Instr Sci (2018) 46:921–946
https://doi.org/10.1007/s11251-018-9456-2

ORIGINAL RESEARCH

Practice makes proficient: teaching undergraduate students to understand published research

Trina C. Kershaw[1] · Jordan P. Lippman[2] · Jennifer M. B. Fugate[1]

Received: 16 November 2016 / Accepted: 10 May 2018 / Published online: 16 May 2018

Abstract Scientific knowledge, including the critical evaluation and comprehension of empirical articles, is a key skill valued by most undergraduate institutions for students within the sciences. Students often find it difficult to not only summarize empirical journal articles, but moreover to successfully grasp the quality and rigor of investigation behind the source. In this paper, we use instructional scaffolds (reading worksheets, RWs, with tutorials) to aid students in being able to comprehend, and ultimately transfer, the skills necessary in critically evaluating primary sources of research. We assess students' learning of these skills on a multiple-choice assessment of Journal Article Comprehension (JAC). Students in experimental classes, who received instructional scaffolds, improved on the JAC post-test compared with students in control classes. This result shows that students are acquiring fundamental research skills such as understanding the components of research articles. We also showed that improvement on the JAC post-test for the experimental class extended to a written summary test. This result suggests that students in the experimental group are developing discipline-specific science process skills that allow them to apply JAC skills to a near-transfer task of writing a summary.

Keywords Reading empirical articles · Instructional scaffolds · Assessment · Learning outcomes · Cognitive psychology

It is important for undergraduate students to learn to comprehend and analyze published research (c.f. Coil et al. 2010) because it exposes them to the means by which scientific knowledge is justified and evaluated. As reported by Oldenburg (2016), over 70% of psychology courses at liberal arts colleges include primary source readings. Although

✉ Trina C. Kershaw
tkershaw@umassd.edu

[1] Department of Psychology, University of Massachusetts Dartmouth, North Dartmouth, MA 02747, USA

[2] The TeamBuilders Group, Pittsburgh, PA, USA

undergraduate students are often asked to summarize research articles (e.g., Anisfield 1987; Gillen 2006; Karcher 2000; Levine 2001; Suter and Frank 1986), they find the task difficult (e.g., Taylor 1983) and often make mistakes (e.g., Day 1983) because they are interacting with materials that are written for expert scientists (c.f. Goldman and Bisanz 2002). In addition, while comprehension of research articles is important, faculty often do not have time to teach science process skills which are necessary to comprehend research articles (Coil et al. 2010).

A number of studies have shown that, due to lack of experience, students have difficulty with understanding critical components of empirical research, including understanding the argument structure of an Introduction section (Newell et al. 2011; Van Lacum et al. 2014), identifying independent and dependent variables (Dasgupta et al. 2014), and understanding statistical concepts such as correlation (Zieffler and Garfield 2009), probability, and statistical significance (Dasgupta et al. 2014). Because students lack experience with these science process skills, they are slow to understand what information in a research article is relevant (c.f. Morris et al. 2012) and thus have difficulty developing coherent mental representations of research (c.f. Chinn and Brewer 2001).

Despite students' difficulty in effectively summarizing and using research articles, many researchers recommend using journal articles in instruction (Anisfield 1987; Gillen 2006; Karcher 2000; Levine 2001). To make the process of learning with research articles more tractable, two main types of approaches have been proposed: (a) using texts that are adapted to make them appropriate for students and (b) creating instructional scaffolds that provide guidance for students on how to approach reading the original articles.

The first approach, called adapted primary literature (APL; cf. Yarden et al. 2001; Yarden 2009), involves simplifying articles to remove technical details so they can be read and understood by students while still maintaining the article's argument structure. Although APL has shown to increase critical reading of science beyond that of secondary literature such as textbooks (Baram-Tsabari and Yarden 2005), it has primarily been used in biology and the life sciences rather than the social sciences.

The second approach is to design instructional scaffolds (cf. Reiser 2004). For example, scaffolds for reading can involve problematizing an aspect of the task by adding additional post-reading work (such as questions to answer), restructuring the activity by using signaling devices such as prominent headings (Lorch and Lorch 1996), or providing structural information about the text such as graphical organizers or outlines (Newell et al. 2011). Scaffolds may also include prompts to direct students' attention to key features of a text (Hmelo-Silver et al. 2007) or directly teach students about the structure of a particular section of a text, such as the rhetorical moves used in the argument structure of the Introduction (Van Lacum et al. 2014).

While scaffolds can take multiple forms, they are often implemented as structured modifications of instruction (Hmelo-Silver et al. 2007) and thus satisfy best practices for instructional design. For example, Goldman and Pellegrino (2015) noted that "… contexts for learning should pose challenging tasks and provide guidance and supports that make the task manageable for learners" (p. 36). The literature concerning the principles that underlie cognitive skill acquisition (e.g., Dunlosky et al. 2013) also suggests several best practice principles that can enhance students' learning. Deliberate practice and distributed practice are particularly relevant principles of skill acquisition that are appropriate for the application of scaffolds to aid learning complex skills such as understanding journal articles.

Deliberate practice is operationalized as purposeful and effortful activity to improve a specific aspect of performance (c.f. Ericsson and Charness 1994). This type of sustained, effortful practice is a positive predictor of educational outcomes (Macnamara et al. 2014), such as improvement on standardized assessments of critical thinking (van Gelder et al. 2004). In the area of science, it is particularly important for students to engage in deliberate practice, because reasoning about science develops slowly and requires instruction, support, and practice (Morris et al. 2012).

In a similar way, it is also important that students engage in distributed practice by spacing out their learning sessions, or assignments, over time. The use of distributed practice has been associated with various learning outcomes such as improved memory performance in the laboratory (Cepeda et al. 2006) and increased academic performance in educational settings (Son and Simon 2012).

Overall, the literature suggests that students can effectively learn how to read research articles when instructional supports such as scaffolds are aligned with the properties of texts, including headings that match a disciplinary structure such as APA style. The use of disciplinary structure can help students develop schema for understanding articles. Unfortunately, many researchers (Anisfield 1987; Gillen 2006; Levine 2001; Locke et al. 1998; Suter and Frank 1986) have simply provided scaffolds that students can use to find information within an article without any theoretical framing for why they should work to improve performance nor any empirical evidence about the effectiveness of the guides for improving comprehension.

Several studies have, however, provided empirical evidence that students learn to comprehend research articles. For example, Bachiochi et al. (2011) designed a set of open-ended questions that students answered about a condensed psychological research article. Students completed this assessment during the first week of the semester, and then completed the same assessment again toward the end of the semester. Overall, students improved on the second assessment, with the most improvement seen among advanced students (e.g. who were currently enrolled in a second research methods class). As a second example, Sego and Stuart (2016) had students complete a set of open-ended and standardized questions to assess their understanding of multiple empirical articles from the Psi Chi Journal of Undergraduate Research. Students either answered these questions by completing a worksheet, or by writing a summary that addressed the questions, depending on their course section. Students' scores increased between the first and last worksheet (or between the first and last summary assignment) over the semester.

While several studies have presented empirical evidence that instructional scaffolds boost learning, these scaffolds often consist of questions that apply to just one specific research article (Bachiochi et al. 2011; Christopher and Walter 2006; but see Van Lacum et al. 2014 and Sego and Stuart 2016 for examples of questions that apply to multiple articles) or examine learning outcomes that are not specific to understanding research articles, such as exam scores (Christopher and Walter 2006), course grades (Robertson, 2012), or opinions about the safety and efficacy of medical procedures (Russell et al. 2004). When learning outcomes focus on understanding research articles, researchers have often compared performance on assignments from the beginning and end of a semester (Bachiochi et al. 2011; Sego and Stuart 2016) to assess learning that has occurred rather than using separate assessments of comprehension, which does not allow for comparisons with students who have not completed the assignments. Further, previous research used condensed versions of research articles (Bachiochi et al. 2011) or research articles that were written specifically for students (Sego and Stuart 2016).

We have designed an instructional scaffold, the reading worksheet (RW) assignment, to facilitate students' learning that leverages what we know about the structure of scientific genre, high-quality instructional design, and how people learn by building schema. Following the suggestions of Hmelo-Silver et al. (2007) and Goldman and Pellegrino (2015), we aimed to support students' learning through a scaffold that pushed the complex task of research article comprehension into students' zones of proximal development. We were inspired by published guides for reading and understanding research. We aimed to develop questions that were more general than the 12 steps suggested by Locke et al. (1998), yet more specific than the vague suggestion to simply summarize the study, as suggested by Gottfried et al. (2009). Our RW has a set of standard questions which can be applied to any empirical journal article (Lippman et al. 2008). Four out of the six questions in our RW assignment (see description in the Method) draw from the major sections of a typical APA-style research article: Introduction, Method, Results, and Discussion. As argued by Madigan et al. (1995), the structure of an APA-style research article helps guide students' understanding of research. This standard set of questions provided an appropriate scaffold for students to be able to read research articles that were published for psychologists, thus providing students with the opportunity to engage in authentic scientific practices (c.f. Chinn and Malhotra 2002).

The development of the RW assignment was also inspired by what is known about the cognitive principles that underlie knowledge acquisition (e.g., Dunlosky et al. 2013). Students completed the RW assignment at least seven times throughout the semester and reviewed the answers in class. These principles were applied to design a rigorous approach to using the RWs during instruction. For example, by completing the RWs, students engaged in deliberate practice. As a second example of a cognitive principle, students completed a RW assignment every 1–2 weeks throughout the semester, thus engaging in distributed practice. In addition, unlike previous research (Bachiochi et al. 2011; Christopher and Walter 2006; Sego and Stuart 2016), we used a multiple choice assessment that was separate from instruction, called the Journal Article Comprehension (JAC) assessment. The JAC contained seven multiple-choice questions that covered the main parts of a research article: purpose, participants and procedure, independent variables, dependent variables, results, conclusions, and criticisms. These multiple-choice questions corresponded to information that was covered in the standard questions for the RW assignment. The JAC assessment thus allowed us to show transfer between instruction and assessment, and also addressed the practical concern of being able to effectively administer and score performance assessments in larger classes.

Pilot development of the Journal Article Comprehension assessment

We developed the JAC assessment through two pilot studies. To create alternate forms of the JAC assessment, we used two empirical journal articles. We included articles with topics that are typically covered in all cognitive psychology courses and that were published in *Psychological Science* because this is the source of many of the journal articles used for RW assignments. Specifically, we chose the article by Lee et al. (2006) about top-down influences on perception, and the article by Beilock and Carr (2005) about effects of working memory on performance. The multiple-choice questions were matched across the alternate forms of the JAC so that the same number and type of answer options were available for each article.

In the first pilot study, the JAC assessment was administered as a pre-test at the beginning of the semester and as a post-test at the end of the semester: we counterbalanced which form (Beilock and Carr 2005 or Lee et al. 2006) was presented at each time point (see Appendix 1 for more details about the method of the first pilot study). We found an unexpected effect of order, in that there was a larger increase between pre- and post-test when students had the Beilock and Carr (2005) JAC form as the post-test (see Appendix 1 for an overview of the results and analyses). After statistically correcting for this order effect, we found that the experimental group, who completed RWs, showed significant improvement between their pre- and post-test, whereas the control group, who did not complete RWs and only used a textbook, did not show any improvement. We conducted a second pilot study because we wanted to experimentally remove the order effect found in the first pilot study.

In the second pilot study, we administered both forms of the JAC at each time point (pre- and post-test) so we could determine whether one of the alternate forms was a better assessment of student comprehension of research articles over time. We only used an experimental group who completed RWs in the second pilot study (see Appendix 1 for more method details about the second pilot study). We also manipulated the content of the tutorial to match one of the alternate forms to test the relative effectiveness of each article for training.

In the second pilot study, performance on the alternate forms of the JAC was not correlated at post-test, so each form was analyzed separately to assess the impact of tutorial content on change over time (see Appendix 1 for an overview of the results and analyses). For the Beilock and Carr JAC form, the only significant result was that scores improved over time (from pre- to post-test) for all students. For the Lee et al. JAC form, we only found a significant interaction between time and tutorial content: if the Lee et al. tutorial was received, students' scores increased (albeit not significantly), whereas if the Beilock and Carr tutorial was received, students' scores decreased (albeit not significantly).

It was clear from the results of the pilot studies that the two JAC forms were not equivalent in their efficacy of measuring student comprehension of research articles. The form of the JAC that used the Beilock and Carr (2005) content was more robust because performance over time was less affected by differences in the content of the tutorial. An examination of the structure of the two articles used in the alternate JAC forms revealed that Beilock and Carr used standard APA-style headings whereas Lee et al. (2006) did not. All the articles used in RWs had the standard APA-style headings. For these reasons, we decided to use the Beilock and Carr JAC form for the current experiment described below.

Overview of the current experiment

In the following experiment, we describe the implementation of our instructional scaffold (the RW assignment) to show how students can improve their comprehension of research articles within a target content area, cognitive psychology. To do this, we compare an experimental group and a control group on their changes in performance over time on both our JAC assessment and a more traditional summary assessment.

We decided to include an additional summary task because this kind of task is considered to be an effective measure of learning, and we wanted to include multiple measures of understanding. Bretzing and Kulhavy (1979) found students' summaries were correlated with their performance on later tests containing this information. The content of student summaries has also been shown to be correlated with multiple measures of the same information (Dyer et al. 1979). Dyer et al. (1979) found that students were more likely to answer test questions correctly if the information necessary to answer a particular question was included in their summaries. Additionally, the quality of student summaries has been associated with better performance on other tests of the same content. Bednall and Kehoe (2011, Experiment 2) found that summaries which contained more information and links to previously-learned information were associated with higher performance on subsequent tests.

In the current experiment, we tested the effectiveness of our RW instructional scaffold for increasing research article comprehension skill. We operationalized distributed and deliberate practice within multiple RWs and used multiple measures of comprehension skill including the JAC assessment and a summary task. In addition, we assessed the relationship between performance changes on the JAC and the summary task.

Method

Participants

One hundred thirty-six students enrolled in undergraduate cognitive psychology courses at the University of Massachusetts Dartmouth completed the JACs as course assignments. The experimental group consisted of 101 students who were taught by the first author. These students completed seven or nine RW assignments across the course of the semester. There were no differences in the JAC or summary task in the experimental student groups who completed seven or nine RW assignments, and thus these groups were combined. As with the samples used in the pilot studies, the experimental group did not have a course textbook and only used research articles as course readings. As with the experimental group in the first pilot study, these students received a tutorial after the pre-test, but before completing the first RW, about how to find information in a journal article, using an article that was not part of the JAC.

The control group consisted of 35 students who were taught by the third author. The primary reading materials for these students were from Reisberg's *Cognition* (5th edition) textbook. No research articles were assigned as readings. Students in the control course had one assignment involving a research article during the semester: They had to find and summarize an article related to embodiment in cognitive psychology. Students were told to include purpose/goals, procedural descriptions, IVs and DVs, and results, and implications (the same things that are asked in the RWs) as part of their summary. The key difference is that they were not asked about each part of their empirical article explicitly, nor did they perform this summary multiple times over the course of the semester on a set of class-shared articles. In addition, the control students did not receive a tutorial about how to locate information in a research article.

The key difference between the experimental and control groups was the number of writing assignments that involved interaction with research articles. The experimental group had seven or nine writing assignments, while the control group had one. Because both groups were in the same course at the same institution, the instructors had to meet the same curricular goal, which was for students to be able to understand research data in order to be able to relate it to theories and practical applications. The instructors both met this goal through their choice of assignments. Further, comparison of course syllabi and teaching materials revealed that the first and third author covered 90% of the same units (attention, memory, language, etc.) and approximately 75% of the same topics within these units. Discussion of teaching approaches revealed that both instructors had face-to-face lecture courses with instructional materials available through online course management software.

Both instructors took a "building blocks of cognition" approach to the course by beginning with basic cognitive processes and moving towards more complex cognitive processes, and both explained research studies, including their methodology and results, during lecture. Given the similarities in course content and teaching style between the experimental and control groups, we believe that the differences in outcomes explained in the Results below are due to the scaffolding and number of research article assignments that each group completed.

Although the first and third authors were the instructors for the experimental and control groups, respectively, student data were not examined or analyzed until after the semester was over. Data were labeled with participant numbers and were analyzed without any identifying information. We followed this procedure to minimize instructor bias on data analysis. In addition, the third author was not part of the research team until more than 6 months after data were collected from her course. Data were originally collected from the third author's course as part of departmental curricular assessment. When teaching her cognitive psychology course, she was unaware of the goals of the overall project, and designed her course to meet curriculum requirements, without knowledge of the first author's approach or the research project that was already in progress.

Materials

Reading worksheet assignments (RWs)

Participants in the experimental group completed seven or nine RWs over the course of the semester, each on a different empirical article taken from a cognitive psychology journal and related in content to the topics of the class. Brief articles of 10 pages or less were selected to align with course topics (see Table 1). The articles were all published within 15 years of when the course was taught and contained a single experiment or quasi-experiment.

The five standard questions for the RW assignment were:

1. What was the purpose of this research?
2. What did the researchers do? (Summarize method, including participants, procedure, independent variables, and dependent variables).
3. What were the main results?
4. What is the take-home message (conclusion)?
5. What criticisms do you have of this research? Is there anything you would do differently?

In addition to the five standard questions, a specific application question was assigned for each article in which students were asked to apply the findings of the article to their own lives or to areas of public interest. For example, the application question for one of the attention articles (Cooper and Strayer 2008) asked students to apply the ideas of divided attention from the article to cell phone and driving laws passed in the state.

Table 1 Course topics, article topics, and chosen articles for the reading worksheet assignment

Course topic	Article topic	Article used for RW assignment
Perception and pattern recognition	How visual recognition is shaped by expertise	Hargreaves, I. S., Pexman, P. M., Zdrazilova, L., & Sargious, P. (2012). How a hobby can shape cognition: Visual word recognition in competitive Scrabble players. *Memory & Cognition, 40*, 1–7[a,*]
Attention and automaticity	Divided attention for driving and talking on a cell phone and the role of experience in these two tasks OR How the timing of an interruption affects reading comprehension	Cooper, J. M., & Strayer, D. L. (2008). Effects of simulator practice and real-world experience on cell-phone-related driver distraction. *Human Factors,50*, 893–902[a] OR Pashler, H., Kang, S.H.K., & Ip, R.Y. (2013). Does multitasking impair studying? Depends on timing. *Applied Cognitive Psychology, 27*, 593–599[b]
Working memory	Individual differences in working memory and text comprehension OR How individual differences in working memory affect the integration of information across multiple texts	Sanchez, C.A., & Wiley, J. (2009). To scroll or not to scroll: Scrolling, working memory capacity, and comprehending complex texts. *Human Factors, 51*, 730–738[a] OR Banas, S., & Sanchez, C.A. (2012). Working memory capacity and learning underlying conceptual relationships across multiple documents. *Applied Cognitive Psychology, 26*, 594–600[b]
Long-term memory storage	Implicit priming from texting OR How individual differences in working memory impact organization in LTM	Topolinksi, S. (2011). I 5683 you: Dialing phone numbers on cell phones activates key-concordant concepts. *Psychological Science, 22*, 355–360[a] OR Unsworth, N., Spillers, G.J., & Brewer, G.A. (2012). The role of working memory capacity in autobiographical retrieval: Individual differences in strategic search. *Memory, 20*, 167–176[b]

Table 1 (continued)

Course topic	Article topic	Article used for RW assignment
Long-term memory encoding and retrieval	Creating false memories using a doctored photograph OR (Using testing to enhance encoding of information into LTM AND Effects of emotional reaction on memory for Red Sox-Yankees 2004 ALCS Game 7)	Wade, K.A., Garry, M., Read, D.J., & Lindsay, D.S. (2002). A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic Bulletin & Review, 9*, 597–603[a] OR (Roediger, H.L, III., & Karpicke, J.D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255[b] AND Kensinger, E., & Schacter, D.L. (2006). When the Red Sox shocked the Yankees: Comparing negative and positive memories. *Psychonomic Bulletin & Review, 13*, 757–763[b])
Language	Reading comprehension and expectation of character behavior OR How differences in spoken language affect LTM	Egidi, G., & Gerrig, R.J. (2006). Readers' experiences of characters' goals and actions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 1322–1329[a] OR Fausey, C.M., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review, 18*, 150–157[b]
Reasoning and decision making	How word choice impacts difficulty of Wason conditional reasoning task	Nickerson, R.S., & Butler, S.F. (2008). Efficiency in data gathering: Set size effects in the selection task. *Thinking & Reasoning, 14*, 60–82[b]
Problem solving and creativity	Effects of pressure and individual differences in working memory on strategy choice in problem solving OR How time of day effects on inhibitory control processes affects insight and analytical problem solving	Beilock, S.L., & DeCaro, M.S. (2007). From poor performance to success under stress: Working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 983–998[a] OR Wieth, M.B., & Zacks, R.T. (2011). Time of day effects on problem solving: When the non-optimal is optimal. *Thinking & Reasoning, 17*, 387–401[b]

Table 1 (continued)

Course topic	Article topic	Article used for RW assignment
Skill acquisition and expertise	Use of deliberate practice in non-expert performance OR How deliberate practice and individual differences in working memory affect piano sight-reading performance	Keith, N., & Ericsson, K.A. (2007). A deliberate practice account of typing proficiency in everyday typists. *Journal of Experimental Psychology: Applied, 13*, 135–145[a] OR Meinz, E.J., & Hambrick, D.Z. (2010). Deliberate practice is necessary but not sufficient to explain individual differences in piano sight-reading skill: The role of working memory capacity. *Psychological Science, 21*, 914–919[b]

[a] Readings for experimental group that completed seven RWs

[b] Readings for experimental group that completed nine RWs

*Hargreaves et al. was an assigned, graded reading for the seven RW group, and a practice assignment for the nine RW group

JAC assessment

All participants completed the JAC twice. The same article, Beilock and Carr (2005), was used for both the pre-test and post-test JAC assessment. The JAC assessment questions and scoring rubric can be found in Appendix 2. This article, and all the journal articles chosen for the RW assignments across the semester, followed a standard APA-style structure. Thus, the JAC assessment for the Beilock and Carr article was a fair test of what students were learning across the semester.

Summary task

All participants completed the summary task twice. Participants were asked to write a summary of the article. Specifically, they were told that their summaries should address the following guiding questions:

1. What are the most important points of the study in this article? Include enough detail about the goals, design, and method of the study to put those points in context.
2. What are the real-world implications of the findings of this study?
3. Are there any potential weaknesses in the design or procedure of this study that limit how much you trust the findings?

Participants were instructed to take no more than 50 min to read the article and write the summary, and to not write a summary that was longer than one typed page (500 words).

Procedure

The pre-test was completed during the second week of the semester, and the post-test was completed during the 14th week of the semester. Both assessments were completed at each time through online survey software. Students received assignment credit for completing the assessments but their responses were not graded. Participants always wrote the summary first, then completed the multiple-choice JAC assessment.

Analysis

Summary scoring

We chose to analyze the aspects of the summaries that addressed the first guiding question, as these aspects best aligned with the information that was tested in the JAC assessment. We developed a coding scheme to address details about the goals, design, method, results, and conclusions of the article that students included in their summaries. Seven codes were developed to capture the main types of information we expected to see in the summaries: goals/hypothesis, sample, procedure, IVs, DVs, findings/results, and conclusions/interpretations. These codes, and the scoring system, were developed to be similar to the way in which RW assignments were graded for the experimental group. For each of these codes, students could receive a 0 (information missing), a .5 (incomplete or partially correct information), or a 1 (complete and correct information). In addition, for IVs and DVs, students could be awarded a .75 if they included information about variables that could count as IVs or DVs but did not label this information as such (see Table 2). The maximum possible score a student could get for the article summary was 7. Examples of coded summaries are provided in Appendix 3.

The first and third author coded 10 summaries for practice to make sure they understood the coding scheme, as suggested by Chi (1997). They then coded 37 summaries [16% of the screened data (see below)], which included a mix of experimental and control group data, as well as a mix of summaries that were written during the pre-test or post-test. Because random participant numbers were used to choose the summaries, the authors were blind to group membership and when the summary was written. The authors' inter-rater agreement for this set of 37 summaries was $\kappa = .81$, SE = .03, 95% CI .75–.86.

In order to have at least the research standard of 20% of the data coded (M.T.H. Chi, personal communication, June 7, 2017; S.R. Goldman, personal communication, June 10, 2017), the first and third author coded an additional 20 randomly chosen summaries. The inter-rater agreement for the 20 additional summaries was $\kappa = .87$, SE = .03, 95% CI .81–.94. The inter-rater agreement for all 57 summaries, or 25% of the data, was of $\kappa = .81$, SE = .02, 95% CI .77–.86. This level of inter-rater agreement meets Gwet's (2014) standards for reporting inter-rater reliability, and the level of agreement reached, both with the exact kappa value and the 95% confidence interval, meets Landis and Koch's (1977) criteria for substantial agreement and Fleiss' (1981) criteria for excellent agreement. The first author then coded the remaining summaries, blind to group membership and whether the summary was written during the pre-test or post-test.

Participant screening

Of the original 136 participants, 22 were removed from analysis for not completing the JAC assessment, the summary task, or neither at either pre-test or post-test. Specifically, 10 participants from the experimental group did not complete the summary and JAC assessments at post-test and one experimental group participant completed the summary but did not complete the JAC at post-test. One participant from the experimental group and two from the control group did not complete the JAC assessment at pre-test, yet completed the summary task. One control group participant did not complete the summary task at post-test, yet completed the JAC assessment. Three participants from the control group did not complete the JAC assessment nor the summary task, one at pre-test and two at post-test. In addition, we removed four participants, three from the experimental group and one from the control group, who had outlying scores on either the JAC or summary task. Thus, 114 participants were left for analysis, 86 from the experimental group (68 female, 18 male) and 28 from the control group (22 female, 6 male).

To test differences between pre- and post-test among the groups, we used factorial analyses of variance (ANOVAs) with the time of test as the repeated measures variable, and group (experimental or control) as between-subjects variable. Because ANOVA can be sensitive to violations of normality, such as unequal sample sizes, we consulted Tabachnick and Fidell (2013). According to Tabachnick and Fidell, "[t]here is no need for a formal test of homogeneity of variance…since the ratio sample sizes is less than 4:1…and there are no outliers" (p. 232). Our ratio of sample sizes is 3.07 (86/28) after the data were screened for outliers. Thus, we report results using unadjusted ANOVAs.

Table 2 Coding scheme for summaries

Item	Key	Missing (0)	Incomplete or partially correct (.5)	Complete and correct (1)
Hypothesis/goals	The authors' purpose was to explain how individual differences in WM capacity affect susceptibility to choking under pressure in mathematical problem solving			
Methods: Sample	93 undergraduate students			
Methods: Procedure	Screened for WM capacity did low- and high-demand modular arithmetic problems under low- and high-pressure conditions			
IVs	Problem demand (low vs. high), pressure (low vs. high), working memory capacity (low vs. high) Note: Assign .75 if participant mentions correct variables and describes them as IVs without using the term IV Assign 1 if completely correct and using correct term			
DVs	Accuracy on modular arithmetic problems and RT on correct problems Note: Assign .75 if participant mentions correct variables and describes them as DVs without using the term DV Assign 1 if completely correct and using correct term			
Results	High WM subjects showed lower accuracy on high-demand problems in the high-pressure condition (low WMs not affected by pressure) and all subjects were slower on high-demand problems and under high-pressure Note: If something like "x group was better" but not specific, give .5			
Interpretations	People with high WM capacity are more likely than people with low WM capacity to choke under pressure in high-demand situations. Note: Don't pull interpretation out of what participant marks as implications			

Results

JAC assessment

A mixed analysis of variance (ANOVA) was used with the time of test (pre and post) as the repeated measures variable, and group (experimental or control) as the between-subjects variable. Descriptive statistics are in Table 3 (additional item-level descriptive statistics are in Appendix 4). The main effect of time of test did not reach significance, F (1, 112) = 1.18, p = .28, η_p^2 = .01. There was a significant main effect of group, F (1, 112) = 23.56, p = .0001, η_p^2 = .17, indicating that the experimental group had higher overall scores. The interaction between time of test and group was also significant, F (1, 112) = 5.39, p = .02, η_p^2 = .05. Follow-up paired-samples t-tests indicated that the experimental group improved between pre-test and post-test, t (85) = − 4.50, p = .0001, d = .50, while the control group's scores did not change between pre-test and post-test, t (27) = .47, p = .64, d = .08.

Summary task

A mixed analysis of variance (ANOVA) was used with the time of test as the repeated measures variable and group, experimental or control, as the between-subjects variable. Descriptive statistics are in Table 3 (additional item-level descriptive statistics are in Appendix 5). There were significant main effects of time of test, F (1, 112) = 6.99, p = .01, η_p^2 = .06, and group, F (1, 112) = 6.68, p = .01, η_p^2 = .06. These main effects were qualified by a significant interaction, F (1, 112) = 5.19, p = .03, η_p^2 = .04. Follow-up paired-samples t-tests indicated that the experimental group improved on the summary task between pre-test and post-test, t (85) = − 4.86, p = .0001, d = .50, while the control group did not, t (28) = − .23, p = .82, d = .06.

Relationship between JAC and summary assessments

We tested the relationship between the two assessments using correlations. Given the group differences found within the assessments, we examined correlations within each group. These correlations are summarized in Table 4. The experimental group showed a significant positive correlation between JAC scores at pre-test and post-test and summary task scores at pre-test and post-test. Post-test JAC scores and summary task scores were positively correlated as well, but there was no significant correlation between these measures at pre-test.

The control group, in contrast, showed a significant negative correlation between JAC scores at pre-test and post-test, and a significant positive correlation between summary task

Table 3 Mean accuracy on JAC assessment and summary task at pre- and post-test by group

Group	JAC		Summary task		*n*
	Pre	Post	Pre	Post	
Experimental	.75 (.11)	.81 (.10)	.42 (.16)	.52 (.18)	86
Control	.71 (.13)	.68 (.15)	.38 (.18)	.39 (.16)	28
Total	.74 (.11)	.78 (.12)	.41 (.17)	.49 (.19)	114

Values are proportion scores represented in the form M (SD)

Table 4 Intercorrelations between pre- and post-test JAC assessment scores and summary task scores as a function of group

Assessment	1	2	3	4
1. JAC pre		− .42*	− .42*	.04
2. JAC post	.33*		.34	.13
3. Summary pre	.13	.17		.53*
4. Summary post	.16	.30*	.36*	

Intercorrelations for participants in the control group (n = 28) are presented above the diagonal and intercorrelations for the participants in the experimental group (n = 86) are presented below the diagonal

*p < .05

scores at pre-test and post-test. JAC scores and summary task scores were negatively correlated with each other at pre-test, and not significantly correlated at post-test.

Discussion

Across our experiment and two pilot studies, we established that students who are taught how to understand research articles through a scaffolded course assignment improved on our assessment of Journal Article Comprehension (JAC). The experimental group showed consistent improvement while students in control classes, who were not directly taught how to understand research articles, did not improve. The control group received the same assessments at pre-test and post-test as the experimental group, so our results show that the improvement of the experimental group is not due to testing effects. Rather, students in the experimental group acquired fundamental science process skills (Coil et al., 2010) including identifying the purpose of the research from the Introduction (Newell et al. 2011; Van Lacum et al. 2014), identifying independent and dependent variables (Dasgupta et al. 2014), and identifying the main findings from the Results (Dasgupta et al. 2014; Zieffler and Garfield 2009).

In addition, our research contributes to what is known about the cognitive principles that underlie knowledge acquisition (Dunlosky et al. 2013). The RW assignment serves as a scaffold for a complex, higher-order cognitive task of reading and understanding research articles. The RW directs students' attention to key features of the text (c.f. Hmelo-Silver et al. 2007) and problematizes the research article comprehension process (c.f. Reiser 2004) as students have to do extra work of answering questions they would not otherwise have to answer about an article to complete the RW assignment. Performance on this task improved through deliberate practice (Ericsson and Charness 1994) as students actively engaged in structured identification of key elements of research articles. As shown by our data and by Macnamara et al. (2014), deliberate practice has small yet significant effects on educational outcomes. Practice was also distributed over time as students completed a RW assignment every 1–2 weeks throughout the semester. As noted by Morris et al. (2012), scientific reasoning develops slowly, therefore distributing learning has been shown to increase performance in academic settings (Son and Simon 2012).

Students in the experimental group also improved on the summary task, while the control group did not. The experimental group students were able to respond to the open-ended summary questions because they had engaged in extensive deliberate practice of identifying key information in research articles. The experimental group's post-test summary task scores were correlated with their post-test JAC scores but there was no such correlation for the control group. This result suggests that students in the experimental group developed discipline-specific science process skills (Coil et al. 2010) that allowed them to apply Journal Article Comprehension (JAC) skills to a near-transfer task of writing a summary. The ability of students in the experimental group to transfer these skills to a new context suggests that they developed coherent research schemas (Chinn and Brewer 2001) and learned what information in a research article is important to attend to (c.f. Morris et al. 2012; Van Lacum et al. 2014).

Our research goes beyond previous empirical evidence of research article comprehension (Bachiochi et al. 2011; Christopher and Walter 2006; Sego and Stuart 2016) by measuring student learning of scientific literacy skills with an assessment that is not simultaneously a course assignment. As such, our assessment strategy allowed us to compare performance of experimental and control groups, and allowed us to examine the impact of class assignments on learning. Further, unlike previous research, we presented students with primary sources that were written by practicing scientists for other scientists rather than using readings that were crafted for students (c.f. Sego and Stuart 2016; Yarden 2009). Our research shows that with sufficient instructional scaffolds, students are able to read and understand research articles that were written for expert scientists, thus meeting the instructional design principle that instructional scaffolds should make a challenging task manageable for learners (Goldman and Pellegrino, 2015; Hmelo-Silver et al. 2007).

Limitations

There are limitations to our research. First, we only targeted cognitive psychology courses. Our findings would have been more generalizable if we had included additional levels of psychology instruction or if we had included courses from different fields. Although the first author has adapted the RW assignment for introductory psychology courses, data have not been collected for those groups. We are currently planning to apply the RW instructional method to the fields of biology and applied linguistics.

Second, we have shown improved student learning in only one aspect of text comprehension, the textbase. The RW assignment was not designed to measure students' ability to build a situation model of cognitive psychology research. Each RW is a separate assignment for a specific article. Students are not asked to compare and contrast different research articles. In a different manuscript, however, we further explored improvement in students' research article comprehension by examining the types of criticisms that they generate over the course of the semester (Kershaw et al. in preparation). One of the standard RW questions asks students to generate criticisms of the research article or suggestions for future research. We found, over the duration of a semester, that students produced fewer critiques of external validity and more critiques of internal validity (Kershaw et al. in preparation). This finding suggests that students are learning how to extract important information about research studies from journal articles, as well as and what types of critiques are appropriate for empirical research in cognitive psychology.

Third, the JAC assessment only contained one item to measure each research knowledge construct present in the instrument. Creation of additional items per construct would allow for an assessment of internal reliability to compare the efficacy of the JAC assessment to other published assessments. For example, Christopher and Walter (2006) report a Chronbach's alpha of .78 for their assessment, which consisted of four questions about interpreting one statistical analysis. In future work, we plan to develop additional questions to assess each construct of the JAC.

Fourth, student characteristics could have influenced the outcome of the experiment. Participants in our sample can take cognitive psychology at different points in in the curriculum. Students must complete a statistics course as a prerequisite, but students in the course range from sophomores to graduating seniors. This was the case for both the experimental and control groups, so any potential effects should apply equally to both groups. Thus, students in both groups likely differed in background knowledge of psychology, experience with psychological research, and the number of research articles that they had ever read. Previous research has shown that students' prior knowledge can affect learning outcomes (cf., Shapiro 2004). Thus, future research should measure and control for the amount of knowledge and experience students have concerning psychological research.

Implications for instruction

Our RWs required deliberate practice and distributed practice, thus providing an instructional scaffold that promoted student learning. Because Coil et al. (2010) noted that instructors want students to learn how to comprehend research, yet often feel that they do not have time to do so, we address how our particular scaffold can be transformed into general instructional guidelines.

A first guideline is that assessment materials must be chosen that follow the same structure as the training materials. As shown in the second pilot study, the use of a research article for assessment that did not follow the structure of research articles failed to produce a significant change between pre-test and post-test JAC scores. Assessment materials with the same structure as training materials lead to the activation of students' "research article" schemas and thus allowed transfer to occur. In our experience, research articles that follow standard APA style help students to engage in text comprehension processes that build expectations for upcoming information in the article (Madigan et al. 1995). APA style forms a common structure that students can use to navigate and understand research articles, despite differences in content.

A second guideline is that practice needs to be distributed throughout the semester (c.f. Son and Simon 2012), although the exact amount is an empirical question. We found no differences on our assessments between students that completed seven or nine RWs, but we do know that more than one is necessary (as evidenced by the lack of improvement in the control group who summarized one article over the semester). Future research should explore how much practice is needed for successful transfer. We are currently conducting a study that compares a group of students that completes nine RWs to a group of students that completes four RWs plus five research article assignments that ask them to connect the findings of the experiments to previous knowledge or personal experience. This new study is especially important because it equates instruction and exposure to research articles across the two groups. The main experiment reported in this paper did not equate groups in that way, and thus we cannot tease apart the effects of instruction vs. exposure to research articles on the gains shown by the experimental group. Yet, as previously mentioned, students in both the experimental and control groups were at differing points in their career. Therefore, individuals in each group should vary similarly with regards to the amount of previous exposure to journal articles.

A third guideline is that instruction in one skill may transfer to related tasks, as previous data have shown (Gadgil and Nokes-Malach 2012). This was also the case with our participants who improved in the summary task although they did not have direct instruction on writing open-ended summaries. It is possible that more specific guiding questions may be sufficient training tools for open-ended written summaries (cf. Sego and Stuart 2016).

Overall, we believe that the use of instructional scaffolds, namely our RWs, provided students with a structured way to engage in dissecting research articles which translated across specific articles and assessment formats. As we have shown, RWs with instructional feedback are relatively easy to employ in the classroom, yet reap large rewards in students' abilities to understand research. In addition, although not highlighted here, the use of research articles rather than a standard textbook, can save students money and proves to be more effective in achieving at least one common goal among upper-level, discipline-specific classes—the understanding of empirical research.

Acknowledgements The authors thank Amy Shapiro and Scott Hinze for access to their cognitive psychology courses. We thank James Bradley for his assistance in creating the summary coding scheme. Judy Sims-Knight gave invaluable advice on several statistical issues, and Susan Goldman and Micki Chi provided important suggestions for best practices in establishing inter-rater reliability. We also thank several anonymous reviewers for their feedback. Preliminary results from the pilot studies were presented at the 15th Annual European Association for Research on Learning and Instruction conference (EARLI 2013, Munich, Germany) and the 4th Biennial Conference of the International Society for the Psychology of Science and Technology (ISPST, 2012, Pittsburgh, PA).

Appendix 1: Method, analysis, and results details, pilot studies

Pilot study 1

The participants of the first pilot study included 78 students from an undergraduate cognitive psychology course in the experimental group who completed RWs, and 78 students from a different undergraduate cognitive psychology course in the control group who only used a textbook as their instructional materials. Students completed one form of the JAC as a pre-test at the beginning of the semester and one form as a post-test as the end of the semester with order of forms counterbalanced. After completion of the JAC pre-test, but before completion of the first RW assignment, the experimental group received a 1-h tutorial covering strategies for locating information within journal articles needed to complete a RW assignment. This tutorial used a research article that was not part of the JAC assessment. The control group did not receive a tutorial.

The first pilot study showed an unexpected effect of order, in that there was a larger increase between pre- and post-test when students had the Beilock and Carr (2005) JAC for the post-test. To account for these order effects, we included order as a covariate in a factorial ANCOVA with a repeated measures factor of time (pre-test vs. post-test) and a between-subjects factor of group (experimental vs. control). All main effects were significant as well as the interactions, as shown in the table below.

Source	SS	MS	F	p	η_p^2
Time	.46	.46	27.42	.0001	.15
Group	.61	.61	28.33	.0001	.16
Order	.24	.24	11.28	.001	.07

Source	SS	MS	F	p	η_p^2
Time × group	.13	.13	7.85	.01	.05
Time × order	.41	.41	24.55	.0001	.14
Within cells	2.56	.02			
Between cells	3.27	.02			

df = 1, 153

To further explore the time x group interaction, paired-samples t-tests were conducted within the experimental and control groups. The experimental group showed significant improvement between their pre-test (*M* = .73, SD = .14) and post-test proportion scores (*M* = .79, SD = .14), *t* (77) = −2.53, *p* = .01, *d* = .32, while the control group's scores did not change significantly between pre-test (*M* = .68, SD = .14) and post-test (*M* = .68, SD = .16), *t* (77) = .24, *p* = .81, *d* = .05.

Pilot study 2

In the second pilot study there was no control group. Participants included 97 undergraduate psychology students in a cognitive psychology course who completed RWs. Similar to the experimental group of the first pilot study, all participants received a tutorial after the pretest, but unlike the first pilot study, the article used for the tutorial was either Lee et al. (2006) or Beilock and Carr (2005).

Performance on the JAC forms was not correlated at post-test (*r* = .15, *p* = .15), so each form was analyzed separately. Using a repeated measures ANOVA on the Beilock and Carr JAC, we found that all students improved from pre-test (*M* = .73, SD = .14) to post-test (*M* = .84, SD = .11) no matter which article they reviewed during the tutorial, *F* (1, 95) = 47.73, *p* = .0001, η_p^2 = .32. There was no main effect of tutorial type, and no interaction, as shown in the table below.

Source	SS	MS	F	p	η_p^2
Time	.65	.65	47.73	.0001	.32
Review type	.01	.01	.69	.41	.01
Time × review type	.03	.03	2.07	.15	.02
Within cells	1.29	.01			
Between cells	1.77	.02			

df = 1, 95

For the Lee et al. JAC, we found no main effect of time and no main effect of review content (see following table). There was, however, a significant interaction between time and review content, *F* (1, 95) = 4.06, *p* = .05, η_p^2 = .04. Specifically, if students received the Lee et al. tutorial, their proportion scores did not significantly increase between pre-test (*M* = .72, SD = .13) and post-test (*M* = .75, SD = .13), *t* (51) = − 1.26, *p* = .21, *d* = .16. If students received the Beilock and Carr tutorial, their proportion scores decreased between pre-test (*M* = .71, SD = .13) and post-test (*M* = .68, SD = .12), although not significantly, *t* (44) = 1.76, *p* = .09, *d* = .23.

Source	SS	MS	F	p	η_p^2
Time	.00	.00	.002	.97	.00
Review type	.07	.07	3.73	.06	.04
Time × review type	.05	.05	4.06	.05	.05
Within cells	1.26	.01			
Between cells	1.79	.02			

df = 1, 95

Appendix 2: Beilock and Carr (2005) JAC questions and scoring rubric

1. Which of these is the **best** statement of the purpose of this research?

    a. The authors' purpose was to examine why some people choke under pressure but others don't. **isn't as specific as answer C** ***partial credit***
    b. The authors' purpose was to explain how math anxiety negatively affects performance on math tests. **although the authors discuss math anxiety in their intro, this is not the purpose of the study** ***no credit***
    c. The authors' purpose was to explain how individual differences in WM capacity affect susceptibility to choking under pressure in mathematical problem solving. ***full credit***
    d. The authors' purpose was to examine how individuals with high WM capacity excel in testing situations. **this is the purpose of many WM articles, but not this one** ***no credit***

2. Which of the following **best** describes the participants and what they did?

    a. 93 undergraduate students completed the operation span and reading span tests and then were split into low and high working memory groups. **this is true but not as good of an answer as B** ***partial credit***
    b. 93 undergraduate students were screened for their WM capacity and then performed low- and high-demand modular arithmetic problems under low- and high-pressure conditions. ***full credit***
    c. 48 undergraduate students completed 93 modular arithmetic problems. **this is wrong** ***no credit***
    d. 93 undergraduate students completed a series of math problems that were adopted from the SAT. **this isn't true at all** ***no credit***

3. Which of the following is/are the independent variable(s)?

    a. Level of math anxiety **not measured at all** ***no credit***
    b. Accuracy on modular arithmetic problems and RT on correct problems **these are the DVs** ***no credit***
    c. Whether the modular arithmetic problem had a large number or required a borrow operation or not **this is the definition of problem demand** ***partial credit***
    d. Problem demand (low vs. high), pressure (low vs. high), and working memory capacity (low vs. high) ***full credit***

4. Which of the following is/are the dependent variable(s)?

   a. Accuracy on modular arithmetic problems and RT on correct problems ***full credit***
   b. Score on the operation span and reading span tests **these scores are used to create an IV** ***no credit***
   c. Performance on the modular arithmetic problems **students need to be more specific** ***partial credit***
   d. Problem demand (low vs. high), pressure (low vs. high), and working memory capacity (low vs. high) **these are the IVs** ***no credit***

5. Which of the following **best** summarizes the important results?

   a. Low WM subjects were slower than high WM subjects to solve the modular arithmetic problems. **while the LWMs were slower on high-demand problems, this is not the best summary** ***partial credit***
   b. High WM subjects showed higher accuracy and were faster in the high-pressure condition. **prediction that was not supported** ***no credit***
   c. High WM subjects showed lower accuracy on high-demand problems in the high-pressure condition (low WMs not affected by pressure) and all subjects were slower on high-demand problems and under high-pressure. ***full credit***
   d. High WM subjects showed higher accuracy on high-demand problems in the high-pressure condition (low WMs not affected by pressure) and all subjects were faster on high-demand problems and under high-pressure. **this is wrong, just trying to have an answer that is the same length as the correct answer** ***no credit***

6. Which of the following is/are valid criticism(s) of the research?

   a. There is demographic information missing about the participants **this is an ok, but not great criticism—the authors do skip a lot of demo. info.** ***partial credit***
   b. Math test and high-pressure condition are unlike what one would experience in the real world **both of these are valid criticisms** ***full credit***
   c. Subjects' baseline math ability was not tested. **this is a nuisance variable and also no Ss would have experience with the modular arithmetic task** ***partial credit***
   d. Only undergraduates were tested. **although WM might change with age, this isn't really a valid criticism for the current study** ***no credit***

7. Which of the following statements is most likely to be true, based on this research?

   a. People only do well on math problems when they are extremely anxious. **wrong** ***no credit***
   b. People choke under pressure because they are worried about what others think about them. **this could be true for the HWM group but not as good as C** ***partial credit***
   c. People with high WM capacity are more likely than people with low WM capacity to choke under pressure in high-demand situations. ***full credit***
   d. People with high WM capacity outperform people with low WM capacity on math problems. **this is true only under low pressure** ***partial credit***

Appendix 3: Sample coded summaries

Subject 1048, experimental group, pre-test

The most important points of study in this article are those claiming that the presence of anxiety interferes with a person's ability to think as diligently as it normally would under normal circumstances. The article is saying that although a person may have the ability to solve math problems and perform well, if the working memory is interrupted by the negative emotion of anxiety, its ability to work at its best is decreased. When anxiety poses a threat to the working memory, it cannot put all its efforts into thinking and solving a problem, it needs to also put effort into figuring out a way to deal with the anxious feeling that is clouding the person's mind. The article says that for people high in working memory, the presence of anxiety poses a threat because so much of their thought processes is used for solving the specific problem and not used for coping with anxious nerves in order to focus on the specific problem.

Subject 1048, experimental group, post-test

The goal of this study was to see if individual differences in WMC may have something to do with a person choking under pressure. 93 undergraduate students were divided into two groups, LWM and HWM, each person was tested on MA problems and used a computer to do so. Each person was given a partner and was told that they could win an award of $5, but that their partner had already completed the task and that it was a team effort. Their scores were dependent on time and accuracy. Results revealed that HWM was not affected by pressure on low-demand problems, however HWM's performance on high-demand started to decline under pressure. In addition, all groups were slower in the low-pressure test than in the high-pressure test. The implications from this study are that HWM does not have an advantage over LWM during high-pressure demand situations.

Subject 5002, control group, pre-test

The research study performed by Beilock and Carr was about whether or not pressure and anxiety during a situation like answering arithmetic problems would affect High Working Memory (HWM) individuals and Low Working Memory (LWM) individuals. It was hypothesized that participants with low working memory are more susceptible to crack under pressure than participants with high working memory due to limited capacity to obtain information and figure out problem solutions. It was also hypothesized that participants with HWM are more susceptible to failure under pressure while answering arithmetic problems than LWMs because the pressure of the situation may deny them the resources/ working memory that they are used to relying on while not in an anxiety provoked situation. The important parts to this study include the findings and results on how participants with HWM did in fact perform worse under pressure during the tests than they would without having pressure. On the other hand, participants with LWM performed equally worse on both the tests with or without added pressure to the situation. So for the LWM group it did not matter if there was added pressure in the situation. Under normal conditions HWM performs better than the LWM group as well because they have high levels of attentional capacities. However, when the attentional capacity was affected, from the pressure induced situation, the HWS advantage disappears. The most important finding is basically that those who have the highest capacity for success or HWM are the individuals who are more susceptible to failing while under pressure.

Subject 5002, control group, post-test

In this study, researchers are interested in whether individuals who rely more on their working memory are influenced by performance pressure while solving mathematics problems than individuals low in working memory. The findings suggest that there was no significant difference in individuals with low capacity working memory and being influenced by performance pressure. The findings further suggest that there was a significant relationship between individuals with high capacity working memory and performance pressure.

Scores received for assessment areas, sample coded summaries

Assessment area	Summary			
	Subject 1048 (experimental)		Subject 5002 (control)	
	Pre	Post	Pre	Post
Overall accuracy	1.5	4.5	3	2
Hypothesis/goals	.5	.5	1	1
Sample	0	1	0	0
Procedure	0	.5	0	0
IVs	0	.75	.5	.5
DVs	0	.75	0	0
Results	0	1	.5	.5
Interpretation	1	0	1	0

Appendix 4

Mean accuracy on each question of the JAC assessment at pre- and post-test by group

Question	Experimental group		Control group	
	Pre	Post	Pre	Post
Purpose	.89 (.27)	.95 (.19)	.79 (.35)	.82 (.34)
Participants and procedure	.84 (.25)	.87 (.22)	.84 (.24)	.84 (.31)
Independent variables (IVs)	.72 (.31)	.84 (.22)	.67 (.38)	.60 (.43)
Dependent variables (DVs)	.52 (.30)	.73 (.28)	.51 (.32)	.45 (.35)
Results	.89 (.31)	.89 (.31)	.79 (.42)	.79 (.40)
Conclusions	.95 (.17)	.95 (.15)	.89 (.25)	.84 (.27)
Criticisms	.46 (.21)	.45 (.21)	.46 (.21)	.46 (.23)
Total JAC score	.75 (.11)	.81 (.10)	.71 (.13)	.68 (.15)

Values are proportion scores represented in the form M (SD). Experimental group, $n=86$. Control group, $n=28$

Appendix 5

Mean accuracy on each item of the summary task at pre- and post-test by group

Item	Experimental group		Control group	
	Pre	Post	Pre	Post
Hypothesis/goals	.80 (.31)	.70 (.38)	.75 (.29)	.79 (.35)
Sample	.37 (.46)	.68 (.41)	.45 (.46)	.46 (.47)
Procedure	.52 (.29)	.61 (.29)	.41 (.27)	.38 (.32)
Independent variables (IVs)	.59 (.22)	.65 (.22)	.52 (.24)	.54 (.26)
Dependent variables (DVs)	.11 (.27)	.30 (.39)	.05 (.20)	.08 (.24)
Results	.35 (.31)	.45 (.32)	.29 (.29)	.30 (.25)
Interpretation	.17 (.29)	.24 (.37)	.21 (.37)	.18 (.34)
Total summary task score	.42 (.16)	.52 (.18)	.38 (.18)	.39 (.16)

Values are proportion scores represented in the form M (SD). Experimental group, $n=86$. Control group, $n=28$

References

Lippman, J. P., Kershaw, T. C., Pellegrino, J. W., & Ohlsson, S. (2008). Beyond standard lectures: Supporting the development of critical thinking in cognitive psychology courses. In D. S. Dunn, J. S. Halonen & R. A. Smith (Eds.), *Teaching critical thinking in psychology: A handbook of best practices* (pp. 183–198). Boston: Blackwell Publishing.

Kershaw, T. C., Lippman, J. P., & Kolev, L. N. (in preparation). Learning to critique published psychological research.

Anisfeld, M. (1987). A course to develop competence in critical reading of empirical research in psychology. *Teaching of Psychology, 14,* 224–227. https://doi.org/10.1207/s15328023top1404_8.

Bachiochi, P., Everton, W., Evans, M., Fugere, M., Escoto, C., Letterman, M., et al. (2011). Using empirical article analysis to assess research methods courses. *Teaching of Psychology, 38,* 5–9. https://doi.org/10.1177/0098628310387787.

Baram-Tsabari, A., & Yarden, A. (2005). Text genre as a factor in the formation of scientific literacy. *Journal of Research in Science Teaching, 42,* 403–428. https://doi.org/10.1002/tea.20063.

Bednall, T. C., & Kehoe, E. J. (2011). Effects of self-regulatory instructional aids on self-directed study. *Instructional Science, 39,* 205–226. https://doi.org/10.1007/s11251-009-9125-6.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science, 16,* 101–105. https://doi.org/10.1111/j.0956-7976.2005.00789.x.

Bretzing, B. H., & Kulhavy, R. W. (1979). Notetaking and depth of processing. *Contemporary Educational Psychology, 4,* 145–153. https://doi.org/10.1016/0361-476X(79)90069-9.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin, 132,* 354–380. https://doi.org/10.1037/0033-2909.132.3.354.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6,* 271–315. https://doi.org/10.1207/s15327809jls0603_1.

Chinn, C. A., & Brewer, W. F. (2001). Models of data: A theory of how people evaluate data. *Cognition and Instruction, 19,* 323–393. https://doi.org/10.1207/S1532690XCI1903_3.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic inquiry in schools: A theoretical framework for evaluating inquiry tasks. *Science Education, 86,* 175–218. https://doi.org/10.1002/sce.10001.

Christopher, A. N., & Walter, M. I. (2006). An assignment to help students learn to navigate primary sources of information. *Teaching of Psychology, 33,* 42–45.

Coil, D., Wenderoth, M. P., Cunningham, M., & Dirks, C. (2010). Teaching the process of science: Faculty perceptions and an effective methodology. *CBE—Life Sciences Education, 9,* 524–535. https://doi.org/10.1187/cbe.10-01-0005.

Cooper, J. M., & Strayer, D. L. (2008). Effects of simulator practice and real-world experience on cell-phone-related driver distraction. *Human Factors, 50*, 893–902.

Dasgupta, A. P., Anderson, T. R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE—Life Sciences Education, 13,* 265–284. https://doi.org/10.1187/cbe.13-09-0192.

Day, J. D. (1983). Teaching summarization skills: Influences of student ability level and strategy difficulty. *Cognition and Instruction, 3,* 193–210. https://doi.org/10.1207/s1532690xci0303_3.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. https://doi.org/10.1177/1529100612453266.

Dyer, J. W., Riley, J., & Yekovich, F. R. (1979). An analysis of three study skills: Notetaking, summarizing, and rereading. *Journal of Educational Research, 73,* 3–7. https://doi.org/10.1080/00220671.1979.10885194.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49,* 725–747. https://doi.org/10.1037/0003-066X.49.8.725.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Gadgil, S., & Nokes-Malach, T. J. (2012). Collaborative facilitation through error-detection: A classroom experiment. *Applied Cognitive Psychology, 26,* 410–420. https://doi.org/10.1002/acp.1843.

Gillen, C. M. (2006). Criticism and interpretation: Teaching the persuasive aspects of research articles. *CBE—Life Sciences Education, 5,* 34–38. https://doi.org/10.1187/cbe.05-08-0101.

Goldman, S. R., & Bisanz, G. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J. A. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 19–50). Mahwah, NJ: Lawrence Erlbaum Associates.

Goldman, S. R., & Pellegrino, J. W. (2015). Research on learning and instruction: Implications for curriculum, instruction, and assessment. *Policy Insights from the Behavioral and Brain Sciences, 2,* 33–41. https://doi.org/10.1177/2372732215601866.

Gottfried, G. M., Johnson, K. E., & Vosmik, J. R. (2009). Assessing student learning: A collection of evaluation tools. *Office of Teaching Resources in Psychology*. Retrieved from http://teachpsych.org/resources/Documents/otrp/resources/gottfried09.pdf.

Gwet, K. L. (2014). *Handbook of inter-rater reliability* (4th ed.). Gaithersburg, MD: Advanced Analytics LLC.

Hmelo-Silver, C., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42,* 99–107. https://doi.org/10.1080/00461520701263368.

Karcher, S. J. (2000). Student reviews of scientific literature: Opportunities to improve students' scientific literacy and writing skills. In S. J. Karcher (Ed.), *Tested studies for laboratory teaching. Proceedings of the 22nd Workshop/Conference of the Association for Biology Laboratory Education* (pp. 484–487).

Landis, J. R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159–174.

Lee, L., Frederick, S., & Ariely, D. (2006). Try it, you'll like it: The influence of expectation, consumption, and revelation on preferences for beer. *Psychological Science, 17,* 1054–1058. https://doi.org/10.1111/j.1467-9280.2006.01829.x.

Levine, E. (2001). Reading your way to scientific literacy. *Journal of College Science Teaching, 31,* 122–125.

Locke, L. F., Silverman, S. J., & Spirduso, W. W. (1998). *Reading and understanding research*. London: Sage.

Lorch, R. F., Jr., & Lorch, E. P. (1996). Effects of organizational signals on free recall of expository text. *Journal of Educational Psychology, 88,* 38–48. https://doi.org/10.1006/ceps.1996.0022.

Macnamara, B. N., Hambrick, D. Z., & Oswald, F. L. (2014). Deliberate practice and performance in music, games, sports, education, and professions: A meta-analysis. *Psychological Science, 25,* 1608–1618. https://doi.org/10.1177/0956797614535810.

Madigan, R., Johnson, S., & Linton, P. (1995). The language of psychology: APA style as epistemology. *American Psychologist, 50,* 428–436. https://doi.org/10.1037/0003-066X.50.6.428.

Morris, B. J., Croker, S., Masnick, A. M., & Zimmerman, C. (2012). The emergence of scientific reasoning. In H. Kloos, B. J. Morris, & J. L. Amaral (Eds.), *Current topics in children's learning and cognition* (pp. 61–82). Rijeka, Croatia: InTech.

Newell, G. E., Beach, R., Smith, J., & VanDerHeide, J. (2011). Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly, 46,* 273–304. https://doi.org/10.1598/RRQ.46.3.4.

Oldenburg, C. M. (2016). Use of primary source readings in psychology courses at liberal arts colleges. *Teaching of Psychology, 32*, 25–29.

Reiser, B. J. (2004). Scaffolding complex learning: The mechanisms of structuring and problematizing student work. *The Journal of the Learning Sciences, 13,* 273–304. https://doi.org/10.1207/s15327809jls1303_2.

Robertson, K. (2012). A journal club workshop that teaches undergraduates a systematic method for reading, interpreting, and presenting primary literature. *Journal of College Science Teaching, 41,* 25–31.

Russell, J. S., Martin, L., Curtin, D., Penhale, S., & Trueblood, N. A. (2004). Non-science majors gain valuable insight studying clinical trials literature: An evidence-based medicine library assignment. *Advances in Physiological Education, 28,* 188–194.

Sego, S. A., & Stuart, A. E. (2016). Learning to read empirical articles in general psychology. *Teaching of Psychology, 43,* 38–42. https://doi.org/10.1177/0098628315620875.

Shapiro, A. M. (2004). How including prior knowledge as a subject variable may change outcomes of learning research. *American Educational Research Journal, 41,* 159–189. https://doi.org/10.3102/00028312041001159.

Son, L. K., & Simon, D. A. (2012). Distributed learning: Data, metacognition, and educational implications. *Educational Psychology Review, 24,* 379–399. https://doi.org/10.1007/s10648-012-9206-y.

Suter, W. N., & Frank, P. (1986). Using scholarly journals in undergraduate experimental methodology courses. *Teaching of Psychology, 13*, 219–221.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Taylor, K. K. (1983). Can college students summarize? *Journal of Reading, 26,* 524–528.

van Gelder, T., Bissett, M., & Cumming, G. (2004). Cultivating expertise in informal reasoning. *Canadian Journal of Experimental Psychology, 58,* 142–152. https://doi.org/10.1037/h0085794.

Van Lacum, E. B., Ossevoort, M. A., & Goedhart, M. J. (2014). A teaching strategy with a focus on argumentation to improve undergraduate students' ability to read research articles. *CBE—Life Sciences Education, 13,* 253–264. https://doi.org/10.1187/cbe.13-06-0110.

Yarden, A. (2009). Reading scientific texts: Adapting primary literature for promoting scientific literacy. *Research in Science Education, 39,* 307–311. https://doi.org/10.1007/s11165-009-9124-2.

Yarden, A., Brill, G., & Falk, H. (2001). Primary literature as a basis for a high-school biology curriculum. *Journal of Biological Education, 35,* 190–195.

Zieffler, A. S., & Garfield, J. B. (2009). Modeling the growth of students' covariational reasoning during an introductory statistics course. *Statistics Education Research Journal, 8,* 7–31.