The keyword effect: A conceptual replication, effects on bias, and an optimization

Julia Waldeyer¹ • Julian Roelle¹

Received: 8 December 2019 / Accepted: 7 July 2020 / Published online: 3 August 2020
© The Author(s) 2020

Abstract

Prompting learners to generate keywords after a delay is a promising means to enhance relative judgment accuracy in learning from texts. However, to date, conceptual replications of the keyword effect without the involvement of the researcher who originally proposed it are still scarce. Furthermore, it is unclear whether generating delayed keywords could reduce bias and whether the benefits of generating delayed keywords could be optimized by having learners compare their keywords with expert ones. Against this background, we conducted an experiment with $N = 109$ university students who read four expository texts and then were randomly assigned to one of three experimental conditions: (a) Generation of keywords after reading, (b) generation of keywords after reading and a comparison with external standards in the form of expert keywords, (c) no keyword generation (control condition). We found that generating delayed keywords significantly increased relative accuracy but did not reduce bias. Furthermore, we found that the comparison with expert keywords enhanced relative accuracy beyond the established keyword effect. However, we also found that the comparison with expert keywords increased bias (here: underconfidence). Overall, these findings suggest that generating and comparing keywords is an effective means to enhance relative accuracy.

Keywords Self-regulated learning · Metacomprehension · Relative accuracy · Bias · Keyword effect

In learning from expository texts, prompting learners to generate keywords that capture the main content of the texts after a delay is an effective means to enhance the accuracy of learners' judgments of learning (JOLs). This so-called *keyword effect* has been found in studies

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s11409-020-09235-7>) contains supplementary material, which is available to authorized users.

✉ Julia Waldeyer
julia.waldeyer@ruhr-uni-bochum.de

¹ Faculty of Philosophy and Educational Research, Ruhr University Bochum, Universitätsstraße 150, 44801 Bochum, Germany

with both high school and university students and with different learning materials (e.g., De Bruin et al. 2011; Thiede et al. 2003; Thiede et al. 2005, 2017; but see Engelen et al. 2018). Despite these promising results, however, there are still some important open issues that need to be addressed.

First, to the best of our knowledge, conceptual replications without the involvement of the researcher who originally proposed the keyword effect (K. W. Thiede) are still scarce (for a rare exception, see Shiu and Chen 2013). In view of the replication crisis in psychological research (e.g., Open Science Collaboration 2015), such *external* replications would substantially strengthen the keyword effect. Second, to date, mainly effects on *relative accuracy* (i.e., discrimination between the level of comprehension for different texts) have been investigated, whereas potential effects on *bias* (i.e., the extent to which learners over- or underconfidently judge their level of comprehension) have widely been ignored. However, for the purpose of optimally supporting learners to come to accurate JOLs, knowing whether generating keywords not only increases relative accuracy but also affects bias would be highly useful.

A third open issue relates to a potential optimization of the keyword effect. The assumed underlying mechanism of the keyword effect is that generating keywords after a delay requires learners to access their situation model-level representation of the respective texts (Griffin et al. 2008; Thiede et al. 2010). Accessing the situation model produces information (usually referred to as ‘cues’) that is predictive for learners’ comprehension (e.g., Thiede et al. 2010). Learners who generate keywords can incorporate these cues into their JOLs, which, in turn, enhances judgment accuracy. We argue that one concrete process via which learners could gather diagnostic information is evaluating the quality of the keywords. More specifically, learners could gather diagnostic information from evaluating how well their generated keywords cover the essence of the respective texts. However, in the absence of *external standards*, learners likely cannot exploit the full potential of this source of information because they cannot accurately judge the degree to which their keywords actually capture the essence of the texts. *Expert keywords* that are provided after learners have generated keywords on their own, could potentially fix this problem and thus contribute to optimizing the keyword effect.

In the present study, we addressed these three open issues. For this purpose, after they read four expository texts, university students either (a) were prompted to generate keywords after a delay, (b) were prompted to generate keywords after a delay and subsequently compare them with expert keywords, or (c) were not prompted to generate keywords. Relative accuracy and bias were used as the main dependent variables.

The keyword effect

An effective way to substantially increase judgment accuracy when learning from texts is to prompt learners to generate keywords that cover the main content of the texts after a delay (e.g., De Bruin et al. 2011; Shiu and Chen 2013; Thiede et al. 2003, 2005, 2017). According to the *situation model approach to metacomprehension* (e.g., Griffin et al. 2008; Wiley et al. 2005), the underlying mechanism via which generating keywords after a delay fosters judgment accuracy is that it requires learners to access their *situation model-level representation* of the respective texts. Situation model-level representations include not only the idea units of a text but also connections among these units and connections with prior knowledge (e.g., Kintsch 1988). When learners access this representation during keyword generation, they can gather cues that inform them about the quality of their situation model-level representation

(e.g., the fluency with which they can generate keywords using this representation). As comprehension tests usually require learners to access their situation model-level representations as well, these cues that emerge during keyword generation entail high diagnosticity regarding learners' level of text understanding (De Bruin et al. 2017).

It is important to note that generating keywords immediately after reading a text does not serve this function (e.g., Thiede et al. 2003, 2005). Immediately after they have read a text, learners can generate keywords by simply accessing their surface- and textbase-level representations of the text. As these representations usually are not tapped by comprehension tests, the cues that learners gather via accessing these representations are not diagnostic with regard to their level of comprehension and thus scarcely foster judgment accuracy. The surface- and textbase-level representations decay faster than the situation model-level representation (e.g., Kintsch et al. 1990). Hence, with an increasing delay of the keyword generation task, the probability that learners access their situation model-level representation during keyword generation increases (Thiede et al. 2010). Consequently, delayed keyword generation provides learners with access to diagnostic cues, whereas immediate keyword generation does not.

The keyword effect has been confirmed in several studies and among different levels of education (e.g., De Bruin et al. 2011; Shiu and Chen 2013; Thiede et al. 2003, 2005, 2017; for similar results using delayed summaries or concept maps see Thiede and Anderson 2003; Van De Pol et al. 2019). Nevertheless, there are still some open issues concerning the effects of generating delayed keywords. First, conceptual replications without the involvement of the researcher who originally proposed the keyword effect (i.e., K. W. Thiede) are still scarce. In view of the replication crisis in psychological research, such *external* replications would substantially strengthen the keyword effect. External replications can be perceived as an indispensable component for making convincing and reliable statements about psychological effects and the resulting theoretical and practical implications (Francis 2012; Makel et al. 2012; Open Science Collaboration 2015; Simons 2014).

Second, to the best of our knowledge, so far studies on the keyword effect have mainly investigated effects on *relative accuracy* but did not consider effects on *bias* (i.e., a measure of absolute accuracy, see e.g., Schraw 2009). Hence, it is unclear whether generating keywords reduces the degree to which learners over- or underconfidently judge their level of comprehension.

Why generating keywords might increase relative accuracy but scarcely decrease bias

Judgment accuracy differs in terms of whether it is calculated *relative* or *absolute* (e.g., Schraw 2009; see also Dunlosky and Thiede 2013). On the one hand, relative accuracy refers to the degree to which learners can evaluate the differential level of knowledge for some learning materials (e.g., texts) versus others and thus giving information about a within-subjects measure of the relationship between several comprehension judgments and the correctness of each response (e.g., De Bruin et al. 2017; Maki et al. 2005; Schraw 2009). Hence, learners with high relative accuracy can accurately rank texts in terms of their level of comprehension. Relative accuracy is mostly calculated by correlation coefficients (e.g., Goodman-Kruskal gamma correlation), whereby values close to zero indicate poor accuracy and values up to one indicate higher accuracy. At a higher level, relative accuracy is often contrasted with absolute accuracy, which is the degree to which the magnitude of a comprehension judgment matches

the magnitude of the actual performance on a criterion task, thus giving information on the difference between the target and the actual score, whereby values close to zero indicate high accuracy and values above indicate poorer accuracy (Maki et al. 2005; Nietfeld et al. 2005). Absolute accuracy is typically separated into measures of *calibration* and *bias*. The former is a measure that is derived from calibration indices reflecting how close a calibration curve mimics the curve of perfect calibration (see Keren 1991), whereas the latter is a participant's average judgment minus the average actual performance. It is important to note that both kinds of accuracy, relative and absolute, are conceptually as well as statistically largely independent of one another. This, however, can lead to the fact that, in one and the same situation, absolute accuracy might be high, while relative accuracy might be low, or vice versa (e.g., Koriat et al. 2002; for a review see Dunlosky et al. 2016; for a discussion on methodological and theoretical differences between relative and absolute accuracy see Dunlosky and Thiede 2013).

The above-mentioned studies on the keyword effect have all focused on relative accuracy measures. Thus, generating keywords after a delay enhances the degree to which learners can discriminate between the differential comprehension for the texts they have read. It is an open question, however, whether generating keywords impacts the degree to which learners judge their level of comprehension accurate in absolute terms. From a practical perspective, however, insight into the effects on absolute accuracy would be highly relevant as well. While high relative accuracy supports learners in deciding which texts they should prioritize concerning the investment of additional effort (i.e., a facet of regulation; Thiede et al. 2003), it does not entail guidance concerning the amount of effort that needs to be invested (another facet of regulation). In view of the finding that learners often dramatically overestimate their text comprehension (e.g., Dunlosky and Lipko 2007; Glenberg and Epstein 1987; Lin and Zabrocky 1998; Maki 1998; Miesner and Maki 2007; Zhao and Linderholm 2008), it could thus be the case that learners who have generated delayed keywords can rank texts in terms of their level of knowledge but would still substantially overestimate their absolute level of knowledge concerning the texts, which would result in suboptimal regulation afterwards.

From a theoretical view, effects of keyword generation on absolute accuracy—e.g., on bias, which is widely considered one of the main absolute accuracy measures (see Schraw 2009)—should occur only to a limited extent. As mentioned above, one mechanism via which generating keywords after a delay is assumed to enhance relative accuracy is that it requires learners to access their situation model-level representation of the texts, which provides them with diagnostic cues that they can incorporate into their JOLs. More specifically, one concrete process via which learners could gather diagnostic information that helps them discriminate their level of comprehension for different texts is assessing the fluency with which they can generate the required keywords. Texts concerning which learners can generate keywords with high fluency can be considered to be better understood than texts concerning which learners cannot fluently generate keywords. Another potential process could be that learners evaluate the quality of their keywords. The higher learners evaluate the quality of their keywords to be, the higher learners might judge their level of comprehension.

It is reasonable to assume that both processes help learners discriminate their level of comprehension of different texts and hence foster relative accuracy. Learners can compare the fluency of keyword generation or the (subjective) keyword quality regarding different texts and rank the texts accordingly. In terms of reducing bias (i.e., the degree of over- or underconfidence), by contrast, it is mainly evaluating the keyword quality which should be useful, but not necessarily the evaluation of the fluency of the keyword generation. Decreases in bias presuppose that learners are able to judge their level of comprehension in *absolute*

terms. The fluency with which learners generate keywords arguably is hardly informative in this respect—the fluency of keyword generation and the absolute level of comprehension are substantially different variables which should be hard to reconcile. By contrast, the keyword quality should at least to some extent serve as a diagnostic cue for learners' absolute level of comprehension. Keywords that fully capture the main content of the text should correspond with “full” comprehension, keywords that capture only two thirds of the main content of the text should indicate that only two third of the main content is understood, and keywords that do not relate to the main content of the text show that learners have not understood the respective content at all (see Dunlosky and Rawson 2012).

In view of these considerations, at first glance it could be argued that requiring learners to generate keywords after a delay should beneficially affect both relative accuracy and bias. However, a crucial prerequisite for the effect on bias to occur is that learners can accurately determine the quality of their generated keywords. Yet in the absence of external standards, learners usually fail in accurately judging the quality of their generative products (e.g., Dunlosky et al. 2011; Roelle and Renkl 2020; Zamary et al. 2016). It follows that learners should scarcely be able to accurately judge the quality of their keywords. Consequently, it can be hypothesized that the keyword effect should mainly be pronounced regarding relative accuracy and hardly at all in terms of bias reduction. A potential remedy for this blind spot of requiring learners to generate keywords could be providing learners with expert keywords that serve as external standards after they have generated keywords on their own.

How the provision of expert keywords could improve the keyword effect

Previous research has shown that learners are often not able to accurately judge the quality of their generative products (e.g., Dunlosky et al. 2011; Roelle and Renkl 2020; Zamary et al. 2016). Hence, a potential restriction of requiring learners to generate keywords could be that learners cannot exploit the full potential of the cues that would in principle be accessible to them; specifically, the quality of the generated keywords, which should be a fruitful cue in reducing bias in particular, remain largely unexploited. A promising remedy for this restriction could be providing learners with external standards. For instance, in a study by Rawson and Dunlosky (2007), learners studied key terms from Introductory Psychology, recalled these terms from memory, and afterwards were provided with correct answers (i.e., external standards) before judging their level of comprehension. In comparison to learners who were not provided with external standards, learners who received external standards showed significantly lower bias (for similar results see also Dunlosky et al. 2011; Lipko et al. 2009). Concerning the keyword effect, these findings suggest that providing learners with expert keywords after they have generated keywords on their own could enhance the metacognitive benefits of keyword generation. Comparing their own keywords with expert keywords could help learners to evaluate whether they have captured the main content of the texts and are thus accurately judging the keyword quality. In terms of reducing bias, exploiting this cue should optimize the keyword effect.

It is important to note, however, that the potential benefits of expert keywords should not be limited to a reduction of bias but can be expected to enhance the benefits of relative accuracy as well. Learners can compare the quality of their keywords concerning the respective texts, which should support them in accurately discriminating their level of comprehension of the texts. Hence, the provision of expert keywords as external standards after learners have

generated keywords on their own should be reflected in both a reduction of bias and an increase in relative accuracy. The underlying mechanism should be that the alignment between the actual quality of learners' keywords and their JOLs should increase when learners can compare their keywords with expert ones (i.e., the correlation between these two measures should be enhanced).

The present study

In view of the outlined theoretical considerations and empirical gaps, in the present study we pursued three main goals. First, we were interested in conceptually replicating the keyword effect. For this purpose, we tested the benefits of the established keyword generation procedure in a new sample and with new learning materials. We assumed that requiring learners to generate keywords after a delay would increase relative accuracy in comparison to no keyword generation (*replication hypothesis*).

Second, we were interested in whether the generation of delayed keywords would reduce bias. Based on the notion that learners should have difficulties in accurately judging the quality of their keywords, we assumed that the keyword effect should hardly be pronounced regarding bias (*keyword-generation-does-not-reduce-bias-hypothesis*).

Third, we were interested in whether providing learners with external standards in the form of expert keywords would optimize the keyword effect in terms of both relative accuracy and bias (*optimization-via-external-standards-hypothesis*). We expected that this potential optimization would be accompanied by an increase in the correlation between the actual keyword quality and learners' JOLs (*increased-correlation-between-keyword-quality-and-JOLs-hypothesis*).

Method

Sample and design

The sample comprised $N = 109$ students who were enrolled in different courses at a German university. Due to non-compliance with the formal instructions, one student was excluded from the study so that the final sample comprised $N = 108$ students (76 females; $M_{Age} = 24.16$ years, $SD_{Age} = 3.59$ years). The voluntary participation was compensated with 10 € and written informed consent was collected from all students.

The study followed an experimental between-subject design. All participants had to read four expository texts and either (a) generate keywords after a delay (*keyword condition*), (b) generate keywords and compare their keywords with expert keywords (*keyword + external standard condition*), or (c) not generate keywords (*control condition*). The participants were randomly assigned to the three conditions.

After reading the texts, the participants in the keyword condition as well as in the keyword + external standard condition were prompted with one text title after the other and were instructed to generate five keywords per text that captured the essence of the text. The participants were instructed as follows: "Now that you have read all of the texts, please generate five keywords for each of the texts which capture their central contents." During this time the control group had a break. After the two keyword conditions had generated keywords,

the participants of all three conditions were asked to evaluate their comprehension. Afterwards, the participants in the keyword+external standard condition were prompted to compare their own keywords with expert keywords. Again, one text title after the other was presented and the participants were shown their generated keywords and the expert keywords simultaneously. The participants were instructed as follows: “Now that you have generated five keywords for all the texts, we would like to present you five keywords for each text that capture their central contents according to expert ratings. Please read these expert keywords attentively. Thoroughly compare the expert keywords with the keywords you generated yourself. Assess how many of your self-generated keywords overlap with the expert keywords (identical content of the keywords or very similar keywords). The order of the keywords does not play a role in this matter. If, for example, your second self-generated keyword has the same content as the third expert keyword, please consider this as a match, even if the keywords stand at different positions.” During this time the keyword condition as well as the control group had a break. After the keyword+external standard condition had compared their keywords, all participants were told to evaluate their comprehension again.

Materials and measures

Expository texts We used expository texts from the domain of physics (thunderstorm formation), medieval history (the crusades), educational psychology (academic self-concept), and biology (neuronal stress response), whereby texts had a length ranging between 855 and 1271 words. Varying topics were chosen in order to minimize differences in interest and prior knowledge.

Expert keywords We chose five expert keywords for each of the four texts. Firstly, all keywords were generated by one expert who had very high-level knowledge concerning the respective texts. Therefore, each text was divided into five sections and the keywords were defined as an umbrella term for each of these five sections. Next, these keywords were discussed with two more experts and partially adapted up to the point at which *all* experts agreed that the main content of each text was covered in an all-encompassing manner.¹ Based on these expert keywords, the posttest was designed. Each of the five posttest questions concerning each text were closely aligned to one of the expert keywords. By this means, it was assured that the posttest assessed the main content of each of the texts (see below; for all posttest questions and corresponding expert keywords see [Supplemental Materials](#)).

Control variables: Prior knowledge, interest, and ease of learning To test whether the random assignment resulted in comparable groups, we assessed prior knowledge, interest, and ease of learning judgement as control variables. Specifically, the participants were asked to estimate their level of knowledge (e.g., “about the topic *thunderstorm formation*, I already know ...”), their interest (e.g., “the topic *neuronal stress response* interests me ...”), and their ease of learning (e.g., “I consider my ability to learn new things about the topic *academic self-concept* to be ...”) on a scale from 1 to 5, with a 1 representing “very low” and a 5 representing “very high” for the prior knowledge and the ease of learning ratings, and a 1 representing “very few” and a 5 representing “very much” for the interest ratings respectively. We used self-

¹ Even though all experts had a very high level of knowledge of all texts, it should be noted that the possibility cannot be entirely ruled out that other experts would have set different keywords for those texts.

reported prior knowledge and refrained from implementing a pretest in order to not focus the students on certain content items during the initial reading phase.

JOLs The participants judged their text comprehension two times on a scale from 1 to 5, with a 1 representing “the text was not understood at all” and a 5 representing “the text was very well understood”. The two measurement time points of the text comprehension judgments were: JOL immediately after generating keywords or having a break (JOL_{T1}), and JOL immediately after comparing the keywords or having a break, respectively (JOL_{T2}).

Posttest In order to assess learning outcomes, we developed a posttest that included five inference questions (single choice format with four options to respond) for each of the four texts (i.e., 20 questions in total; see [Supplemental Materials](#)). In line with Thiede et al. (2005), the questions were explicitly designed such that they required learners to access their situational model-level representations of the texts. Furthermore, as mentioned above the questions were closely aligned with the expert keywords. Answers were scored as correct (1) or incorrect (0). The final performance score was calculated by adding up the number of correctly answered items, so that a maximum of 20 points and a minimum of zero points could be reached. The test showed sufficient internal consistency (Cronbach’s $\alpha = .61$).

Judgment accuracy In order to assess judgment accuracy at both measurement time points, we used students’ JOLs after generating keywords or having a break (judgment accuracy at measurement time point one) and after comparing keywords or having a break (judgment accuracy at measurement time point two) and their posttest performances to form two measures. First, in line with prior research on the keyword effect (e.g., De Bruin et al. 2011; Thiede et al. 2003, 2005), intra-individual *gamma correlations* between students’ performance and their JOLs were calculated as a measure for relative accuracy—which is the degree to which a person’s JOL discriminates between texts that they will perform better on vs. worse on during a performance test. Specifically, we computed Goodman-Kruskal’s gamma correlations (Goodman and Kruskal 1954), whereby values in a range between minus one and close to zero indicate a poor relation of posttest performance and JOLs and values up to one indicate a better relation (i.e., JOLs increasing from text one to text two, three, and four, and the test performance increasing similarly across the same order of the four texts).²

Second, we calculated *bias*—which is the degree to which learners over- or underconfidently judge their performance in absolute terms—across the results of all performance questions and all JOLs.³ More specifically, for each participant we calculated a bias score by subtracting the *z*-standardized performance scores from the *z*-standardized JOLs. The participants were considered to judge their performance accurate when their JOLs were as high as could be expected based on their performance scores (i.e., equal *z*-scores on both measures; numerically said = 0). When the JOLs were higher than could be expected on the basis of their performance scores (higher *z*-score on the text comprehension rating than on the performance test; numerically said >0), the participants were considered to be too optimistic/ overconfident; when the JOLs were lower than could be expected on the basis of their performance scores

² Relative accuracy was separately calculated for both measurement time points of the JOLs.

³ Bias was separately calculated for both measurement time points of the JOLs.

(lower z -score on the text comprehension rating than on the performance test; numerically said <0), the participants were considered to be too cautious/ underconfident.⁴

Relation-of-keyword-quality-and-JOLs measure In order to analyse if the relation between the actual keyword quality and participants' JOLs increases after the external standards were set, we computed intra-individual Goodman-Kruskal's gamma correlation between students' keyword quality and their JOLs for both measurement time points, whereby values close to zero indicate a poor relation of keyword quality and JOLs and values up to one indicate a good relation.

The keyword quality was assessed by comparing the generated keywords with the expert keywords, giving two points for an identical keyword, one point for a similar keyword, and zero points for keywords that did not at all relate to the expert keywords. For example, if the expert keywords for the thunderstorm text were "lightning", "electric fields", "cloud formation", "pre- and main-discharge" and "thunder", and a participant came up with "cloud(s)" as one keyword, it was rated with one point because it was similar or at least to some extent related to the respective expert keyword. However, if a participant came up with a keyword like "ice crystals" or "warmth", which was completely unrelated to the expert keywords, it was rated with zero points. Two independent raters scored the quality of the generated keywords per text for each participant. The two raters achieved very good interrater reliability, as determined by the intraclass coefficient with measures of absolute agreement, for all keywords per text (all ICCs $> .85$).

Procedure

The participants took part in the experiment in group sessions in which each participant worked individually on a personal computer. First, the participants signed an informed consent form, gave general information on their gender and age, and answered the questions on prior knowledge, interest, and ease of learning. Second, all participants were instructed that they would have to read texts and would have to answer test questions on each text afterwards. Furthermore, in order to not induce differential reading goals, in line with the procedure of Thiede et al. (2003) all participants were informed that they might be asked to generate keywords that are meant to capture the main content of each of the texts. For a better understanding of the following procedure, the students got an example of keywords for a text on the Titanic (identical to Thiede et al. 2003; e.g., for a text of the Titanic one might write "iceberg", "shipwreck" etc.), they first read an introductory example text with a length of 850 words, and answered two sample questions (single choice format with four options, in line with the posttest questions), before finally reading the four actual texts successively (see [Supplemental Material](#) for the example text and questions). After reading the texts, the participants of the keyword and the keyword+external standard condition were asked to generate a list of five keywords for each of the texts, whereas the control group had a six-minute break with no specific instruction. This design was chosen to ensure equal delays in

⁴ Importantly, it should be highlighted that our bias measure differed from the established and recommended bias measure (e.g., Keren 1991) in that we did not require the learners to provide their judgments on the same scale as the performance was measured. Rather, in our study the scales for posttest performance (initial scale: 0–5 for each topic) and judgments of comprehension (initial scale: 1–5 for each topic) were only made comparable by z -standardization. We critically discuss this measure in the Discussion section.

time between reading the texts and the comprehension judgments in all three conditions. Moreover, we decided to use no filler task during this break for the control group since we did not intend to prevent them from executing the processes they would spontaneously execute in this setting. Afterwards all groups were asked to judge their comprehension with respect to the four text topics (JOL_{T1}). Third, the students in the keyword+external standard condition were asked to compare their own keywords with expert keywords, whereas the keyword condition and the control group again had a short break (of four minutes) with no specific instruction or filler task respectively before *all* participants had the opportunity to revise their first comprehension judgments (JOL_{T2}). Finally, all participants took the posttest.

All texts were rated for text comprehension and tested in the same order as they were presented for reading. However, to control for potential position effects, the order of the four expository texts and the order of the respective subsequent measures were varied within each condition (24 different orders within each condition). All items/questions had to be answered completely and students had as much time as they needed for the whole procedure (except for the breaks that were timed). An overview of the procedure is provided in Table 1.

Results

Table 2 gives an overview of the descriptive statistics for the three groups on all measures. We used an α -level of .05 for all analyses and partial η^2 as the effect size measure. Following Cohen (1988), values of $< .06$ indicate a small effect, values in the range between .06 and .14 indicate a medium effect, and values $> .14$ indicate a large effect.

Preliminary analyses

Control variables In the first step, we tested whether the random assignment resulted in comparable groups. We did not find any statistically significant differences between the groups regarding ease of learning judgment, $F(2, 104) = 0.79, p = .458, \eta_p^2 = .02$, prior knowledge, $F(2, 104) = 0.02, p = .982, \eta_p^2 = .00$, reading time $F(2, 104) = 1.51, p = .226, \eta_p^2 = .02$, or interest $F(2, 104) = 2.50, p = .087, \eta_p^2 = .05$, which indicated that the randomization had resulted in comparable groups.

We also compared the students regarding their posttest score and overall processing time. In this regard, we did not find a statistically significant difference between the groups regarding posttest performance, $F(2, 105) = 2.57, p = .082, \eta_p^2 = .05$, or processing time, $F(2, 105) = 1.51, p = .226, \eta_p^2 = .03$. Hence, generating keywords or generating and comparing keywords had no substantial learning effect and processing time was not different in the three groups.

Replication hypothesis

To address whether the keyword effect regarding relative accuracy could be conceptually replicated in our study, we conducted a contrast analysis. Contrast analysis is recommended to test hypotheses in experimental designs in the American Psychological Association Guidelines for the use of statistical methods in psychology journals (Wilkinson et al. 1999). A major strength of contrast analysis is that it provides the most direct and efficient way to evaluate specific predictions (Furr and Rosenthal 2003). We contrasted the keyword groups to the

Table 1 Overview of the Procedure

Condition	<i>n</i>	Reading texts	Generating keywords/ 6 min break	Judgments of learning T1	Comparing keywords/ 4 min break	Judgments of learning T2	Performance Test
Keyword+external standard group	36	Q R1 ... R4	K1 ... K4	JOL1 _{T1} ... JOL4 _{T1}	C1 ... C4	JOL1 _{T2} ... JOL4 _{T2}	T1 ... T4
Keyword group	36	Q R1 ... R4	K1 ... K4	JOL1 _{T1} ... JOL4 _{T1}	B	JOL1 _{T2} ... JOL4 _{T2}	T1 ... T4
Control group	36	Q R1 ... R4	B	JOL1 _{T1} ... JOL4 _{T1}	B	JOL1 _{T2} ... JOL4 _{T2}	T1 ... T4

Note. Q = filling out all questionnaires that were used to assess covariates; R1 = Participants read text 1...; JOL1 = Participants judged their comprehension of text 1...; T1 = Participants took 5 test questions on text 1...; K1 = Participants generated a list of five keywords for text 1...; C1 = Participants compared their generated keywords to expert-keywords; B = Break

Table 2 Means and (Standard Deviations) of all Measures

	Keyword+external standard group (<i>n</i> = 36)	Keyword group (<i>n</i> = 36)	Control group (<i>n</i> = 36)
Ease of learning judgment	3.28 (0.61)	3.31 (0.64)	3.15 (0.52)
Prior knowledge	3.47 (0.57)	3.47 (0.74)	3.49 (0.65)
Interest	3.05 (0.56)	2.73 (0.61)	2.89 (0.61)
Reading time (in min.)	29.62 (11.13)	29.94 (7.48)	29.60 (7.59)
Posttest	10.56 (3.03)	11.31 (3.14)	9.76 (3.21)
Judgment of learning T1	12.17 (2.15)	12.75 (2.51)	11.44 (2.44)
Judgment of learning T2	11.88 (2.52)	12.58 (2.71)	11.47 (2.60)
Relative accuracy T1	0.43 (0.75)	0.43 (0.63)	0.11 (0.65)
Relative accuracy T2	0.50 (0.67)	0.43 (0.63)	0.12 (0.63)
Bias T1	0.03 (1.00)	0.04 (0.98)	-0.03 (1.27)
Bias T2	-0.15 (1.01)	0.07 (0.99)	0.12 (1.29)
Processing time (in min.)	46.84 (15.21)	52.28 (15.63)	47.26 (13.34)

control group regarding their relative accuracy that was determined on the basis of the JOLs at the first measurement time point and the posttest score (at this time, the keyword groups were still identical to each other; contrast weights assigned to the experimental conditions: 1 for the keyword condition, 1 for the keyword+external standard condition, -2 for the control group). We hypothesized that requiring learners to generate keywords after a delay would increase relative accuracy in comparison to no keyword generation. As can be seen in Fig. 1 (left), we found that the keyword groups reached higher relative accuracy (i.e., ranked their comprehension more accurately) than the control group, $t(93) = 2.06$, $p = .021$ (one-tailed), $\eta_p^2 = .04$.⁵

Keywords-generation-does-not-reduce-bias-hypothesis

To address whether generating keywords could also reduce bias, we contrasted the keyword groups to the control group using the bias that was determined on the basis of the JOLs at the first measurement time point and the posttest score (at this time, the keyword groups were still identical to each other; contrast weights assigned to the experimental conditions: 1 for the keyword condition, 1 for the keyword+external standard condition, -2 for the control group). We hypothesized that the keyword effect should hardly be pronounced regarding bias. As can be seen in Fig. 1 (right), no significant effect was found for bias, $t(105) = 0.29$, $p = .384$ (one-tailed), $\eta_p^2 = .00$.

Optimization-via-external-standards-hypothesis

To address whether external standards in the form of expert keywords could optimize the keyword effect, we computed planned contrasts with students' judgment accuracy (relative accuracy and bias) as the dependent variable (within-subject-factor: first vs. second measurement time point), and condition as the between-subject-factor (contrast weights assigned to the experimental conditions: -1 for the keyword condition, 1 for keyword+external standard condition, 0 for control group). We hypothesized that providing learners with external standards in the form of expert keywords would optimize the keyword effect in terms of both relative accuracy and bias. Indeed, the contrast test revealed a statistically significant effect

⁵ $N = 12$ of the gamma correlations were undetermined.

Fig. 1 Judgment accuracy for both measures at the first comprehension judgments: relative accuracy (left), bias (right). * $p < .05$.

regarding both accuracy measures, $t(91) = 1.69$, $p = .047$ (one-tailed), $\eta_p^2 = .04^6$ for the relative accuracy and $t(105) = 2.45$, $p = .008$ (one-tailed), $\eta_p^2 = .05$ for bias. Fig. 2 illustrates the results. As expected, relative accuracy of the students who additionally compared their keywords increased from the first to the second measurement time point, whereas it remained unchanged in the keyword group. However, contrary to our expectations, for the students who additionally compared their keywords with expert ones the bias (here: level of underconfidence) increased from the first to the second measurement time point, whereas it remained unchanged in the keyword group.

Increased-correlation-between-keyword-quality-and-JOLs-hypothesis

We assumed that the metacognitive benefits of the keyword comparison would be reflected in an increase of the correlation between the JOLs and the keyword quality from the first to the second measurement time point. To test this, we computed planned contrasts with students' intraindividual gamma correlation between their keyword quality and the JOLs at the first and the second measurement time point as the dependent variable (within-subject-factor: first vs. second measurement time point) and condition as the between-subject-factor (contrast weights assigned to the experimental conditions: 1 for keyword+external standard condition, -1 for keyword condition). As expected, intraindividual correlations between participants' keyword quality and JOLs increased statistically significantly from the first to the second measurement time point for the keyword+comparison group but did not increase for the keyword group, $t(63) = 2.24$, $p = .014$ (one-tailed), $\eta_p^2 = .07^7$ (see Fig. 3).

⁶ $N = 14$ of the gamma correlations were undetermined.

⁷ $N = 5$ of the gamma correlations were undetermined.

Fig. 2 Development of judgment accuracy across both measurement times: relative accuracy (left), bias (right)

Discussion

Our study entails four main contributions concerning the effects of generating delayed keywords. First, our results substantiate the notion that generating delayed keywords is an effective means to enhance relative accuracy. Specifically, using a new sample as well as new learning materials that had not been used in a keyword study before, we found similar results as previous keyword studies concerning relative accuracy (*replication-hypothesis*). Second, our study indicates that generating keywords does not reduce bias (*keyword-generation-does-not-reduce-bias-hypothesis*). Hence, the keyword effect appears to be limited to relative accuracy.⁸ Third, providing learners with external standards in the form of expert keywords and prompting them to compare those keywords with their own is a promising means to optimize the keyword effect. At least in terms of relative accuracy, the learners who generated keywords benefitted from subsequent comparison with expert ones (*optimization-via-external-standards-hypothesis*). The fourth main contribution relates to the underlying mechanism of this optimization regarding relative accuracy. The increased correlation between the JOLs and the keyword quality after learners had compared their own keywords with the expert ones indicates that the optimization was driven by the learners exploiting the potential of using the keyword quality as cue (*increased-correlation-between-keyword-quality-and-JOLs-hypothesis*). Conversely, this indicates that learners without expert keywords do not optimally use the quality of their keywords as a cue for their JOLs.

Benefits and limitations of generating delayed keywords

The current study is one of the few that confirms the keyword effect for relative accuracy beyond the pioneering work of K. T. Thiede (but see Shiu and Chen 2013). In line with our *replication hypothesis*, we successfully replicated and thereby contributed to validating that generating delayed keywords is a promising activity to enhance the accuracy with which learners can discriminate between their differential comprehension for certain texts compared

⁸ This result speaks in favor for the proposition that future research on metacomprehension would benefit from more studies that consider measures of both relative and absolute accuracy (for a similar suggestion see also Dunlosky and Thiede 2013).

3

Fig. 3 Correlation between keyword quality and comprehension judgments across both measurement times.

to others (see also e.g., De Bruin et al. 2011; Thiede et al. 2003, 2005, 2017). In view of the above-mentioned replication crisis in psychological research (e.g., Makel et al. 2012; Open Science Collaboration 2015), this replicational finding of the present study can be conceived of as a valuable contribution to strengthening the keyword effect. Besides, going beyond replication, our study is the first on the keyword effect that controlled for potential time-on-task differences between the keyword and no-keyword group by adding a break for the no-keyword group that corresponds with the time of keyword generation. To the best of our knowledge, all above-mentioned studies on the keyword effect had the no-keyword control group judge their comprehension of the texts immediately after reading (or at least did not report on a taken break; but see Thiede et al. 2005 for a discussion on delayed vs. immediate judgments). However, the fact that we did not give the learners any other task during this break may of course have led them to think about possible keywords or to evaluate the text without being asked to do so. As we still found the keyword-effect in this setting, however, it is reasonable to assume that the potential “hidden” keyword-generation on part of the learners in the no-keyword group was relatively rare.

From a practical view, the keyword effect implies that for learners who need support in accurately judging which text they have understood less well than others and therefore should prioritize concerning the investment of additional effort (i.e., regulation), generating delayed

keywords can serve as a beneficial activity. However, in terms of bias our findings also point to an important limitation of generating keywords and thereby also support the above-mentioned notion that relative accuracy and absolute accuracy are conceptually largely independent of one another (see Dunlosky and Thiede 2013; in the present study, the correlation between relative accuracy and bias for the overall sample ranged between $r = .03$ and $r = .11$, both $p > .05$; see also the separate correlations for all three groups and all measurement time points in the [Supplemental Materials](#)). Specifically, our results indicate that, while generating delayed keywords increases relative accuracy, it does not decrease bias (*keyword-generation-does-not-reduce-bias-hypothesis*). Admittedly, however, this finding is hard to interpret. At first glance, it seems to fit our assumption that the keyword effect is less pronounced for bias than for relative accuracy. Accordingly, one explanation for this finding could be that the cues that became accessible through keyword generation (e.g., the fluency of keyword generation and/ or the quality of keywords) were not or merely scarcely reconcilable with the absolute level of comprehension. An alternative explanation, however, could be that in the present study the bias was surprisingly low—even the learners in the control group were able to accurately judge their level of comprehension in absolute terms (see Table 2). Consequently, a comparatively profane explanation for the missing effect could simply lie in the fact that there was hardly any room for improvement concerning bias. Yet, we cannot say for sure whether this low bias was simply due to chance or whether and to what extent both the learning material (i.e., the texts) and the posttest questions that were used contributed to this surprising finding. Future studies certainly are needed to clarify this pattern of results.

A further aspect that needs to be considered with respect to the findings on bias is that the bias measure used in the present study most likely was suboptimal. More specifically, we did not ask the learners to estimate the number of points they could reach on the posttest (i.e., on a scale from 0 to 20 points or on a scale from 0 to 100%) but rather asked them to judge their level of comprehension on a scale from one to five (see Method section). Thus, contrary to the recommendation of Keren (1991; see also Schraw, 2009) the JOLs made by the participants were not measured on the same scale as the performance test. Therefore, this measure (i.e., the JOLs) is hard to reconcile with learners' posttest performance so that it seems worthwhile for future research to let the participants predict their performance on a percentage scale (or let them predict the number of points they could reach on the posttest) to compute a more valid bias measure. However, as the posttest was designed to assess comprehension and the JOLs tapped comprehension as well (see Method section), learners' *z*-standardized JOLs and posttest scores should at least to some extent be reconcilable and thus the bias measures might be valid at least to some extent. To test this assumption, in a small follow-up study with $N = 92$ university students, we had learners provide both judgments of comprehension (similar to the measure used in the present study) and judgments of performance (i.e., a prediction of the number of correctly answered posttest questions). We found a correlation of $r = .75$, $p < .001$ between these two measures. Hence, although this correlation does not imply that both measures yield identical results, it at least shows that the measures closely relate to each other and thus that the bias measure used in the present study is valid to some extent.

Nevertheless, for the above-mentioned reason, the bias measure used in the present study certainly is not flawless; hence, our findings regarding bias need to be interpreted very cautiously, since we cannot state for sure which aspect of judgment accuracy it actually taps.

The use of expert keywords for optimizing the keyword effect

A novel finding of the present study that does not suffer from the potentially suboptimal bias measure is that providing learners with expert keywords that are well aligned with the posttest after learners had generated keywords on their own optimized the keyword effect. More specifically, providing learners with external standards in the form of expert keywords contributed to a significant increase in relative accuracy. One explanation for this finding, which is supported by the finding that the correlation between keyword quality and learners' JOLs increased through the comparison with expert keywords (*increased-correlation-between-keyword-quality-and-JOLs-hypothesis*), is that the expert keywords helped the learners to accurately evaluate the quality of their own generated keywords. This, in turn, supported them in exploiting the potential of this predictive cue in making their JOLs (for similar results of external standards see e.g., Baars et al. 2014; Rawson and Dunlosky 2007; Dunlosky et al. 2011).

However, contrary to our *optimization-via-external-standards-hypothesis*, this optimization did not apply regarding bias. On the contrary, comparing their keywords with expert ones increased bias compared to generating keywords only. As comparing the keywords with expert keywords contributed to underconfidence, it could tentatively be argued that it would have resulted in beneficial effects had the learners in the present study shown the *usual* overconfident bias (see e.g., Dunlosky and Lipko 2007; Lin and Zabracky 1998; Miesner and Maki 2007; Zhao and Linderholm 2008). However, future studies (that also circumvent the above-mentioned criticism regarding the bias measure used in the present study) are certainly needed to test this tentative prediction.

In these future studies, it could furthermore be worthwhile to differentiate different types of bias that are affected by external standards. Specifically, expert keywords might have different effects for students who have performed very well and are sure they have performed well compared to those who have not performed well and are sure they have not performed well (both being indicators of low bias) and furthermore compared to students who tend to over- or underconfidently judge their comprehension. In other words, and in line with a large amount of instructional research in general, providing expert keywords as a means of support to reduce bias might possibly not be equally beneficial under all circumstances.

Limitations and future research

In addition to the above-mentioned unresolved issues, the present study has some important limitations that should be addressed in further research. First, we did not investigate whether generating and comparing keywords not only enhances judgment accuracy but also promotes learning via regulation. Although this link between enhanced judgment accuracy, regulation, and learning is intuitively plausible and assumed by theories of self-regulated learning (e.g., Nelson and Narens 1994; Zimmerman 2008) as well as by empirical studies (e.g., De Bruin et al. 2011; Thiede et al. 2003; Rawson et al. 2011), we did not focus on this aspect in the present study. Future studies should thoroughly analyse the effects of generating and comparing keywords concerning regulation and achievement.

A second aspect that needs to be critically reflected upon relates to the design of expert keywords that we chose as external standards. Specifically, one crucial ingredient of our optimization regarding relative accuracy could be that the expert keywords were closely aligned with the posttest questions. Baars et al. (2014) point out the problem of external

standards that are very similar to what has been learned (or what is going to be tested): Although such external standards are well able to improve reflection and predict performance on the respective learning content, their benefits are limited to just this content. Hence, future studies should investigate the extent to which the benefits of expert keywords concerning relative accuracy depend on the degree of alignment between the expert keywords and the posttest.

Finally, it could be fruitful to more deeply investigate *why* and *how* the cues produced by generating (and comparing) keywords are effective. Although the present study at least to some extent illuminates learners' cue use—for it suggests that learners who only generate keywords scarcely exploit the potential of keyword quality as a cue, whereas learners who also receive expert keywords at least in part exploit this cue—it is still widely unclear which cues learners use to what extent to form their JOLs after they have generated keywords. Future studies that delve into the process of forming JOLs more deeply (e.g., by using think-aloud methodology) should address this open issue (see also Van De Pol et al. 2019). In this regard, it seems also interesting and worthwhile for future studies to incorporate a control condition that does not generate keywords but sees expert keywords (for a similar study see also Thiede et al. 2005, Exp. 4). Also, it could be worthwhile to investigate whether learners benefit from being informed about the purpose of generating keywords. Informed keyword generation might enhance the keyword effect because learners might be more attentive to the cues that become available during keyword generation. However, it might also decrease the keyword effect because it might enhance the degree to which learners consciously interpret and weigh the respective cues when forming their JOLs, which might overload them.

Acknowledgements We thank the students who participated in the study. We furthermore thank the research assistant Anna-Lena Kwelik for her support during conducting the experiment and analyzing the data and Marc Hafner for programming the software. Open Access funding provided by Projekt DEAL.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Baars, M., Vink, S., Van Gog, T., De Bruin, A. B. H., & Paas, F. (2014). Effects of training self-assessment and using assessment standards on retrospective and prospective monitoring of problem solving. *Learning and Instruction*, 33, 92–107. <https://doi.org/10.1016/j.learninstruc.2014.04.004>.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Lawrence Erlbaum Associates, Publishers.

- De Bruin, A. B. H., Thiede, K. W., Camp, G., & Redford, J. (2011). Generating keywords improves metacomprehension and self-regulation in elementary and middle school children. *Journal of Experimental Child Psychology*, 109, 294–310. <https://doi.org/10.1016/j.jecp.2011.02.005>.
- De Bruin, A. B. H., Dunlosky, J., & Cavalcanti, R. B. (2017). Monitoring and regulation of learning in medical education: The need for predictive cues. *Medical Education*, 51, 575–584. <https://doi.org/10.1111/medu.13267>.
- Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: A brief history and how to improve its accuracy. *Current Directions in Psychological Science*, 16, 228–232. <https://doi.org/10.1111/j.1467-8721.2007.00509.x>.
- Dunlosky, J., & Rawson, K. A. (2012). Overconfidence produces underachievement: Inaccurate self evaluations undermine students' learning and retention. *Learning and Instruction*, 22, 271–280. <https://doi.org/10.1016/j.learninstruc.2011.08.003>.
- Dunlosky, J., & Thiede, K. W. (2013). Four cornerstones of calibration research: Why understanding students' judgments can improve their achievement. *Learning and Instruction*, 24, 58–61. <https://doi.org/10.1016/j.learninstruc.2012.05.002>.
- Dunlosky, J., Hartwig, M. K., Rawson, K. A., & Lipko, A. R. (2011). Improving college students' evaluation of text learning using idea-unit standards. *The Quarterly Journal of Experimental Psychology*, 64, 467–484. <https://doi.org/10.1080/17470218.2010.502239>.
- Dunlosky, J., Mueller, M. L., & Thiede, K. W. (2016). Methodology for investigating human metamemory: Problems and pitfalls. In J. Dunlosky & S. K. Tauber (Eds.), *Oxford library of psychology. The Oxford handbook of metamemory* (p. 23–37). Oxford University Press.
- Engelen, J., Camp, G., Van de Pol, J., & De Bruin, A. D. H. (2018). Teachers' monitoring of students' text comprehension: Can students' keywords and summaries improve teachers' judgment accuracy? *Metacognition and Learning*, 13, 287–307. <https://doi.org/10.1007/s11409-018-9187-4>.
- Francis, G. (2012). The psychology of replication and replication in psychology. *Perspectives on Psychological Science*, 7, 585–594. <https://doi.org/10.1177/1745691612459520>.
- Furr, R. M., & Rosenthal, R. (2003). Evaluating theories efficiently: The nuts and bolts of contrast analysis. *Understanding Statistics*, 2, 45–67. https://doi.org/10.1207/S15328031US0201_03.
- Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition*, 15, 84–93. <https://doi.org/10.3758/BF03197714>.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 732–764.
- Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, 36, 93–103. <https://doi.org/10.3758/MC.36.1.93>.
- Keren, G. (1991). Calibration and probability judgements: Conceptual and methodological issues. *Acta Psychologica*, 77, 217–273. [https://doi.org/10.1016/0001-6918\(91\)90036-Y](https://doi.org/10.1016/0001-6918(91)90036-Y).
- Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review*, 95, 163–182.
- Kintsch, W., Welsh, D., Schmalhofer, F., & Zimmy, S. (1990). Sentence memory: A theoretical analysis. *Journal of Memory and Language*, 29, 133–159.
- Koriat, A., Sheffer, L., & Ma'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General*, 131, 147–162. <https://doi.org/10.1037/0096-3445.131.2.147>.
- Lin, L. M., & Zabrucky, K. M. (1998). Calibration of comprehension: Research and implications for education and instruction. *Contemporary Educational Psychology*, 23, 345–391.
- Lipko, A. R., Dunlosky, J., Hartwig, M. K., Rawson, K. A., Swan, K., & Cook, D. (2009). Using standards to improve middle school students' accuracy at evaluating the quality of their recall. *Journal of Experimental Psychology: Applied*, 15, 307–318. <https://doi.org/10.1037/a0017599>.
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, 7, 537–542. <https://doi.org/10.1177/1745691612460688>.
- Maki, R. H. (1998). Predicting performance on text: Delayed versus immediate predictions and tests. *Memory & Cognition*, 26, 959–964. <https://doi.org/10.3758/bf03201176>.
- Maki, R. H., Shields, M., Wheeler, A. E., & Zacchilli, T. L. (2005). Individual differences in absolute and relative metacomprehension accuracy. *Journal of Educational Psychology*, 97, 723–731. <https://doi.org/10.1037/0022-0663.97.4.723>.
- Miesner, M. T., & Maki, R. H. (2007). The role of test anxiety in absolute and relative metacomprehension accuracy. *European Journal of Cognitive Psychology*, 19, 650–670. <https://doi.org/10.1080/09541440701326196>.

- Nelson, T. O., & Narens, L. (1994). Why investigate metacognition. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge: MIT Press.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. <https://doi.org/10.1126/science.aac4716>.
- Rawson, K. A., & Dunlosky, J. (2007). Improving students' self-evaluation of learning for key concepts in textbook materials. *European Journal of Cognitive Psychology*, 19, 559–579. <https://doi.org/10.1080/09541440701326022>.
- Rawson, K. A., O'Neil, R., & Dunlosky, J. (2011). Accurate monitoring leads to effective control and greater learning of patient education materials. *Journal of Experimental Psychology: Applied*, 17, 288–302. <https://doi.org/10.1037/a0024749>.
- Roelle, J., & Renkl, A. (2020). Does an option to review instructional explanations enhance example-based learning? It depends on learners' academic self-concept. *Journal of Educational Psychology*, 112, 131–147. <https://doi.org/10.1037/edu0000365>.
- Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning*, 4, 33–45. <https://doi.org/10.1007/s11409-008-9031-3>.
- Shiu, L.-p., & Chen, Q. (2013). Self and external monitoring of reading comprehension. *Journal of Educational Psychology*, 105, 78–88. <https://doi.org/10.1037/a0029378>.
- Simons, D. (2014). The value of direct replication. *Perspectives on Psychological Science*, 9, 76–78. <https://doi.org/10.1177/1745691613514755>.
- Thiede, K. W., & Anderson, M. C. M. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology*, 28, 129–160. [https://doi.org/10.1016/S0361-476X\(02\)00011-5](https://doi.org/10.1016/S0361-476X(02)00011-5).
- Thiede, K. W., Anderson, M. C. M., & Theriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, 95, 66–73. <https://doi.org/10.1037/0022-0663.95.1.66>.
- Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1267–1280. <https://doi.org/10.1037/0278-7393.31.6.1267>.
- Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*, 47, 331–362. <https://doi.org/10.1080/01638530902959927>.
- Thiede, K. W., Redford, J. S., Wiley, J., & Griffin, T. D. (2017). How restudy decisions affect overall comprehension for seventh-grade students. *British Journal of Educational Psychology*, 590–605. doi: <https://doi.org/10.1111/bjep.12166>.
- Van De Pol, J., De Bruin, A. B. H., Van Loon, A. H., & Van Gog, T. (2019). Students' and teachers' monitoring and regulation of students' text comprehension: Effects of metacomprehension cue availability. *Contemporary Educational Psychology*, 56, 236–249. <https://doi.org/10.1016/j.cedpsych.2019.02.001>.
- Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology*, 132, 408–428. <https://doi.org/10.3200/GENP.132.4>.
- Wilkinson, L., the Task Force on Statistical Inference, & APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604. <https://doi.org/10.1037/0003-066X.54.8.594>.
- Zamary, A., Rawson, K. A., & Dunlosky, J. (2016). How accurately can students evaluate the quality of self-generated examples of declarative concepts? Not well, and feedback does not help. *Learning and Instruction*, 46, 12–20. <https://doi.org/10.1016/j.learninstruc.2016.08.002>.
- Zhao, Q., & Linderholm, T. (2008). Adult metacomprehension: Judgment processes and accuracy constraints. *Educational Psychology Review*, 20, 191–206.
- Zimmerman, B. J. (2008). Investigating self-regulation and motivation: Historical background, methodological developments, and future prospects. *American Educational Research Journal*, 45, 166–183. <https://doi.org/10.3102/0002831207312909>.