidentally stepped on the only tower that was left. Billy cried even more.

Students were encouraged to identify "wave" as a clue word indicating that the story setting was a beach. They were also helped to link "wave" with "tower" in order to infer that the tower was a part of a sandcastle (Yuill and Oakhill 1988, p. 38). Question-generation training

constituted the second part of the intervention. During question-generation training, students took turns acting the part of teacher and composing "who," "where," and "why" inferential questions about a recently read passage in order to quiz peers. Finally, students received "prediction" training. In contrast to the predictions that integrated prior knowledge with text in the studies described earlier, however, prediction training comprised "macrocloze tasks" that required students to read a passage in which sentences were obscured by tape and infer the meaning of the hidden sentence based on clues in surrounding sentences (Yuill and Oakhill 1988, p. 43).

The final study that focused on key words (Yuill and Joscelyne 1988 (g=0.93*)) investigated the impact of just the "lexical inference training" component of the intervention described above. In this study, students were first given single sentences and then whole paragraphs, and taught to identify and integrate key words in order to make inferences about settings and other story elements.

Fritschmann and colleagues (2007; PND=100 %) aimed to teach students to activate prior knowledge as well as to identify and integrate key words or phrases in text. Students learned a five-step INFER strategy: (a) "*I*nteract with the passage and the questions" by previewing the passage, reading the questions, and distinguishing between factual questions and "think-and-seek" questions, which include purpose, main idea/summarization, prediction, and clarification questions; (b) "*N*ote what you know" by activating background knowledge or experiences related to the topic at hand, as well as underlining any key words in the questions; (c) "*F*ind the clues" by carefully reading the passage and underlining clues directly related to keywords in the questions; (d) "*E*xplore more details" by looking for additional clues in the passage that support tentative answers; (e) "*R*eturn to the question" and make sure that an answer has been selected and marked (p. 248).

The last intervention, designed by McMaster et al. (2012) (g=−0.27 on measure of liberal recall), evaluated the impact of two different questioning interventions aimed at improving inference generation skill. Students in a "causal" questioning condition were asked questions that encouraged them to connect an important consequence with a casual antecedent when they reached a point in a passage where a causal inference was necessary (McMaster et al. 2012, p. 104). Students in a "general" questioning condition were asked the question, "How does the sentence you just read connect with something that happened before in the story?" every five to six sentences (once students were familiar with this question, they were simply prompted to "Connect it!") (McMaster et al. 2012, p. 105). There were no significant differences between groups of struggling readers in the three conditions, and effect sizes for all comparisons were quite small. However, the authors found that when they disaggregated scores for two subgroups of struggling comprehenders labeled "elaborators" and "paraphrasers," elaborators benefited more than paraphrasers from causal questioning (g=0.83*), whereas paraphrasers benefited more than elaborators from general questioning (g=1.37*).

The total number of sessions for synthesized interventions ranged from 1 to 40. The mean number of sessions was 15.7. For the five studies that reported session length, total mean intervention dosage ranged from 2.67 to 26.67 h. The mean dosage for these five studies was 12.6 h. Visual inspection of dosage alongside of effect sizes indicated that increased dosage was not associated with larger effect sizes, nor was it associated with smaller effect sizes.

Comparison We labeled as true "comparison" (CO) conditions those conditions that engaged students in business-as-usual instruction: students read passages independently with little or no introduction to text and answered post-reading comprehension questions. When comprehension questions were asked post-reading, then more than half of questions were literal comprehension questions (i.e., less than half required students to make an inference). In one study,

students in the comparison condition were asked literal W-questions (who, what, where, and when) during reading; the answers to these questions were stated explicitly in the text, in the sentence prior to the question (McMaster et al. 2012, p. 105). In this study, all during-reading questions were judged to be literal questions and the condition was judged to be a true comparison condition.

In two studies, there were no true comparison conditions according to the above definitions. In the study reported by Winne et al. (1993), the two conditions were identical except that feedback provided to students after they answered the post-reading inferential question was "inductive" in one and "explicit" in the other. In the "inductive" condition, correct answers were provided and the necessary ingredients for making an inference were underlined—but there was no explicit instruction explaining explicitly how to integrate ingredients and generate the inference. This was judged to be an alternate treatment, rather than a true comparison condition.

In the study reported by Holmes (1985), there was also no true comparison condition according to the above definition. In the condition that was intended to be the comparison condition, students read passages independently with little or no introduction to text and practiced answering post-reading comprehension questions, all of which required students to make an inference. This condition resembled alternative treatment conditions in two studies that employed multiple-treatment designs (McCormick and Hill 1984; Yuill and Oakhill 1988). Interestingly, in the study reported by McCormick and Hill (1984), students in the second treatment condition, who received no instruction but did receive practice answering inferential questions after reading, scored statistically significantly higher than students in the control condition on two researcher-developed post-reading questions measures (g=1.49* for each). These students even statistically significantly outperformed students in the inference instruction treatment condition on the researcher-developed posttest measure (g=0.83*).

Yuill and Oakhill (1988) designed a comparison condition that comprised rapid word recognition instruction. At the beginning of each session, students were told of the importance of rapid word recognition, and then shown a list of words taken from the passage they were about to read. After the experimenter read the words aloud, students practiced reading the list as quickly and as accurately as possible. Then, students took turns reading the day's passage. Finally, each student read the word list again, and the experimenter recorded the time taken on a stopwatch.

Materials and Implementation More than half of synthesized studies instructed students in generating inferences while reading narrative text. Four studies used a mix of narrative and expository texts (Hansen and Pearson 1983; Holmes 1985; McCormick and Hill 1984; Winne et al. 1993). There was not one text type or mixture of types that was associated with increased effects. In one third of the studies (Fritschmann et al. 2007 (PND=100 %); Holmes 1985 (g= 1.85*); Yuill and Oakhill 1988 (g=1.96*)), instruction was provided to small groups of between three and five students. In two studies, instruction was provided to students individually (Winne et al. 1993 (g=.82*); Yuill and Joscelyne 1988 (g=.93*)). Ouellette et al. (1999) (g=0.36) provided instruction to groups of six students and Hansen and Pearson (1983) (g= 1.20*) provided instruction to groups of 10 students. Finally, McMaster et al. (2012)(g=−0.27 for measure of liberal recall) provided instruction to students in groups of between 23 and 28 students.

For more than half of studies, researchers delivered the intervention. In four studies, however, reading teachers (or students enrolled in teacher preparation programs) were the implementers of the intervention (Hansen and Pearson (g=1.20*); McCormick and Hill (g=0.72*); McMaster et al. 2012 (g=.027 for measure of liberal recall); Winne et al. 1993 (g=0.82*)).

Two studies (Holmes 1985; Ouellette et al. 1999) did not report measuring fidelity of implementation. Three studies reported that teachers were audiotaped and or/observed regularly, but did not report specific fidelity measurements (Hansen and Pearson 1983; Winne et al. 1993; Yuill and Joscelyne 1988; Yuill and Oakhill 1988). For the remaining three studies, the percentage of faithfully executed parts of the intervention was between 90 and 98 % for all lessons that were observed (Fritschmann et al. 2007; McCormick and Hill 1984; McMaster et al. 2012).

Outcome Measures

While one study (Yuill and Oakhill 1988) employed only a global, standardized measure of reading comprehension skill, most studies used (either in addition to or in place of a standardized measure) researcher-designed measures of inferential comprehension. Some researchers measured both literal and inferential comprehension, with scores for each type of question disaggregated and effects on each considered separately. It is not uninteresting to consider the effects of inference instruction on literal comprehension. Hansen and Pearson (1983) did find that inference instruction had a significant impact on students' literal comprehension of text, and McMaster et al. (2012) determined that in the general questioning condition one subgroup of struggling readers benefitted more than elaborators on a measure of literal recall. Nevertheless, we found that addressing the impact of inference instruction on literal comprehension skill was beyond the scope of this study. This synthesis focuses primarily on the effects of inference instruction on inferential reading comprehension and global reading comprehension.

More than half of studies employed only researcher-developed measures of inferential reading comprehension. Most researcher-developed measures resembled each other: They consisted of post-reading inferential comprehension questions and were highly aligned with instruction. In only one case was a researcher-developed measure slightly less aligned with instruction: In the intervention reported by Hansen and Pearson (1983), students were taught to make gap-filling, mostly elaborative inferences by interweaving their background knowledge with information in the text in order to infer about character traits, feelings, motivation, and goals. However, at least one question used in the researcher-developed measure of inference skill designed by Hansen and Pearson (1983) measured students' ability to make a text-connecting inference: Students were asked to use context clues to determine the meaning of a phrase in the text.

Three studies transcribed and parsed student recalls of stories as a means of assessing inferential comprehension. McMaster et al. (2012) categorized each recall clause based on how closely it matched the gist of the original text (Kendeou and van den Broek 2005). McMaster et al. (2012) categorized clauses as (a) "conservative," which were literal renderings of original text units that reflected "near-verbatim memory for the text"; (b) "liberal," which were non-literal renderings of the original text that captured "the essence of its meaning" such that they reflected "the extent to which the reader has established a coherent representation of text"; (c) "highly connected," which had five or more causal connections based on the causal network for that text, such that they reflected "the reader's strong sensitivity to the text structure"; (d) "no match consistent," which could "not be matched directly with the gist of a text unit" but "was valid and moderately constrained by the text," such that they reflected "the extent to which the reader has elaborated his/her mental representation through inferences"; and (e) "no match inconsistent," which did not match the gist of a text unit and was invalid or unconstrained by the text (p. 104). Ouellette et al. (1999) measured accuracy of story summary using a checklist patterned after Glazer's (1988) retell analysis guide, but I could not obtain a copy of this checklist or a description of any of the items on it. Yuill and Joscelyne (1988) specify only that "recall scores were based on correct recall for the gist of each idea unit, scored separately for main and subsidiary ideas" (p. 154).

Three of the nine studies employed standardized measures of reading comprehension in addition to researcher-developed measures of inferential comprehension (Holmes 1985; McCormick and Hill 1984; Ouellette et al. 1999). One study (Yuill and Oakhill 1988) used a standardized measure of reading comprehension as its sole outcome measure. In the study reported by Holmes (1985), the standardized measure was the Nelson Reading Skills Test (Hanna et al. 1977), which claims to assess comprehension at three levels: Students must answer "higher level" inference questions, "translational" or lower-level inference questions, and literal questions. McCormick and Hill (1984) used the fifth edition of the Reading Comprehension subtest of the Reading Diagnostic Tests of the Metropolitan Achievement Tests (MAT5; Prescott et al. 1978), and Ouellette et al. (1999) used the sixth edition (MAT6; Prescott et al. 1984). Both the MAT5 and the MAT6 require students to read simple sentences and select the picture that best corresponds with these sentences, as well as to read passages and answer literal and inferential comprehension questions. Yuill and Oakhill (1988) measured improvement in reading comprehension by means of the Reading Comprehension subtest of the Neale Analysis of Reading Ability—Revised British Edition (NARA; Neale 1989). The NARA has students read passages and asks them to answer both literal and inferential questions after reading. All four of these standardized measures assess reading comprehension globally, including students' ability to answer literal and inferential questions. None of the assessments disaggregated students' scores on inferential questions.

For tests that were not scored objectively, information about inter-rater reliability was provided in all but two studies (Holmes 1985; Winne et al. 1993). For all other studies that employed subjective scoring procedures, inter-rater reliability was at or above 89 %.

Study Quality

Group design studies were judged according to What Works Clearinghouse standards (IES 2013, pp. 9–10) and the essential quality indicators for group experimental and quasi-experimental design studies described by Gersten et al. (2005). Using these standards as guidelines, synthesized studies were given an impressionistic quality rating of "high," "medium," or "low" during the coding process. All group design studies employed random assignment, thus meeting the primary quality standard laid out by the What Works Clearinghouse (IES 2013). One study (McCormick and Hill 1984) met all of the additional essential quality indicators described by Gersten et al. (2005), and it was given a high impressionistic quality rating. The study reported by Winne et al. (1993) was judged as being of low quality because it failed to provide information about the process by which participants were classified as having LD and about inter-rater reliability for an objectively scored outcome measure. It also did not report any specific measurements related to fidelity of treatment implementation. All other group design studies met most but not all of the essential quality indicators described by Gersten et al. (2005); they were judged to be of medium quality. The one single case study that was included in this synthesis met almost all quality indicators for single subject research described by Horner et al. (2005). Study quality was not associated with increased effects.

Discussion

The purpose of this study was to provide a synthesis of the research on inference instruction interventions conducted among students who are struggling readers. We aimed to find and compare all intervention studies that evaluated the effects of inference instruction on inferential reading comprehension. A comprehensive search yielded nine studies, one of which employed a single-case design, and eight of which employed randomized group designs.

The fact that only nine intervention studies met the stated selection criteria is perhaps the most important finding of this synthesis. Given the importance of inference generation to reading comprehension (Cain et al. 2004; Kendeou et al. 2008; Oakhill and Cain 2012; Oakhill et al. 2003) and the evidence that struggling readers may receive particular benefit from inference instruction (Hansen and Pearson 1983; McGee and Johnson 2003; Raphael and Pearson 1985; Yuill and Oakhill 1988), it is surprising that there are so few studies that investigate the impact of inference instruction on the inferential comprehension skill of struggling readers. Also worthy of note is the fact that, of the eight group design studies that emerged from the literature in meeting selection criteria, only one could be judged to be of the highest quality. Most studies included in this synthesis lacked carefully described, reliable, and valid measures, and very few employed multiple measures in order to provide a balance between measures closely aligned with the intervention and measures of more generalized performance.

In general, the findings of this synthesis suggest that inference instruction interventions can be effective both when they target prior knowledge activation (and teach students to integrate prior knowledge with information in text) and when they focus only on integrating information in text. The majority of effective interventions taught students to identify clues or key words in the text and to use these key words to furnish answers to post-reading inferential questions. Another approach that proved effective was to activate students' prior knowledge and teach them how to interweave this knowledge with information in the text. Finally, two studies (Holmes 1985; Yuill and Oakhill 1988) taught students to generate questions as a way of identifying gaps in text or confirming tentative inferences.

Interventions were effective across grades 2–9, for readers with the label LD, as well as struggling readers not identified with a specific learning disability, regardless of student socioeconomic status. The effectiveness of included studies did not depend on type of text used during instruction (narrative vs. a mixture of narrative and expository), or on the implementer of the intervention (a researcher vs. a classroom teacher). Interventions were effective when they were conducted one-on-one, in small groups of three to five students and in larger groups of ten students. The one study that was conducted in larger classes of between 23 and 28 students (McMaster et al. 2012) did not produce statistically significant effects in favor of either intervention condition.

There were two other findings worthy of note. While McMaster et al. (2012) did not find statistically significant effects in favor of either intervention condition, they did report differential effects for intervention between two subgroups of struggling comprehenders, which they designated "elaborators" and "paraphrasers" (101). Elaborators benefitted more than paraphrasers from causal questioning on highly connected items recalled (g=0.83*), whereas paraphrasers benefitted more than elaborators from general questioning on unique conservative items recalled (g=1.37*). These results suggest that it may be important to identify subgroups of struggling readers and target interventions at a particular subgroups, rather than to target interventions at struggling readers in the aggregate.

Finally, based on the results reported by McCormick and Hill (1984), simply giving students opportunities to answer a large number of inferential questions after reading provides significant benefit in comparison to assigning them mostly literal comprehension questions after reading. It may even be as effective as strategy instruction combined with fewer opportunities to answer inferential questions. As it is presumably less time-consuming to provide access to questions than it is to provide explicit strategy instruction (at least in the case of some of the multistep strategies described above), it would be worthwhile to confirm the results of this study. In addition, it would be useful to determine the extent to which effects are contingent on following a certain protocol when educators ask post-reading inferential questions and provide feedback about answers.

Limitations

One potential limitation of this synthesis is that inclusion criteria were very specific and might be considered overly restrictive. The purpose for these criteria was to assure reliability in selecting studies and ensure study replicability. For example, the fourth criterion required that "a study needed to state, as its primary purpose, an aim to improve inference-making skill." Had the criterion required only that the study have some (primary or collateral) benefit related to inference making skill, it would be far more difficult to ensure replicability. Nevertheless, by excluding studies that did not state an explicit aim to improve inference-making skill, a wide variety of interventions that had a collateral impact on inference making or general comprehension were eliminated.

There is another potential explanation for the dearth of studies that were found to meet selection criteria for this synthesis. Because of the now well-documented "file drawer" problem (Cooper 1998; Torgerson 2006), it is difficult to find published studies that report negative or null effects. It is possible that this bias prevented the publication of any number of studies that did not find statistically significant differences in favor of an inference instruction treatment.

Another significant limitation is that synthesis findings were limited by the wide variation in comparison groups and dependent measures. As a result, it was not productive to combine and average effect sizes for studies according to particular shared study features, and thus, it was not possible to objectively identify features of instructional interventions (e.g., duration, grade level, type of instruction) that were associated with higher or lower mean effect sizes.

The quality of any synthesis is directly dependent on the quality of included studies. As detailed above, the majority of group design studies synthesized here could not be judged to be of the highest quality. Most studies lacked carefully described, reliable, and valid measures. Very few employed multiple measures in order to provide a balance between measures closely aligned with the intervention and measures of more generalized performance. In order to facilitate comparison across studies, future research would benefit from the development and use of carefully described, validated measures of inferential comprehension that closely resemble those used in other inference instruction studies. Similarly, it will be important for future researchers to compose and carefully describe comparison conditions that closely resemble each other and mirror typical practice in English language arts or content area classrooms.

A large number of studies had to be excluded from this synthesis because they did not meet the second inclusion criterion employed during the search process. Many of these excluded studies reported important results related to the differences between struggling and non-struggling readers who received inference instruction (Carr and Thompson 1996; Carr et al. 1983; Dewitz et al. 1987; McGee and Johnson 2003; Raphael and Pearson 1985; Raphael and Wonnacott 1985). However, they did not provide sufficient information for the purpose of calculating the significance of differences between groups or conditions for struggling readers. It will be valuable for future studies to collect data that enable readers to compare the impact of different inference instruction treatments on struggling readers, in addition to comparing differential effects of interventions on struggling readers and non-struggling readers.

Finally, it was difficult to arrive at generalizations and make recommendations based on the results of only nine studies. There appears to be a need for more intervention research investigating the effects of inference instruction on struggling readers. Alternatively, there may be a need for more publication of studies that find negative or null effects for this type of instruction.

Implications for Practice

This synthesis yields several implications for educators. First, studies indicated that helping struggling readers identify key words in text and use these key words to furnish answers to post-reading inferential questions is associated with improved comprehension outcomes. Studies also suggest that it may be useful for teachers to devote time before reading to building and/or activating students' knowledge related to topics covered in text. In combination with knowledge building and activation, it may be necessary to provide instruction and practice with integrating prior knowledge with information in text. Finally, simply giving students opportunities to answer a large number of inferential questions after reading (instead of asking mostly literal post-reading comprehension questions) may be as effective in improving their inferential comprehension of text as strategy instruction combined with fewer opportunities to answer inferential questions.

In addition, because there is some evidence that there are subgroups of poor comprehenders (such as the "elaborators" and "paraphrasers" described by McMaster et al. 2012, p. 101), educators may benefit from knowing that some of their struggling readers may make greater improvement when they participate in one type inference generation instruction and practice, while others may make greater gains as a result of a different type of instruction.

Future Research

This synthesis may raise more questions than it provides answers. In addition to investigating the extent to which providing opportunities to answer inferential questions is as effective as strategy instruction for the purposes of improving inferencing skill, it would be profitable for researchers to consider the following:

1. Which steps of multi-step interventions (like the one described by Fritschmann et al. 2007) are the ones most closely associated with positive outcomes?
2. To what extent does the effectiveness of given interventions differ based on the reading level or learning characteristics of the participant (as is suggested by McMaster et al. 2012)?
3. Does type of text (e.g., expository or narrative) moderate the effect of interventions? For example, is one type of instruction effective in the context of narrative text, but not in the context of expository text?
4. Are certain interventions more appropriate for teaching students to make particular kinds of inferences (e.g. text-connecting vs. gap-filling, or necessary vs. elaborative)? Does instruction in making particular types of inferences produce specific effects only for similar untaught inferences, or are the effects of inference training more general?
5. Does text complexity matter? Research suggests that overly complex texts may impede inference processing (Singer et al. 1997), while texts that are overly elaborated or explicit promote passive reading and poor comprehension (Gilabert et al. 2005; McNamara et al. 1996). Do teachers need to employ different kinds of instruction when they are teaching students to read texts of high (or low) levels of complexity? How does the text complexity variable interact with student background knowledge and comprehension skill level (McNamara et al. 1996), in the context of inference instruction?
6. Are the effects of inference training sustainable? Only one study among those synthesized here included a follow-up measure of inferential comprehension (Fritschmann et al. 2007). In this study, performances on reading comprehension probes conducted 8 months after the intervention were relatively low. Intervention is far more valuable if it offers long-

lasting benefits; future research should address the question as to if and how interventions can achieve sustainable improvement in inferential comprehension.

7. Finally, to what extent are benefits obtained as a result of inference instruction translatable to overall reading achievement? Is inference instruction more (or less) powerful in the context of multi-component interventions than it is when delivered alone?

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

References

Berkeley, S., Scruggs, T. E., & Mastropieri, M. A. (2010). Reading comprehension instruction for students with learning disabilities, 1995–2006: a meta-analysis. *Remedial and Special Education, 31*, 423–436.

Bowyer-Crane, C., & Snowling, M. J. (2005). Assessing children's inference generation: What do tests of reading comprehension measure? *British Journal of Educational Psychology, 75*, 189–201.

Cain, K. (2010). *Reading development and difficulties*. Chichester: BPS Blackwell.

Cain, K., & Oakhill, J. V. (1999). Inference making and its relation to comprehension failure. *Reading and Writing, 11*, 489–503.

Cain, K., Oakhill, J. V., & Bryant, P. E. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skill. *Journal of Educational Psychology, 96*, 671–81.

Carr, E. M., Dewitz, P., & Patberg, J. P. (1983). The effect of inference training on children's comprehension of expository text. *Journal of Reading Behavior, 15*(3), 1–18.

Carr, S. C., & Thompson, B. (1996). The effects of prior knowledge and schema activation strategies on the inferential reading comprehension of children with and without learning disabilities. *Learning Disability Quarterly, 19*(1), 48–61.

Common Core State Standards Initiative. (2010). *Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects*. Retrieved from http://www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf.

Cooper, H. (1998). *Synthesizing research* (3rd ed.). Thousand Oaks: Sage.

Dewitz, P., Carr, E. M., & Patberg, J. P. (1987). Effects of inference training on comprehension and comprehension monitoring. *Reading Research Quarterly, 22*(1), 99–121.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C. K., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research, 79*(1), 262–300. doi:10.3102/0034654308325998.

Ezell, H. K., Kohler, F. W., Jarzynka, M., & Strain, P. S. (1992). Use of peer-assisted procedures to teach QAR reading comprehension strategies to third grade children. *Education and Treatment of Children, 15*, 205–227.

Flynn, L. J., Zheng, X., & Swanson, H. L. (2012). Instructing struggling older readers: A selective meta-analysis of intervention research. *Learning Disabilities Research and Practice, 27*(1), 21–32. doi:10.1111/j.1540-5826.2011.00347.

Fritschmann, N. S., Deshler, D. D., & Schumaker, J. B. (2007). The effects of instruction in an inference strategy on the reading comprehension skills of adolescents with disabilities. *Learning Disability Quarterly, 30*, 245–262.

Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. S. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children, 71*(2), 149–164.

Gersten, R., Fuchs, L. S., Williams, J. P., & Baker, S. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research, 71*(2), 279–320.

Gilabert, R., Martinez, G., & Vidal-Abarca, E. (2005). Some good texts are always better: Text revision to foster inferences of readers with high and low prior background knowledge. *Learning and Instruction, 15*(1), 45–68.

Glazer, S.M., Searfoss, L.W., & Gentile, L.M. (1988). Re-examining reading diagnosis: New trends and procedures in classrooms and clinics. Newark, DE: International Reading Association.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395.

Gurney, D., Gersten, R., Dimino, J., & Carnine, D. (1990). Story grammar: Effective literature instruction for high school students with learning disabilities. *Journal of Learning Disabilities, 23*(6), 335–348.

Hanna, G., Schell, L.M., & Schreiner, R. (1977) The Nelson reading skills test. Los Angeles: Houghton Mifflin.

Hansen, J., & Pearson, P. D. (1983). An instructional study: Improving the inferential comprehension of good and poor fourth-grade readers. *Journal of Educational Psychology, 75*(6), 821–829.

Holmes, B. C. (1985). The effects of a strategy and sequenced materials on the inferential comprehension of disabled readers. *Journal of Learning Disabilities, 18*(9), 542–546.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*(2), 165–179.

Institute of Education Sciences. (2008). *What Works Clearinghouse study review standards*. Retrieved from http://ies.ed.gov/ncee/ wwc/pdf/reference_resources/wwc_version1_standards.pdf.

Institute of Education Sciences (2013). What Works Clearinghouse procedures and standards handbook (Version 3.0). Retrieved February 14, 2014, from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19.

Johnson, D. D., & Pearson, P. D. (1978). *Teaching reading vocabulary*. New York: Holt, Rinehart & Winston.

Kendeou, P., & van den Broek, P. (2005). The effects of readers' misconceptions on com- prehension of scientific text. *Journal of Educational Psychology, 97*(2), 235–245.

Kendeou, P., Bohn-Gettler, C., White, M., & van den Broek, P. (2008). Children's inference generation across different media. *Journal of Research in Reading, 31*, 259–72.

Kintsch, W. (1993). Information accretion and reduction in text processing: Inferences. *Discourse Processes, 16*, 193–202.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. New York: Cambridge University Press.

Kispal, A. (2008). *Effective teaching of inference skills for reading: Literature review* (DCSF Research Report 031). London: DCSF

Mason, L. H. (2004). Explicit self-regulated strategy development versus reciprocal questioning: Effects on expository reading comprehension among struggling readers. *Journal of Educational Psychology, 96*(2), 283–296.

Mastropieri, M. A., Scruggs, T. E., Hamilton, S. L., Wolfe, S., Whedon, C., & Canevaro, A. (1996). Promoting thinking skills of students with learning disabilities: Effects on recall and comprehension of expository prose. *Exceptionality, 6*(1), 1–11.

McCormick, S., & Hill, D. S. (1984). An analysis of the effects of two procedures for increasing disabled readers' inferencing skills. *Journal of Educational Research, 77*(4), 219–227.

McGee, A., & Johnson, H. (2003). The effect of inference training on skilled and less skilled comprehenders. *Educational Psychology, 23*(1), 49–59.

McMaster, K. L., van den Broek, P., Espin, C. A., White, M. J., Rapp, D. N., Kendeou, P., Bohn-Gettler, C. M., & Carlson, S. (2012). Making the right connections: Differential effects of reading intervention for subgroups of comprehenders. *Learning and Individual Differences, 22*, 100–111.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43.

Neale, M. D. (1989). The Neale Analysis of Reading Ability—Revised British Edition. Windsor, United Kingdom: NFER-Nelson.

Oakhill, J. V. (1982). Constructive processes in skilled and less-skilled comprehenders' memory for sentences. *British Journal of Educational Psychology, 73*, 13–20.

Oakhill, J. V. (1984). Inferential and memory skills in children's comprehension of stories. *British Journal of Educational Psychology, 54*, 31–39.

Oakhill, J., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading, 16*(2), 91–121. doi:10.1080/10888438.2010.529219.

Oakhill, J. V., Cain, K., & Bryant, P. E. (2003). The dissociation of word reading and text comprehension: Evidence from component skills. *Language and Cognitive Processes, 18*, 443–468.

Ouellette, G., D'Agostino, L., & Carifio, J. (1999). The effects of exposure to children's literature through read-aloud and on inferencing strategy on low reading ability fifth graders' sense of story structure and reading comprehension. *Reading Improvement, 36*(2), 73–89.

Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. (1978). *Metropolitan achievement tests*. New York: The Psychological Corporation.

Prescott, G. A., Balow, I. H., Hogan, T. P., & Farr, R. C. (1984). Metropolitan achievement tests, (6th ed.). New York: The Psychological Corporation.

Raphael, T. E., & Pearson, P. D. (1985). Increasing students' awareness of sources of information for answering questions. *American Educational Research Journal, 22*(2), 217–235.

Raphael, T. E., & Wonnacott, C. A. (1985). Heightening fourth-grade students' sensitivity to sources of information for answering comprehension questions. *Reading Research Quarterly, 20*(3), 282–296.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*(4), 289–312.

Scammacca, N., Roberts, G., Vaughn, S., & Stuebing, K. K. (2013). A meta-analysis of interventions for struggling readers in grades 4–12: 1980–2011. *Journal of Learning Disabilities, advanced online copy.* doi: 10.1177/0022219413504995.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research methodology: Methodology and validation. *Remedial and Special Education, 8*, 24–33.

Singer, M., Harkness, D., & Stewart, S. T. (1997). Constructing inferences in expository text comprehension. *Discourse Processes, 24*(2–3), 199–228.

Solis, M., Ciullo, S., Vaughn, S., Pyle, N., Hassaram, B., & Leroux, A. (2011). Reading comprehension interventions for middle school students with learning disabilities: A synthesis of 30 years of research. *Journal of learning disabilities*, 0022219411402691.

Torgerson, C. J. (2006). Publication bias: the Achilles' heel of systematic reviews? *British Journal of Educational Studies, 54*(1), 89–102.

Wanzek, J., & Vaughn, S. (2007). Research-based implications from extensive early reading interventions. *School Psychology Review, 36*(4), 541–561.

Wanzek, J., Vaughn, S., Wexler, J., Swanson, E., Edmonds, M. E., & Kim, A.-H. (2006). A synthesis of spelling and reading interventions and their effects on the spelling outcomes of students with LD. *Journal of Learning Disabilities, 39*, 528–543. doi:10.1177/00222194060390060501.

Warren, W. H., Nicholas, D. W., & Trabasso, T. (1979). Event chains and inferences in understanding narratives. In R. O. Freedle (Ed.), *New directions in discourse processing* (Vol. 2). Norwood: Ablex Publishing Corporation.

Wendt, O. (2009, May). *Calculating effect sizes for single-subject experimental designs.* Paper presented at the Ninth Annual International Campbell Collaboration Colloquium, Oslo, Norway

Winne, P. H., Graham, L., & Prock, L. (1993). A model of poor readers' text-based inferencing: Effects of explanatory feedback. *Reading Research Quarterly, 28*(1), 53–66.

Yuill, N., & Joscelyne, P. (1988). Effects of organizational cues and strategies on good and poor comprehenders' story understanding. *Journal of Educational Psychology, 80*, 59–67.

Yuill, N., & Oakhill, J. (1988). Effects of inference awareness training on poor reading comprehension. *Applied Cognitive Psychology, 2*, 33–45.