University of Southampton

Faculty of Environmental and Life Sciences

School of Psychology

A Systematic Review and Investigation of Judgment of Learning (JoL) Reactivity.

by

Lloyd Chilcott

Thesis for the degree of Doctorate of Educational Psychology

June 2022

University of Southampton

Abstract

Faculty of Environmental and Life Sciences

School of Psychology

Thesis for the Degree of Doctorate of Educational Psychology

A Systematic Review and Investigation of Judgment of Learning (JoL) Reactivity.

by

Lloyd Chilcott

Judgments of learning (JoLs) are predictions about the likelihood of recalling learnt material. JoLs have been a standard self-report tool in memory research for over 50 years, but recent research has observed that JoLs can affect memory in and of themselves: an effect termed *JoL reactivity*. JoL reactivity is typically observed in word pair experiments, in which participants who give a JoL for related word pairs (e.g., *dog-cat*) recall more targets than participants without a JoL. Thus, JOLs appear to improve recall for related word pairs. However, despite this finding, little is understood about when or why JoL reactivity occurs. Subsequently, this thesis provides an investigation into JoL reactivity across two papers.

The first paper provides a systematic review of the JoL reactivity literature. JoL reactivity research has grown rapidly since the last systematic review, but with contradictions in the literature: some papers report positive reactivity (improved performance), others negative reactivity (impaired performance) and some no reactivity. In addition, contrasting theoretical frameworks have been put forward to explain the mechanisms that result in JoL reactivity. The systematic literature review assesses the evidence and theoretical accounts of JoL reactivity. We observed that word pair relatedness appears to moderate the reactive effect and that there is a growing consensus that JoLs produce positive reactivity with semantically related word pairs. We also observed that relational accounts of reactivity are most common in the literature but have inconsistent evidence. There are emerging non-relational accounts, but these are tentative frameworks. Future areas for research are suggested.

The second paper investigates JoL reactivity in a *transfer appropriate processing* (TAP) paradigm. In an initial encoding phase, we presented participants with related, rhyming, or unrelated word pairs to induce different levels of processing. Half of the participants made a JoL after studying each word pair, while the remaining participants simply studied each word pair for an equivalent duration. Afterwards, all participants completed either standard or rhyme recognition tests. We successfully replicated the TAP effect. In the rhyme recognition test, the participants successfully recognised more rhymes of targets from the rhyming pairs than the related and unrelated pairs. However, no significant evidence of JoL reactivity was seen, regardless of encoding or test condition. The study is the first to investigate JoL reactivity using a TAP paradigm with word pairs and provides a foundation for future work to examine the role of the test on JoL reactivity and JoL reactivity in alternative paradigms.

Table of Contents

Table of Tables

Table of Figures

Research Thesis: Declaration of Authorship

Print name: LLOYD CHILCOTT

Title of thesis: A Systematic Review and Investigation of Judgment of Learning (JoL) Reactivity.

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Signature: .... Date: ……………………..

Acknowledgements

This project would not have been possible without the support of many people. Thank you to my lead supervisor, Tina Seabrooke, who went over and above to provide support and guidance. You somehow made sense of my confusion. Thank you to my second supervisor, Philip Higham, for your advice and insight. And thank you to my voluntary research assistant, Rebecca Everitt, for your support with the systematic review. You generously gave up your time to navigate the weird and wonderful world of JoLs.

Thank you to the tutors and trainees of the 2019-2022 Educational Psychology Doctorate at the University of Southampton. In particular, thank you to my personal tutor, Cora Sargeant, for always providing the right support at the right moment, and to Jamie Wilson, for your friendship and support throughout the highs and lows.

Thank you to my fiancé Bethany for your love and encouragement throughout the doctorate. And lastly, thank you to my dog Doug, whose walks and emotional support kept me sane. You are the bestest boy.

Definitions and Abbreviations

ANOVA	Analysis of variance
Cohen's *d*	Cohen's d effect size
df	Degrees of Freedom
F	*F test statistic*
JBI	Joanna Briggs Institute
LoP	Levels of Processing
M	Mean
n	Total number of participants
OSF	Open Science Framework
p	Significance
SD	Standard deviation
t	t-test statistic
TAP	Transfer Appropriate Processing
α	Cronbach's alpha
η_g^2	Generalised eta squared

Chapter 1 Introduction

1.1 Context for Research

Cognitive and educational psychology has informed the creation and review of simple classroom-based practices, tools and recommendations (Dunlosky et al., 2013a). Everyday classroom strategies, such as rereading and writing summaries, are analysed for their potential benefits to education (Dunlosky et al., 2013b). One increasingly explored example is testing (Schwieren et al., 2017). Summative tests are commonly used in the classroom to assess students' performance and academic achievement. However, rather than being solely a form of assessment, tests can induce *the testing effect,* the phenomenon that learning activities designed as a test benefit memory (Trumbo et al., 2021). The testing effect has been demonstrated in laboratory studies (Karpicke & Roediger, 2007b) and, crucially, with educational materials (Karpicke & Blunt, 2011) and in the classroom (McDaniel et al., 2007; see Trumbo et al., 2021, for a review). Consequently, low-stakes, formative testing is increasingly encouraged as an evidence-based strategy to aid learning (Dunlosky et al., 2013b; Schwieren et al., 2017)

Cognitive and educational psychology has also provided evidence against using some traditional teaching practices: for example, the act of underlining to-be-remembered items. Underlining is a popular learning technique (Peterson, 1991). It is quick and simple and, in theory, creates a distinction towards a to-be-remembered item, benefiting item retention of particularly important information (Dunlosky et al., 2013a). However, underlining rarely isolates essential information (often due to excessive underlining: Lorch et al., 1995) and, empirically, provides no memory benefit over simply reading (Marxen, 1996; see Dunlosky et al., 2013a, for a review). Furthermore, underlining is not a neutral act; in some cases, it can harm learning. Peterson (1991), for example, observed that participants who used underlining focused on single items and failed to create connections between items. Similarly, using underlining comes at an opportunity cost. Time spent underlining is not time spent on a more beneficial strategy, such as testing. Underlining serves as an example of why a learning technique cannot be assumed to be beneficial nor assumed to do no harm.

Research plays a vital role in informing educational practice, as seen with the testing effect and underlining. Therefore, it is vital to turn the research lens towards new practices to understand how they operate, their risks, and their potential before they reach the classroom. One such practice is judgements of learning (JoLs). JoLs are predictions about the likelihood of recalling learnt material (Koriat, 1997). A researcher may elicit a JoL by presenting a word and asking for the likelihood a participant will recall the word in a future memory test, often on a scale (e.g., from 0, not at all confident, to 100, extremely confident; Rhodes, 2016). JoLs have been a standard self-report tool in memory research for over 50 years (Arbuckle & Cuddy, 1969; Rhodes, 2016), but there are increasing suggestions that JoLs could become more than a metamemory measure and have potential as a learning strategy (Janes et al., 2018; Li et al., 2021; Soderstrom et al., 2015). These suggestions were first based on studies demonstrating that JoLs influence how participants use their study time (Bjork et al., 2013), leading to adaptive study decisions, such as studying for longer on items that are judged to be more difficult and less well learnt (Soderstrom & Bjork, 2014). More recently, JoLs have also been found to affect memory in and of themselves: an effect termed *JoL reactivity* (Double et al., 2018; Rhodes & Tauber, 2011).

Reactivity (or *reactive effects*) refers to an intentional or unintentional change in behaviour or performance in response to a measurement or observation (Double & Birney, 2019). JoL reactivity is typically observed in a word pair associate research paradigm (e.g., Mitchum et al., 2016; Soderstrom et al., 2015). In this paradigm, researchers present participants with semantically related (e.g., *blunt-sharp*) and unrelated (e.g., *juice-leap*) cue-target word pairs. Half of the participants study each word pair before providing a JoL, and half simply study each word pair (the total study time is the same). After a distraction task, all participants complete a cued-recall test (e.g., blunt-???). Recent research has observed that participants from the JoL group recall more targets from strongly related pairs than the no-JoL group (Janes et al., 2018; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015). However, this *positive reactivity effect* does not typically extend to the weakly and unrelated word pairs. Thus, JOLs appear to improve cued-recall only for word pairs with a strong pre-existing semantic association.

The increasing number of studies reporting positive reactivity has led numerous authors to highlight the potential of JoLs to improve education (Janes et al., 2018; Li et al., 2021; Soderstrom et al., 2015). We note, however, that there is insufficient research into JoL reactivity in the classroom. Almost all published research using JoLs in an educational setting and with educational materials do not examine if JoLs benefit learning (e.g., Baars et al., 2018; Roelle et al., 2017); only a single paper has explored JoL reactivity with educational materials. In the paper, Ariel et al. (2021) asked participants to read five paragraphs, each roughly 100 words, and provide an aggregate JoL of the text after each paragraph. The authors did not observe reactivity in the follow-up comprehension questions. However, in a follow-up experiment, the authors did observe reactivity when the participants were presented with a recall opportunity (two or three short questions about the text) before the JoL. The authors concluded that JoL reactivity does not readily occur with their complex educational materials, but there is potential for JoLs to aid learning in carefully constructed learning scenarios. Thus, not only is there minimal research on

JoL reactivity in education or with educational materials, but the sole study concludes with caution and caveats.

Research into the real-world application of JoLs will be vital for future applications outside of the laboratory. However, there is still much to learn about JoL reactivity, and it may be premature to talk about JoLs in the context of education. Most JoL reactivity research draws upon a single paradigm (paired-associates: e.g., Janes et al., 2018; Myers et al., 2020; Rivers et al., 2021; Tauber & Witherby, 2019). There is no research into how JoL reactivity is impacted by participant characteristics other than age (Tauber & Witherby, 2019; Zhao et al., 2021), and only a single study has explored the impact of different types of tests (Myers et al., 2020). Myers et al. (2020) argued that JoL reactivity research is underdeveloped compared to other memory phenomena, and more research is required to understand when reactivity occurs. While JoLs for education is an exciting prospect, JoL reactivity research is in its infancy, and there is a need to develop a foundational understanding before exploring JoLs in the classroom.

1.2 Overview of the Present Research

The present research provides an investigation into JoL reactivity across two papers. Here, I provide a brief overview of both papers, including the rational, method and results.

The first paper, presented in chapter two, provides a systematic review of the JoL reactivity literature. JoL reactivity research has grown rapidly since the last systematic review (Double et al., 2018), but with contradictions in the research; different studies have reported positive reactivity (improved performance: Myers et al., 2020; Rivers et al., 2021; Senkova & Otani, 2021; Tekin & Roediger, 2020), negative reactivity (impaired performance: DeYoung & Serra, 2021) and no reactivity (Dougherty et al., 2018; Robey et al., 2017). In addition, contrasting theoretical frameworks have been put forward to explain the mechanisms that result in JoL reactivity (e.g., the cue-strengthening hypothesis versus the changed-goal hypothesis: Mitchum et al., 2016; Soderstrom et al., 2015). The systematic literature review assesses the evidence and theoretical accounts of JoL reactivity to address these issues. We observe that word pair relatedness appears to moderate the reactive effect, and there is a growing consensus that JoLs produce positive reactivity with semantically related word pairs. We also observed that *relational* accounts of reactivity (theories that focus on the relationship between stimuli) are most common in the literature but have inconsistent evidence. There are emerging *non-relational* accounts, but these are tentative frameworks. We conclude by addressing the impact of JoL reactivity research on future JoL research, metacognition research and the potential for JoLs in education.

The second paper, presented in chapter three, investigates JoL reactivity in a *transfer appropriate processing* (TAP) paradigm. Many authors attribute JoL reactivity to JoLs providing elaborate processing for related word pairs (Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015). This is an equivalent line of reasoning to what has been suggested to underpin the *levels of processing* (LoP) effect: tasks that foster deep, elaborate processing should produce longer-lasting retention than tasks that encourage shallow processing (Craik & Lockhart, 1972). Morris et al. (1977) highlighted that memory is not solely about the depth of encoding but also how memory is tested (the TAP effect). Subsequently, we sought to investigate JoL reactivity in a TAP paradigm (Morris et al., 1977) with word pair associates. In an initial encoding phase, we presented the participants with related, rhyming, or unrelated word pairs to induce different LoP. Half of the participants made a JoL after studying each word pair, while the remaining participants simply studied each word pair for an equivalent duration. Afterwards, all participants completed either a standard and rhyme old/new recognition test of the target words (or rhymes of the targets) from the encoding phase. The targets from the related pairs were recognised most often on the standard recognition test, while rhymes of targets from rhyming pairs were best recognised on the rhyme recognition test. Thus, we observed a clear TAP effect. However, we did not observe significant evidence of JoL reactivity, regardless of encoding or test condition. The study is the first to investigate JoL reactivity using a TAP paradigm with word pair associates and provides a foundation for future work to examine the role of the test on JoL reactivity and JoL reactivity in alternative paradigms. Together, the two papers move forward the JoL reactivity literature by synthesising existing findings and contributing a novel insight into JoL reactivity.

1.3 Researcher's Background and Rationale for Engagement

I take a *critical realist* philosophical research position. Critical realism is a philosophy that draws upon *post-positivism*, a position that takes a mostly *positivist* ontology (there is a material reality independent of human minds; Bhaskar et al., 2017) but with a *constructionist* critique (Bhaskar, 2008). Namely, that knowledge is tentative and influenced by bias and perspective (Creswell, 2009). As a researcher, I value critical realism as it attempts to find the balance between pursuing a well-defined shared reality while incorporating the influence of values and political intent into the philosophical underpinnings of research (Salomon, 1991). I believe this is particularly important in educational research, as it is often enormously challenging to simplify and control the many variables in education, and we must understand different perspectives to provide the best education for all (Scotland, 2012).

Despite my critical realist position, much of the present research reflects a positivist approach. The review and my empirical research explore experimental data to contribute to a single, shared understanding of JoL reactivity. However, I do not believe this approach contradicts my critical realist position: JoLs are inherently about eliciting the participants' perspectives; I acted upon my potential bias and included a second reviewer (see 2.3.4 Assessing eligibility); I applied a critical realist lens to my discussion sections, such as my critique of the insufficient details on participant characteristics in the JoL literature (see 2.7.3 Impact on education); and I undertook and designed the research because of what I believe is valuable for education. As such, the present research embodies critical realism by committing to positivism while being informed by constructionism.

Before researching JoL reactivity, I was a primary school teacher who observed and practised many learning techniques with my pupils. Some learning tools could make a positive difference and create a more accessible classroom, whereas poorly considered resources and techniques resulted in lost time, money, and opportunities. Subsequently, once I learnt of the authors discussing JoLs in education, I wanted to help develop the literature (be it to help the journey towards JoLs as an evidence-informed strategy or push back against its application in the classroom). As discussed previously, my search of the literature highlighted that JoL reactivity research is in its infancy, and while I could have decided upon an applied approach (such as bringing JoLs into the classroom) I felt that the next steps for the literature required building upon the existing paradigms to further develop the foundational understanding of JoL reactivity. There are still many pertinent, unanswered questions about JoL reactivity which are best explored with methods that systematically develop upon the existing research. Such basic experimental research will further researchers’ understanding of the mechanisms that drive JoL reactivity, which future research can then build upon in a more applied direction. Hence, the present research serves as a pre-requisite to applied research that aims to bring JoLs closer to the classroom, should the prior basic research suggest it is responsible to do so.

1.4 Ethical Challenges

I recruited most of my participants using Prolific, an online participant recruitment platform (www.prolific.co/), due to not being able to test in the laboratory following the Covid-19 social restrictions. Consequently, I had to decide how much to pay each participant for their contribution. While deciding, I was made aware that other experiments in the same department used the platform's minimum payment of £5/hour. I appreciated this payment was to maximise limited resources and to allow for further research. However, at the time, the UK minimum wage was £6.15 to £8.21 (depending on age). I was aware that paying £5/hour would negatively impact my participants and would fail to meet their entitled income (even if I was not obliged to pay the higher rate). I believe my research has a political function in and of itself (as informed by critical realism), so I decided to pay £7.50/hour to respect my participants and contribute to a culture of increased equality. My decision could limit the availability for future research, but I believe that should not come at the cost of the very participants that make our research possible.

I used R, a statistical programme and programming language, for my analysis. I was introduced to R at the start of the present research and encouraged to consider it instead of alternative statistical programmes. While R has a steep learning curve, it became apparent that R was consistent with my research values. Using R contributes to the financial sustainability of the publicly-funded higher education system. R is a free, open-source programming platform compared to costly, licensed alternatives. Similarly, R emboldens open science. The R analysis script for my research is available on the Open Science Framework (https://osf.io/ak57u/), allowing anyone to run my analysis, see my decision making, and check the outcomes. It is not possible to easily document the same decision making in non-programming alternatives. In sum, R elevated my research. However, R was often beyond my capabilities and was considerably more time-consuming than a more accessible alternative. R has many benefits, but I hope future researchers in a similar inexperienced position will receive more significant support to use R, enabling them to apply their values consistently throughout their practice.

Lastly, I did not exclude outliers from my analysis. Osborne and Overbay (2004) argue that outliers are an inevitable part of research, are often an illegitimate value (e.g., a mistyped number), and may produce overestimated and underestimated outcomes (Kwak & Kim, 2017). However, Welles (2014) argued that omitting outliers potentially discriminates against the representation of all participants in the pursuit of a desired outcome. These two viewpoints represent contrary epistemological positions. Ultimately, I decided to keep the outliers as I could not discern between potential illegitimate values and those deriving from anomalous, legitimate behaviour. For example, my results contained some participants with a zero JoL rating for every word pair, but I did not have the means of knowing *why* they scored accordingly. They may not have engaged with the test, or they may have believed zero reflected their confidence. It is possible I was wrong to keep the outliers (assuming they were illegitimate), and I should have pre-registered my experiment with a defined outlier exclusion criteria (e.g., the exclusion of outliers 3 *SD* above and below the sample average). However, by making my data openly available, and documenting my decision, I invite future researchers to use the present experiment to make an informed decision about outliers in a similar method.

Chapter 2 A Systematic Literature Review of Reactivity to Judgments of Learning

Abstract

A rapidly growing literature base has explored the reactive effects of Judgments of learning (JoLs). Research increasingly shows that JoLs, a metamemory measure, produce reactivity effects in paired-associate and word-list learning paradigms. However, there are contradictions in the literature, with reports of positive reactivity (improved performance), negative reactivity (impaired performance) and no reactivity. In addition, contrasting theoretical frameworks have been put forward to explain the mechanisms that result in JoL reactivity. We report a systematic literature review assessing the evidence and theoretical accounts of JoL reactivity. A search of Scopus, PsychInfo and ProQuest databases was conducted in February 2022. We included studies with an adult neuro-typical and non-clinical population, using immediate JoLs with a no-JoL control group, with an outcome on a recognised cognitive task. We included peer-review and thesis papers and excluded non-English papers. All the studies included for analysis were assessed using the JBI Critical Appraisal Checklist to assess the methodological quality and risk of bias. The themes and findings of the systematic search were collated and presented in a narratively driven summary. Eighteen papers containing 44 experiments and 4891 participants were included in the final analysis. We observed that word pair relatedness appears to moderate the reactive effect and that there is a growing consensus that JoLs produce positive reactivity with semantically related word pairs. We also observed that *relational* accounts of reactivity (the cue-strengthening and changed-goal hypotheses) are most common in the literature but have inconsistent evidence. There are emerging *non-relational* accounts (item-specific processing and elaborate processing hypotheses), but these are tentative frameworks. The review was limited by the absence of a meta-analysis to provide further insight, and the discussion of the theoretical frameworks of JoL reactivity was limited by reference to the studies that were selected for inclusion in accordance with the systematic review. We conclude with the implications of our review on future metacognitive measures and a discussion on the prospect of JoLs for education.

2.1 Introduction

Metamemory is the awareness and understanding of one’s memory content and processes (Schwartz & Metcalfe, 2017). For example, a person who believes they are good at remembering faces, but not names, is reflecting on their metamemory knowledge. Metamemory beliefs affect restudy decisions (Son & Metcalfe, 2000), error monitoring (Yeung & Summerfield, 2012) and strategy selection (Çubukcu, 2008). Metamemory is typically measured using self-report measures that are either prospective (i.e., predicting future performance) or retrospective (i.e., reflecting on the accuracy of past responses; Schwartz & Metcalfe, 2016). Retrospective measures, such as confidence judgements, tend to be more accurate (a similar judgment and performance score) than prospective measures, such as judgments of forgetting (Siedlecka et al., 2016). However, prospective judgments can provide information earlier in the learning process, thereby allowing for the judgment to potentially affect behaviour (Rhodes, 2016).

Judgments of learning (JoLs) have been a common prospective metamemory measure since their inception by Arbuckle and Cuddy (1969) to help investigate how people learn (Rhodes, 2016). Taking JoLs involves asking participants to assess their confidence in their ability to recall recently learnt material (Koriat, 1997). JoLs can be taken for a range of study materials (e.g., word lists, Tekin & Roediger, 2020; or educational texts, Ariel et al., 2021) but are most frequently applied to word pairs (Double et al., 2018). For example, a participant may study the word pair *fun-happy* before being asked to judge the likelihood of recalling the target word (*happy*) in a subsequent cued memory performance test. JoLs are typically recorded using a scale, such as from 0 (not at all confident) to 100 (extremely confident; Rhodes, 2016). JoLs can be made immediately after learning or made following a delay. Delayed JoLs are shown to have greater accuracy than immediate JoLs (Rhodes & Tauber, 2011), an effect dubbed the *delayed JoL effect* (Nelson & Dunlosky, 1991).

Recent research has observed *JoL reactivity*, namely, that JoLs can intentionally or unintentionally affect memory (Double et al., 2018; Rhodes & Tauber, 2011; Soderstrom et al., 2015). Soderstrom et al. (2015) provided the first demonstration of JoL reactivity in a series of experiments that directly compared an immediate JoL group to a no-JoL control group. In an initial study phase, all participants studied word pairs that were either semantically related (e.g., *blunt-sharp*), weakly related (e.g., *boxer-terrible*) or unrelated (e.g., *flag-sack*). The no-JoL control group studied each word pair for 8 seconds, while the JoL group participants studied each word pair for 4 seconds and gave a JoL for 4 more seconds to match the study duration of the control group. All participants then engaged in a 3 minute distraction task before completing a cued-recall test (e.g., *flag-???*). There was no significant difference in recall performance score in the cued-recall test between the JoL and no-JoL groups for the weakly related and unrelated word pairs. In contrast, the JoL group recalled significantly more pairs than the no-JoL group for the strongly related pairs.

Soderstrom et al. therefore concluded that immediate JoLs improve subsequent cued-recall (dubbed *positive reactivity*) for word pairs with a strong semantic association.

While some subsequent studies have also observed JoL reactivity with related word pairs (Janes et al., 2018; Myers et al., 2020; Rivers et al., 2021; Tauber & Witherby, 2019), other experiments have reported conflicting results. Studies that used a similar paired-associated (word pair) paradigm as Soderstrom et al. (2015) have observed no significant reactivity for related pairs (Myers et al., 2020: experiment 2), positive reactivity for unrelated word pairs (Tauber & Witherby, 2019: experiments 3-5), no reactivity for related word pairs and *negative reactivity* (reduced performance) for unrelated pairs (Mitchum et al., 2016). Thus, the literature presents a mixed picture with respect to the effects of administering JoLs on memory.

Double et al. (2018) provided the first systematic and meta-analytic review of the JoL reactivity literature. Overall, the authors found no significant reactivity effect for immediate JOLs. However, reactivity was reliably observed under certain circumstances, with moderate reactivity seen for related word pairs (e.g.*, carpet-rug, sticky-glue, blend-mix*) and lists of single words (e.g., *car, sand, jump*). When participants studied lists of either unrelated pairs (e.g., *switch-lie, wool-small, moon-card*) or mixed lists of related and unrelated pairs, JoLs did not reliably improve recall. Since Double et al.’s review, there has been an upsurge of interest in understanding the effects of collecting JoLs on memory (e.g., Dougherty et al., 2018; Janes et al., 2018; Li et al., 2021; Myers et al., 2020; Rivers et al., 2021; Senkova & Otani, 2021; Tauber & Witherby, 2019; Tekin & Roediger, 2020). Therefore, the aim of this paper is to provide an updated review of the JoL reactivity literature.

2.1.1 The Present Review

The present review provides a systematic review of the immediate JoL reactivity literature. This review was necessary due to the aforementioned upsurge in JoL reactivity research. We identified 44 relevant experiments, a considerable increase from the 17 experiments included in Double et al.’s 2018 review. The present review will synthesise the latest findings of the conditions that give rise to JoL reactive effects.

In addition, the review will provide a summary and investigation of the contradictory results that have been reported in the JoL reactivity literature. While Double et al. (2018) highlighted some of these contrasting outcomes, more recent papers have added to the mixed picture of the presence and direction of JoL reactivity, with different authors showing positive reactivity (Janes et al., 2018; Rivers et al., 2021; Senkova & Otani, 2021; Tekin & Roediger, 2020), negative reactivity (DeYoung &

Serra, 2021) and no reactivity (Dougherty et al., 2018; Robey et al., 2017). The review will explore the divergent outcomes for potential discrepancies that influence the outcome (e.g., does a specific methodological decision reliably result in one direction of reactivity?). In addition, the review will include grey literature to account for potential publication bias (a paper with significant results is more likely published than those with null results: Easterbrook et al., 1991). Grey literature often provides data not within commercially published literature, often with null or negative results, and may reduce publication bias while providing a more balanced picture of the evidence (Paez, 2017).

Lastly, the present research will provide the first systematic review of the evidence for and against the theoretical frameworks of JoL reactivity. Recent research has increasingly explored the mechanisms that produce JoL reactivity (Mitchum et al., 2016; Rivers et al., 2021; Senkova & Otani, 2021; Soderstrom et al., 2015; Tekin & Roediger, 2020). Therefore, this review will synthesise the leading theoretical frameworks and provide suggestions for future research.

2.2 Method

Bespoke systematic literature review training, provided by the University of Southampton, was completed in anticipation of the present review. The training was informed by the work of Boland et al., 2017.

2.2.1 Study Eligibility

2.2.1.1 Participant Types

The present review included studies with adult participants and excluded studies using a clinical population (e.g. adults with caffeine induced cravings, see Palmer et al., 2017) or a neurodivergent population (e.g., adults with autism, see Grainger et al., 2016). This was to reduce the confounding variables explaining JoL reactivity.

JoL reactivity studies with children (those under 18-years-old) is a newly emerging field with very few publications (Zhao et al., 2021). Consequently, the present review excluded children to better understand the conditions that give rise to JoL reactivity in the population present in the vast majority of the literature (adults). This allows for a more equitable comparison between papers by removing a significant participant characteristic (children versus adults) and replicates the decision-making in the previous JoL reactivity review (Double et al., 2018). Future reviews may look to include children as the field becomes more populous.

2.2.1.2 Study Types

We only included experiments comparing memory performance between an immediate JoL and no-JoL control condition. We also included studies with an additional metacognitive rating comparison group (e.g., a retrospective confidence judgment (RCJ) group; Robey et al., 2017), but the additional group was excluded from the summary statistics. We limited the sample to studies that examined immediate JoLs (rather than delayed JoLs) to focus our review on the recent strong interest in immediate JoLs (e.g., Double et al., 2018; Maxwell & Huff, 2022; Myers et al., 2020; Rivers et al., 2021), to contribute to the understanding of the baseline for reactivity effects (Double et al., 2018) and to facilitate an achievable review, as the inclusion of the delayed JoL reactivity literature would have exceeded the limits of the present research.

We only included studies that randomly assigned participants to the JoL and no-JoL conditions. Some experiments manipulated the JoL group as a between-subject variable, with participants completing JoLs or not (e.g., Senkova & Otani, 2021), or as a within-subject variable, with all participants giving JoLs and no-JoLs in different encoding opportunities (e.g., Myers et al., 2020, Experiment 4). Each experiment included an initial encoding phase where the participants studied the stimuli, such as a word list (e.g., *plant, tree, bird*) or a series of word pairs (e.g., *sand-sky*), before completing a final memory test (see below). Most of the experiments also included a filled retention interval (e.g., arithmetic questions) between the encoding and test phases to minimise serial order effects (the increased recall of first and last items) on the final test. Usually, the retention interval lasted just a few minutes, but reactivity effects have also been tested with longer retention intervals (see Witherby & Tauber, 2017).

2.2.1.3 Outcome Types

The included studies all assessed memory performance using a cued-recall, recognition, or free-recall criterion test. When the final test was cued-recall, the participants initially studied a cue and target together (e.g., *sun-moon*). They were then presented with the cues (e.g., *sun-???*) and asked to recall the targets (e.g., *moon*). When the final test format was a recognition test, the participants were typically required to discriminate the target words presented at encoding from intermixed new foils (e.g., “is ‘moon’ one of the target words presented earlier?”). Finally, the free-recall tests involved the participants recalling the material studied at encoding without cues or prompts.

2.2.2 Study Selection

A search of Scopus, PsychInfo and ProQuest databases was conducted on the 18th February 2022. Scopus and PsychInfo provide two of the largest indexes of psychological science and ProQuest provides the largest collection of dissertations and theses (ProQuest, 2022). We did not include further databases due to the limited scope of the review. The search included peer-reviewed studies and grey literature. Grey literature often provides data not within commercially published literature, often with null or negative results, and may reduce publication bias while providing a more balanced picture of the evidence (Paez, 2017). The ProQuest search included grey literature but not any peer-reviewed papers, while the other searches included both.

All the studies included for analysis were assessed using the JBI Critical Appraisal Checklist (Tufanaru et al., 2020) to assess the methodological quality and ensure methodological rigour (see Appendix A). The JBI Critical Appraisal Checklist provided a 13-step appraisal tool with an accessible 4-part rating system. The tool was selected over alternative quantitative checklists, such as the 27-step Downs and Black checklist (Downs & Black, 1998), as it provides a sufficient analysis of potential bias and limitations within a more concise number of questions. Consequently, the reduced number of items resulted in a more accessible appraisal checklist table (Appendix A).

Search terms for JoL ("judgements of learning" OR "judgment of Learning" OR "JOL" OR "JOLs" OR "metacognitive judgements" OR "metacognitive ratings" OR "metamemory judgements" OR "metamemory ratings") were combined with cognitive ability search terms ("reactivity" OR "cognitive performance" OR "cognitive ability" OR "memory"). Two additional studies were identified from scanning the reference lists of the final studies included for synthesis.

2.2.3 Assessing Eligibility

One reviewer screened the 654 title and abstracts of the search results for eligibility. Another reviewer independently cross-reviewed a random sample of 20 studies from the 654 available studies to inform the reliability of the eligibility decision-making. The second reviewer could not review a larger sample due to time restraints. There was an inter-rater reliability of 95% agreement (one study of the 20 cross-reviewed papers resulted in a discrepant inclusion decision). The two reviewers discussed and resolved the divergent decision before the first reviewer proceeded to the full-text eligibility search. One author was contacted to clarify the distinction between their thesis paper and a published article. The study's inclusion and exclusion procedures are presented in Figure 1.

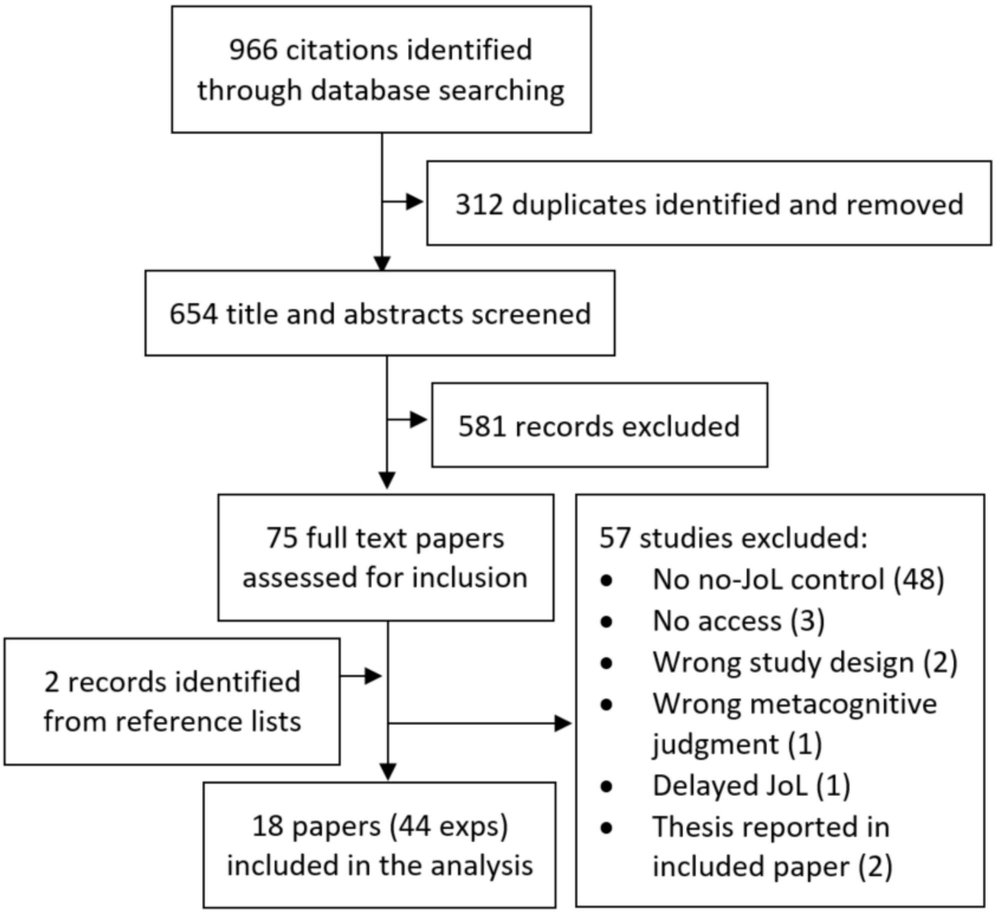
2.2.4 Analytic Approach

The analysis of the 18 final papers comprised of two processes. First, selected study details were extracted and included in the summary of study characteristics of included studies (Table 1). We selected which details to extract based on the study characteristics presented in the last immediate JoL reactivity review (Double et al., 2018) in addition to further characteristics considered of value for comparison. The extracted study characteristics were the country of study, study design (within or between), number of JoL and no-JoL participants, the type of stimuli, the word type, number of words, study time, length of retention interval, test type and the presence or direction of reactivity. From the extracted data, the present review reports descriptive statistics as presented in the results section below. The present review did not report on effect size as effect sizes were not consistently reported in the included studies, nor was a consistent effect size measure used within the literature.

The second analytic process involved identifying themes and details within the included studies and reporting a narratively driven summary. A narratively driven summary allowed the themes to be discussed and considered in relation to one another and allowed for a more accessible presentation of results. The present review incorporates these themes and details, in addition to the descriptive statistics, to discuss the factors that affect JoL reactivity and the theoretical frameworks of JoL reactivity (as presented in the discussion section, respectively).

Figure 1

Study Inclusion and Exclusion Procedure


966 citations identified through database searching
312 duplicates identified and removed
654 title and abstracts screened
581 records excluded
75 full text papers assessed for inclusion
2 records identified from reference lists
18 papers (44 exps) included in the analysis
57 studies excluded:
• No no-JoL control (48)
• No access (3)
• Wrong study design (2)
• Wrong metacognitive judgment (1)
• Delayed JoL (1)
• Thesis reported in included paper (2)

2.3 Results

2.3.1 Description of Studies

Eighteen papers containing 44 experiments and 4891 participants were included in the final analysis. Thirty-six experiments manipulated the JoL vs no-JoL conditions in a between-subjects design (i.e., one group provided JoLs and a separate group did not). The mean number of participants in the JoL and No-JoL groups was 58.97 (*SD* = 26.98) and 57.36 (*SD* = 27.97), respectively. Eight experiments used a within-subject design (i.e., a single group gave JoLs and no-JoLs in different encoding opportunities) with a mean sample size of 70.30 (*SD* = 74.21). Table 1 provides a summary of the final set of studies.

Twenty-eight experiments were based in the United States and three experiments were based in China. There were 29 laboratory experiments and nine online studies. The participants for the online experiments were recruited via Amazon Mechanical Turk, a participant recruitment platform. The authors of the remaining experiments did not state their geography or if the experiment was online.

All 18 papers were peer-reviewed, despite the inclusion criteria allowing for thesis papers. Two thesis papers met the final inclusion criteria (Mitchum, 2011; Witherby, 2016), but the experiments were reported in subsequent published papers already within the final papers for synthesis (Mitchum et al., 2016; Witherby & Tauber, 2017). We excluded both thesis papers to avoid reporting duplicate experiments.

Table 1

Summary of Study Characteristics of Included Studies

First author	Year	Exp.	Country	JoL design	JoL pps.	No-JoL pps.	Word type	Stimuli	No. words	Study time	JoL time	Retention interval	Test-type	Reactivity
DeYoung	2021	5	U.S. (web)	BS	41*	45*	RWP v WRWP	Pairs	40	5s	SP	Immediate	Cued-recall	Negative (LRWP)
Li	2021	2	China	BS	32		UWL	List	160	6s / 3s	3s	5 min	Recognition	Positive
		3	China	BS	21		RWP	Pairs	80	8s / 4s	4s	5 min	Cued-recall	Positive
Rivers	2021	1	U.S.	BS	30	30	RWP v UWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive (RWP)
		1	U.S.	WS	32		RWP v UWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive (RWP)
		2	U.S.	WS	246		RWP v UWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive (RWP)
		3	U.S.	WS	64		RWP v UWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive (RWP)
Senkova	2021	1	U.S.	BS	68	68	CWL v UWL	List	32	5s	7s	2 min	Free-recall	Positive (CWL)
		2	U.S.	BS	80	80	CWL v UWL	List	32	5s	7s	2 min	Free-recall	Positive (CWL)
Myers	2020	1	U.S.	BS	46	40	RWP v UWP	Pairs	60	12s / 5s	7s	5 min	Cue v Free-recall	Positive (RWP)
		2	U.S.	BS	48	48	RWP v UWP	Pairs	60	12s / 5s	7s	5 min	Cue v Free-recall	No
		3	U.S.	BS	60	61	RWP v UWP	Pairs	60	12s / 5s	7s	5 min	Cue v Recognition	Positive (RWP)
		4	U.S. (web)	WS	161		RWP v UWP	Pairs	60	12s / 5s	7s	5 min	Recognition	Positive (RWP)

First author	Year	Exp.	Country	JoL design	JoL pps.	No-JoL pps.	Word type	Stimuli	No. words	Study time	JoL time	Retention interval	Test-type	Reactivity
Tekin	2020	1	U.S. (web)	BS	42	42	LOP	List	60	5s	SP	10 min	Recognition	Positive
		2	U.S. (web)	BS	42	42	LOP	List	60	5s	SP	10 min	Recognition	Positive
		3	U.S. (web)	BS	42	40	UWL	List	80	7s	UC	2 days	Recognition	Positive
Tauber	2019	1	U.S.	BS	78	80	RWP	Pairs	60	10s / 5s	5s	3 min	Cued-recall	Positive
		2	U.S.	BS	80	80	RWP	Pairs	60	10s / 5s	5s	3 min	Cued-recall	Positive
		3	U.S.	BS	79	80	UWP	Pairs	60	10s / 5s	5s	3 min	Cued-recall	Positive
		4	U.S.	BS	89	90	UWP	Pairs	60	10s / 5s	5s	3 min	Cued-recall	Positive
		5	U.S.	BS	91	91	UWP	Pairs	60	5s / 2.5s	2.5s	3 min	Cued-recall	Positive
Dougherty	2018	1	UC	BS	52	51	UWP	Pairs	56	5s	UC	Immediate	Cued-recall	No
Janes	2018	1	UC	BS	35	35	RWP v UWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive (RWP)
		1	UC	BS	36	35	RWP v UWP	Pairs	60	SP	SP	3 min	Cued-recall	Positive (RWP)
		2	U.S. (web)	BS	96**	96**	RWP/UWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive
		3	UC (web)	BS	147	142	RWP/UWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive
Robey	2017	1	U.S.	BS	51	50	UWP	Pairs	56	5s	UC	Immediate	Cued-recall	No
		2	U.S.	BS	55	53	UWP	Pairs	56	5s	UC	Immediate	Cued-recall	No

First author	Year	Exp.	Country	JoL design	JoL pps.	No-JoL pps.	Word type	Stimuli	No. words	Study time	JoL time	Retention interval	Test-type	Reactivity
Witherby	2017	1	UC	BS	80	80	RWP	Pairs	60	8s / 4s	4s	3m v 2day	Cued-recall	Positive
		2	UC	BS	80	80	RWP	Pairs	60	8s / 4s	4s	3m v 2day	Cued-recall	Positive
		3	UC	BS	80	80	RWP	Pairs	60	8s / 4s	4s	3m v 2day	Cued-recall	Positive
Mitchum	2016	1	U.S.	BS	103	102	RWP v UWP	Pairs	60	SP	UC	Immediate	Cued-recall	Negative (RWP)
		2	U.S.	BS	25	25	RWP v UWP	Pairs	44	SP	UC	Immediate	Cued-recall	Negative (RWP)
		3	U.S.	BS	26	25	RWP v UWP	Pairs	48	SP	UC	Immediate	Cued-recall	Negative
		4a	U.S.	BS	28	27	UWP	Pairs	48	SP	UC	Immediate	Cued-recall	No
		4b	U.S.	BS	68	66	UWP	Pairs	48	SP	UC	Immediate	Cued-recall	No
		5	U.S.	BS	33	31	RWP v UWP	Pairs	48	5s	UC	Immediate	Cued-recall	Negative
Soderstrom	2015	1a	U.S	BS	40	20	RWP v WRWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive (RWP)
		1b	U.S. (web)	BS	30**	30**	RWP v UWP	Pairs	60	8s / 4s	4s	3 min	Cued-recall	Positive (RWP)
		2	U.S. (web)	BS	40	20	RWP	Pairs	50	8s / 4s	4s	3 min	Cued-recall	Positive
Yang	2015	1	China	WS	26		UWL	List	360	2s	UC	1 day	Recognition	Positive
Tauber	2012	5	U.S.	BS	40	40	UWL	List	30	4s	UC	3 min	Free-recall	No
Dougherty	2005	2	U.S.	BS	62	60	UWP	Pairs	52	3s v 12s	UC	Immediate	Cued-recall	Positive
Kelemen	1997	1	U.S.	WS	49		RWP/UWP	Pairs	60	8s	UC	UC	Cued-recall	No
		3	U.S.	WS	48		RWP/UWP	Pairs	60	6s	UC	UC	Cued-recall	No
Zechmeister	1980	1	U.S.	WS	24		UWL	List	24	6s	6s	1 min	Free-recall	Positive

Note. Exp = Experiment; UC = unclear; RWP = related word pair; WRWP; weakly related word pair; UWP = unrelated word pair; LOP = Levels of Processing; SP = Self-paced; UWL = unrelated word list; CWL = categorised word list. Multiple study times indicate the time for the control vs the JoL group (i.e., 6 seconds of study time for the control group and 3 seconds for the JoL group). The participant numbers with an asterisk were confirmed by the researcher in direct correspondence. The participant numbers with a double asterisk reported the total participant numbers, not the group numbers, so the total was shared between the groups. Rivers et al. (2021) experiment 1, and Janes et al. (2018) experiment 1, were represented in Table 1 as two rows to present an additional between-subject factor for the judgment condition.

2.3.2 Study Design

All 44 experiments started with an encoding phase. The researchers presented the participants with a series of word pairs (*n* = 35 experiments) or a series of individual words from a list (*n* = 9 experiments), one item at a time. Each word pair consisted of a cue and target word. For example, *run-walk* consists of the cue *run*, and the target *walk.* The participants in the no-JoL groups studied each item for an average of 7.22 seconds (*SD* = 2.59 seconds), while the JoL groups studied each item for an average of 4.23 seconds (*SD* = 0.66 seconds) before making the JoL with the cue-target pair still typically present. The participants in six experiments had self-paced, unlimited study time partly due to the researchers exploring the impact of a JoL on study decisions (Janes et al., 2018; Mitchum et al., 2016). In all experiments, the participants in the JoL groups provided a JOL immediately after studying each item. The mean time allocated to make a JoL was 4.80 seconds (*SD* = 1.36 seconds), and in four studies, the participants had self-paced time to make a JoL.

After the encoding phase, 29 experiments had all participants engage in a short retention interval (distraction period: *M* = 3.76 minutes, *SD* = 2 minutes), while four experiments used a long retention interval (2 days (*n* = 3) or 1 day (*n* = 1)). The researchers used the retention interval (e.g., arithmetic questions) to minimise serial order effects (the increased recall of first and last items) on the final test. Eleven experiments did not have a retention interval (in Figure 1, this is shown as *immediate* under *retention internal*), and two experiments did not provide the relevant details (in Figure 1, this is shown as *UC (unclear)* under *retention interval*). The authors of the studies without a retentional interval did not provide an explanation for their ommision of an retention interval (Dougherty et al., 2005, 2018; Mitchum et al., 2016; Robey et al., 2017).

In the final stage of each experiment, each participant engaged in a memory test. Thirty-one experiments used a cued-recall test. Participants observed the cue word from their learnt word pair (e.g., *run* from the earlier example) and attempted to recall the associated target word (e.g., *walk* from the earlier example). Six experiments used old/new target recognition tests, in which the participant answered “yes” if they recognised the presented word from the encoding phase or “no” if they did not. Lastly, four experiments used free-recall procedures in which participants attempted to recall the words from the encoding phase. Two experiments manipulated test type with cued-recall vs free-recall, while another experiment manipulated cued-recall vs recognition.

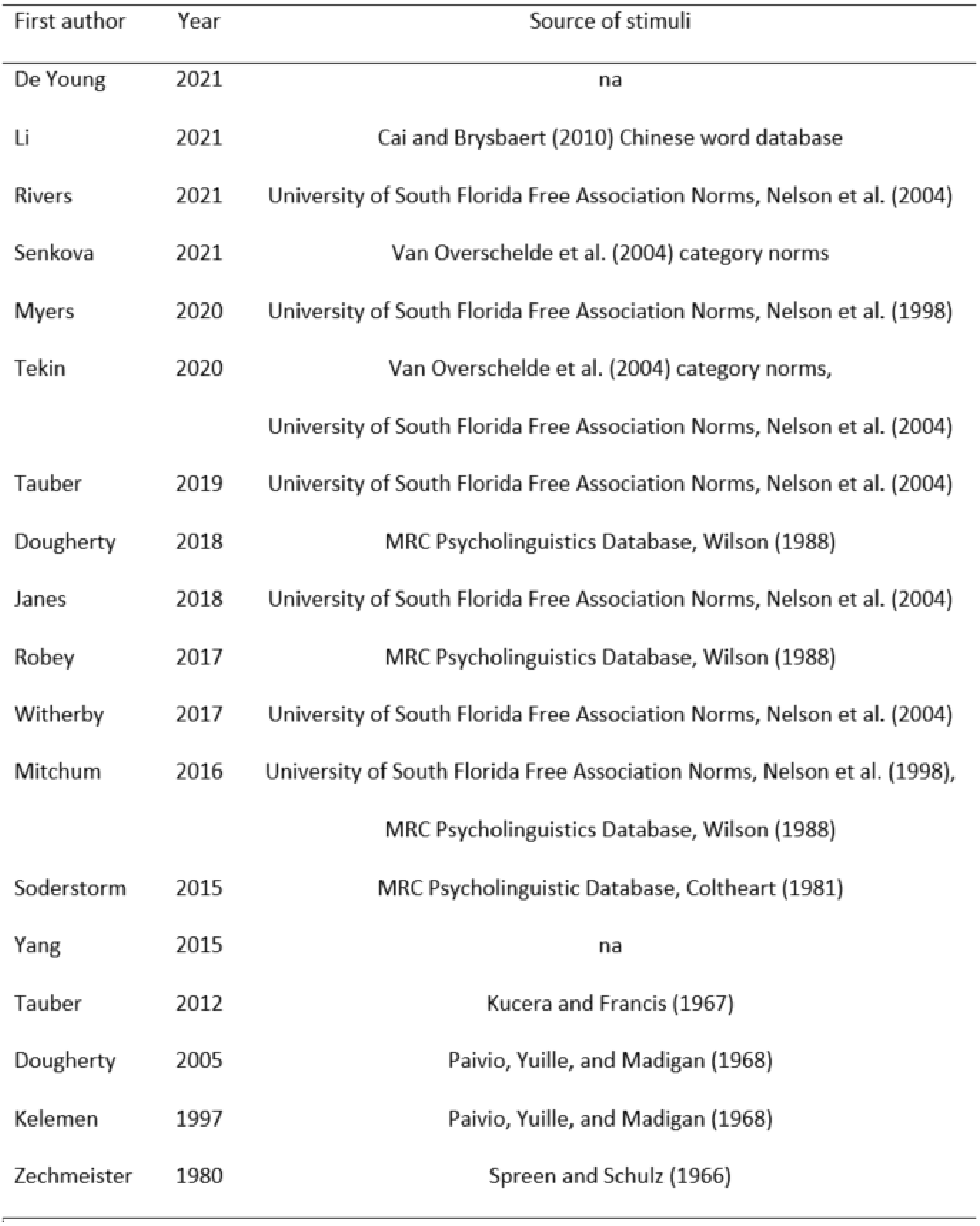
2.3.3 Stimuli

Across all experiments, the participants studied an average of 53.43 (*SD* = 5.76) word pairs, or 93.10 (*SD* = 108.57) words from a word list. Seven experiments used only related word pairs (e.g., *cat-dog*), nine experiments used only unrelated word pairs (e.g., *tree-moon*), 19 experiments used both related and unrelated word pairs, and two experiments used related and weakly related word pairs (a less semantically related pair than a related word pair but not completely unrelated; e.g., *boxer–terrible*: Soderstrom et al., 2015). Seven experiments used unrelated word lists (words with no associative strength), and two used categorised (words ordered by a shared taxonomy, e.g., rabbit, cat, dog) and unrelated word lists. In Table 1, each experiment's stimuli is recorded under *word type*.

There was considerable variation in how each paper reported the stimuli features (e.g., length, frequency), with varying levels of detail. Consequently, stimuli features were omitted from Table 1. Nonetheless, where stimuli details were reported, eight papers reported using English nouns (Dougherty et al., 2005, 2018; Kelemen & Weaver, 1997; Mitchum et al., 2016; Tauber et al., 2015; Tekin & Roediger, 2020; Zechmeister & Shaughnessy, 1980) and two papers reported using Chinese nouns (Li et al., 2021; Yang et al., 2015). Stimuli length was reported to be between 3-8 letters (Myers et al., 2020) 4-8 letters (Dougherty et al., 2018; Robey et al., 2017) and 5-8 letters (Senkova & Otani, 2021). Table 2 shows the source of the stimuli reported in each paper. Due to the variation in the included and omitted stimuli details, we do not include stimuli details in the discussion.

Table 2

Source of Stimuli



First author	Year	Source of stimuli
De Young	2021	na
Li	2021	Cai and Brysbaert (2010) Chinese word database
Rivers	2021	University of South Florida Free Association Norms, Nelson et al. (2004)
Senkova	2021	Van Overschelde et al. (2004) category norms
Myers	2020	University of South Florida Free Association Norms, Nelson et al. (1998)
Tekin	2020	Van Overschelde et al. (2004) category norms, University of South Florida Free Association Norms, Nelson et al. (2004)
Tauber	2019	University of South Florida Free Association Norms, Nelson et al. (2004)
Dougherty	2018	MRC Psycholinguistics Database, Wilson (1988)
Janes	2018	University of South Florida Free Association Norms, Nelson et al. (2004)
Robey	2017	MRC Psycholinguistics Database, Wilson (1988)
Witherby	2017	University of South Florida Free Association Norms, Nelson et al. (2004)
Mitchum	2016	University of South Florida Free Association Norms, Nelson et al. (1998), MRC Psycholinguistics Database, Wilson (1988)
Soderstorm	2015	MRC Psycholinguistic Database, Coltheart (1981)
Yang	2015	na
Tauber	2012	Kucera and Francis (1967)
Dougherty	2005	Paivio, Yuille, and Madigan (1968)
Kelemen	1997	Paivio, Yuille, and Madigan (1968)
Zechmeister	1980	Spreen and Schulz (1966)

2.3.4 Reactivity

Thirty-one experiments reported positive reactivity, five experiments reported negative reactivity, and nine experiments reported no reactivity. For related word pairs, 19 experiments observed positive reactivity, four negative reactivity and three no reactivity. For unrelated pairs, by contrast, seven experiments observed positive reactivity, two negative reactivity and 18 no reactivity.

Thirty-seven experiments adopted experimenter paced encoding time, with 28 reporting positive reactivity, two negative reactivity and seven with no reactivity. Six experiments adopted self-paced encoding time, with one reporting positive reactivity, three negative reactivity and two with no reactivity.

For the studies that included a retention interval (a minimum of a one-minute delay between the encoding and test phase), 30 reported positive reactivity, two reported no reactivity, and none reported negative reactivity. In contrast, for the studies without a retention interval, one reported positive reactivity, five reported negative reactivity, and five reported no reactivity.

Most experiments used cued-recall tests (*n* = 33), with 21 reporting positive reactivity, seven no reactivity, and five negative reactivity. For the experiments using free-recall, three reported positive reactivity, and one no reactivity. All experiments that assessed reactivity with old/new target recognition test (*n* = 6) reported positive reactivity. The three experiments that compared test outcomes (Myers et al., 2020) reported positive reactivity for cued-recall tests across two experiments and no reactivity in one experiment, positive reactivity for a recognition test in one experiment, and no reactivity for free-recall across two experiments.

2.4 Discussion

In the present work, we sought to provide an up-to-date and comprehensive review of JoL reactivity effects. In line with the conclusions of Double et al. (2018) – the last systematic review of immediate JoL reactivity – word pair relatedness appeared to moderate the JoL effect, with participants most likely to benefit from completing JoLs when studying related word pairs. In addition, we also found tentative evidence to suggest that JoL reactivity is moderated by the type of encoding time and retention interval.

The remaining section is divided into two parts. We first consider the variables that appear to influence whether JoL reactivity is observed and, if so, whether the benefits of collecting JoLs is positive or negative (relative to a no-JoL condition). We then consider the evidence for and against the key theoretical frameworks of JoL reactivity, given those moderating factors.

2.4.1 Factors Affecting JOL Reactivity

The present review shows that related word pairs are more likely to result in positive JoL reactivity than unrelated word pairs (the *increased relatedness effect*; Janes et al., 2018). However, given the discrepancies in the literature, it is important to consider methodological differences between the studies. For example, Soderstrom et al. (2015) and Mitchum et al. (2016)

both carried out word paired associate experiments comparing related and unrelated word pairs with a cued-recall test, but Soderstrom et al. reported positive reactivity for related pairs, whereas Mitchum et al. reported negative reactivity for unrelated pairs. Mitchum et al. suggested the difference may be due to the use of a self-paced encoding procedure instead of Soderstrom et al.'s experimenter-paced procedure. However, in a follow up study (Experiment 5), Mitchum et al. continued to report negative reactivity even with an experimenter-paced procedure. In an attempt to resolve these discrepant results, Janes et al. (2018) attempted to replicate both Soderstrom et al.'s and Mitchum et al.'s experiments. However, Janes et al. observed positive reactivity when the encoding phase was both experimenter-paced and self-paced and hence failed to replicate Mitchum et al.'s key result (negative reactivity for unrelated pairs). It may be a further methodological difference can account for Mitchum et al.'s findings, or the findings may result from a Type I error. Nevertheless, the negative reactivity reports are anomalous amongst the broader JoL reactivity literature. Thus, despite some negative reactivity reports, most of the evidence from word pair associate experiments suggests related word pairs consistently produce positive JoL reactivity.

The experiments in the present review suggest that the retention interval may be a moderator of JoL reactivity. Experiments with a delayed test most often resulted in positive reactivity, whereas immediate tests (i.e. no retention interval) resulted in no or negative reactivity. However, the immediate test experiments are overly represented by Mitchum et al. (2015) with five experiments. As discussed, Mitchum et al. used a different methodology compared to most of the studies in this review, including the use of self-paced study time. It is not possible in this review to differentiate the strength of influence for each variable, and hence there must be caution in attributing the effect to only one cause (i.e., test delay over study time). Furthermore, Witherby and Tauber (2017) manipulated the retention interval with a word pair associate experiment comparing a three minute and two-day delay. Both conditions resulted in positive reactivity with no significant difference. It is possible that the influence of the retention interval suggested in this review may be more apparent than real.

The experiments in the present review showed little difference between the reactive effects of different test types. Myers et al. (2020) was the only experiment to directly test the influence of test type, with a comparison of JoL reactivity from word pairs with cued-recall, free-recall and old/new target recognition tests across four experiments. A meta-analysis of their four experiments reported a medium effect size for positive JoL reactivity with cued-recall, no effect for free-recall, and a small positive effect for recognition tests. The authors proposed that the recognition test provided an advantage over free-recall as it provided the ability to retrieve other elements of the original encoding context, such as the complete studied word pair, thereby accessing the strengthened relationship. Future research should continue to explore the role of the test format.

Tauber and Witherby (2019) provided the only study in this review to examine JoL reactivity in different populations. The authors compared JoL reactivity between younger and older adults, as most JoL reactivity findings derived from an undergraduate population. Across five experiments, younger adults consistently demonstrated positive JoL reactivity, whereas older adults failed to show JoL reactivity. These results occurred with related and unrelated word pairs, detailed and simple JoL instructions and reduced study time. The authors caution against concluding that JoL reactivity is not present in older adults, given the unique context of the word pair research paradigm. However, the findings highlight the potential for population variation for JoL reactivity. Such variation is already demonstrated for other metacognitive monitoring tasks, such as the impact of participant self-confidence (Double & Birney, 2017) and working memory (Griffin et al., 2008). Almost all JoL reactivity research in this review provided insufficient consideration of participant characteristics. Future research would benefit from exploring how personal characteristics inform JoL reactivity.

2.4.2 Theoretical Frameworks of JoL Reactivity

Theoretical frameworks of JoL reactivity attempt to explain the mechanisms that produce JoL reactivity. Such frameworks provide the foundation for future research and offer insight into *how* JoL reactivity is produced. However, despite the importance of the frameworks, there is yet to be a review of the theoretical accounts of immediate JoL reactivity. Consequently, the present review investigated the 18 included papers for any discussion of theoretical frameworks and collated the findings for a narratively driven summary.

The theoretical accounts in this discussion are divided between relational and non-relational accounts of reactivity – a novel categorisation. Relational accounts provide a framework for JoL reactivity in relational contexts (e.g., word pairs) and are the most widely discussed account within the JoL reactivity literature (Janes et al., 2018; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015). However, such relational frameworks fail to account for reactivity in non-relational contexts (e.g., word lists). The following discussion includes relational and non-relational accounts to ensure a broad understanding of the mechanisms driving JoL reactivity. Although, the theoretical frameworks are bound by reference only to the studies that were selected for inclusion in accordance with the systematic review. It is possible that additional theoretical accounts are available in papers not included in this review.

2.4.2.1 Relational Frameworks of JoL Reactivity

2.4.2.1.1 Cue-strengthening Hypothesis

Soderstrom et al. (2015) proposed a cue-strengthening hypothesis to explain JoL reactivity which built upon and combined two existing encoding theories. The first is Koriat's (1997) cue-utilisation approach. Here, participants draw upon predominately intrinsic cues (such as the perceived association between cues and targets) when making a JoL. For example, a participant will perceive the word pair *light-dark* as strongly related, resulting in highly rated JoLs. The second theory is de Winstanley's (1996) account of generation effects. Here, the act of generation results in greater recall. For example, participants will have greater recall after generating a word pair (e.g., *sad-???*) rather than reading a word pair (e.g., *sad-happy;* Bertsch et al., 2007). From these two theories, Soderstrom et al. formed a cue-strengthening hypothesis that states that the act of making a JoL strengthens the cues informing the judgment, providing a generative effect that will enhance performance should a test be sensitive to the strengthened information (e.g., a cued-recall test: *light-???*). It follows that JOLs will boost memory for related pairs (relative to no-JoLs) more than unrelated pairs.

As discussed in the Introduction, Soderstrom et al. (2015) conducted a series of experiments demonstrating positive JoL reactivity for related word pairs and no significant reactivity for weakly related and unrelated word pairs. Soderstrom et al. also observed that JoLs attenuated the effects of a generation task, suggesting JoLs provide similar benefits to learning as the act of generation. The authors suggest their results show that the performance-enhancing effect of JoLs is dependent upon the degree of cue-target association: related word pairs benefited from JoLs whereas weakly and unrelated pairs did not. Soderstrom et al. concluded that their findings provided initial evidence for the cue-strengthening hypothesis and a potential explanation for the increased relatedness effect (related word pairs are more likely to result in positive JoL reactivity than unrelated word pairs).

Many studies in the present review report the increased relatedness effect (Janes et al., 2018; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015), with some authors attributing the effect to the cue-strengthening hypothesis (Myers et al., 2020; Rivers et al., 2021; Witherby & Tauber, 2017). However, this attribution remains hypothetical, as no experiments provide evidence of the strengthened relationship between cues and targets (i.e., there is no evidence that the association between *light-dark* is stronger after making a JoL). The hypothesis can only draw upon correlational data (word relatedness coincides with reactivity). As such, the pattern of results can be explained by an alternative hypothesis (e.g., see changed-goal hypothesis below).

A further limitation of the cue-strengthening hypothesis is that the theory is not clear how strengthening occurs. Rivers et al. (2021) suggested that strengthening may be a product of attentional processes, namely reduced mind-wandering or increased attention to the word pairs during encoding. While there is limited evidence for this suggestion, two studies in this review measured the role of attention with JoLs using a dot-probe task (Dougherty et al., 2018; Robey et al., 2017). In this task, four boxes surrounded the word pair, with an asterisk fading into one of the squares after revealing the word pair. The participants then pressed a button as soon as they saw the asterisk, with a quick press indicating greater attention. There was no significant difference between the JoL and no-JoL participants' reaction scores, indicating that JoLs neither increased nor decreased attention. Rivers et al. (2021) further suggested that JoLs could strengthen cues by encouraging participants to change encoding strategies. To foreshadow a discussion below (see changed-goal hypothesis), in the present review, there is no evidence of participants changing strategy in response to a JoL.

A final limitation of the cue-strengthening hypothesis is that it can only account for the increased relatedness effect. The present review showed that unrelated word pairs occasionally resulted in reactivity (Dougherty et al., 2005; Li et al., 2021; Tauber & Witherby, 2019), and item-by-item word lists consistently resulted in positive reactivity (Li et al., 2021; Yang et al., 2015; Zechmeister & Shaughnessy, 1980). If cue-strengthening accounts for reactivity in relational contexts, a separate mechanism is required to explain reactivity in non-relational contexts.

2.4.2.1.2 Changed-goal Hypothesis

Mitchum et al. (2016) proposed the changed-goal hypothesis to explain JoL reactivity. The authors suggest that JoLs change the participant's study goals by bringing attention to the most memorable learning materials (i.e., *stop-pause* is a more memorable word pair than *sock-voice* due to the greater associative strength). Participants are then suggested to prioritise their efforts towards the easier learning material, resulting in greater performance of the prioritised items compared to the unprioritised items. Thus, related word pairs will outperform unrelated word pairs. As discussed previously, Mitchum et al., (2016) supported the changed-goal hypothesis with a series of experiments reporting negative reactivity for unrelated word pairs and no reactivity for related word pairs when both types of word pairs were mixed within a list. In contrast, no reactivity occurred when the experimenters only presented unrelated pairs. The authors suggested that the mixed word pairs allowed the participants to compare the word pairs, resulting in prioritisation and reactivity of related word pairs. The isolated pairs could not change the participant's study goals as there was no comparison, resulting in no reactivity. Janes et al. (2018) supported the changed-goal hypothesis with a replication of Mitchum et al. (2016). Janes et al. observed JoL reactivity when related and unrelated pairs were presented in a mixed list design and no reactivity for pure lists of only related or unrelated pairs. Other experiments in the present review similarly report no reactivity for pure word pairs (Dougherty et al., 2018; Mitchum et al., 2016; Robey et al., 2017).

Mitchum et al. (2016) argued that the changed goal hypothesis would predict negative reactivity for difficult stimuli (e.g., unrelated pairs) when presented alongside easier stimuli (e.g., related pairs). However, subsequent studies supporting the changed-goal hypothesis contest Mitchum et al.'s conclusion. Janes et al. (2018) reported positive reactivity for related pairs in a mixed list, suggesting the participants prioritise the more memorable related pairs and benefit from their changed-goal orientation. This creates a positive reactivity effect. In contrast, DeYoung and Serra (2021) reported no reactivity for weakly related word pairs and negative reactivity for related pairs. The authors suggested the participants may have reduced their efforts toward the easier pairs, resulting in negative reactivity. At present, there is no consensus of the direction of reactivity for the changed-goal hypothesis.

The changed-goal hypothesis was challenged by the observations of Rivers et al. (2021), who asked the same participants to give JoLs for some items, and not for others, with mixed lists of related and unrelated word pairs. The authors proposed the changed-goal hypothesis would predict an improvement in all related word pairs (even those without a JoL) because the presence of JoLs for half the time would result in a prioritisation of all related word pairs (a global change in goal orientation). The related word pairs with a JoL was the only condition with positive reactivity, suggesting JoLs failed to create a global change in goal orientation. Rivers et al. replicated their findings in two follow up experiments, which produced the same outcome when putting the word pairs in a blocked design (rather than mixed). The authors concluded that the changed-goal hypothesis could not explain the results because only the related pairs with JoLs increased. If the participants changed goal orientation, all the related pairs (JoL and no-JoL) would benefit as the JoL pairs were mixed with no-JoL pairs. However, Rivers et al. acknowledged that their conclusion relied upon a participant making a global change in their learning goals (i.e., participant priorities change for all word pairs). If a participant makes local changes in their learning goals (i.e., a participant prioritises change for only those with a JoL), then the changed-goal hypothesis could explain the results.

How could a change in goal orientation result in reactivity? Mitchum et al. (2016) proposed that JoLs encourage metacognitive monitoring (assessment of the ongoing learning), which may bring attention to the participant's study strategy (e.g., imagery or creating sentences for the words). The authors suggested a change in strategy may be evidenced by participants' use of study time, assuming a change in self-chosen study time reflects the participant's change in study strategy. Mitchum et al. (2016) observed that JoL participants spent less time studying unrelated pairs than no-JoL participants when they could pace their study. However, a replication of Mitchum et al.'s experiment by Janes et al. (2018) failed to replicate the change in study time. Mitchum et al. (2016) also asked the participants to complete a questionnaire rating their use of various memory strategies. There was no significant difference between the strategies of the JoL and no-JoL conditions. Rivers et al. (2021) asked participants to share their memory strategy after a recall test but similarly found no significant difference between learning strategies used with or without a JoL. Tauber and Witherby (2019) explored whether JoL instructions implicitly induce memory strategies (the author's JoL instruction asked participants to give judgements based on the distinctiveness of the item in their memory). The authors compared simple JoL instructions with detailed JoL instructions but found no significant difference between the groups. If the instructions were inducing a change in strategy, it did not significantly impact reactivity. It is possible that the methods used by Rivers et al. (2021) and Tauber and Witherby (2019) do not capture how JoLs change learning strategies. For example, Rivers et al. suggested that participants using imagery may have generated a more detailed image with a JoL than without, but in both instances used the same strategy, thus not resulting in a reported change of strategy. Such measurement limitations may obfuscate finding a change in strategy. However, overall, there is very little evidence for JoL reactivity resulting from a change in study strategy.

The changed-goal hypothesis can only explain reactivity in relational contexts (i.e., items with a comparison, such as a mixed list of related and unrelated word pairs), not non-relational contexts (i.e., items without a comparison, such as a block of unrelated word pairs, or an uncategorised word list). As with cue-strengthening, the changed-goal hypothesis cannot account for the positive JoL reactivity findings for pure lists of related or unrelated word pairs (Dougherty et al., 2005; Li et al., 2021; Tauber & Witherby, 2019; Witherby & Tauber, 2017) and positive reactivity of word lists (Li et al., 2021; Yang et al., 2015; Zechmeister & Shaughnessy, 1980). There is no comparison available in these non-relational designs. Thus, a separate mechanism is required to explain reactivity in non-relational contexts.

2.4.2.2 Non-relational Theoretical Frameworks

The cue-strengthening and change-goal hypotheses attempt to explain JoL reactivity in a relational context (e.g., word pairs), but a separate mechanism is required to explain reactivity in non-relational contexts (e.g., word lists). Three non-relational frameworks were reported within the literature of the present review: JoLs as a retrieval opportunity and two theories that explain JoL reactivity as a form of elaborate attentional encoding. The following section discusses each of the three non-relational theoretical frameworks to further understand

2.4.2.2.1 Retrieval Opportunity

Spellman and Bjork (1992) hypothesised that when participants make a JoL, they engage in covert retrieval of the item. This theory was based on early studies demonstrating that participants attempt to retrieve targets before making a JoL about the target (Nelson & Dunlosky, 1991). Dougherty et al. (2005) built upon Spellman and Bjork's hypothesis and suggested that the retrieval opportunity could boost memory by strengthening the memory trace and improving subsequent retrieval. Therefore, JoLs serve as a form of retrieval practice (Dougherty et al., 2018), potentially providing learning benefits akin to the testing effect (see Karpicke & Roediger, 2007).

Dougherty et al. (2005) reported that JOLs improved memory even when the participants also had an opportunity to practice retrieval. This result suggests that JoLs may benefit memory beyond a retrieval opportunity. However, subsequent replications found that JoLs provided no additional benefits to a retrieval opportunity (Dougherty et al., 2018; Robey et al., 2017), leading the authors to attribute the original results to a Type I error. Dougherty et al. (2018) argued that covert retrieval is unlikely for immediate JoLs as there is no opportunity between learning and a JoL for discrete recall. Overall, there is limited evidence supporting JoLs as a retrieval opportunity resulting in reactivity in the present review.

2.4.2.2.2 Item-specific Processing

Senkova and Otani (2021) hypothesised that JoLs enhance item-specific processing, resulting in reactivity. Specifically, JoLs direct attention to a given item and enhance the item's distinctiveness in memory. This contrasts with relational processing (such as the cue-strengthening hypothesis), which increases organisation (Hunt & Einstein, 1981).

Senkova and Otani (2021) compared JoLs to a pleasantness rating and imagery task, as both are known to enhance item-specific processing. In addition, Senkova and Otani paired JoLs with categorised lists. Hunt and Einstein (1981) observed that an item-specific processing task paired with stimuli promoting relational processing (e.g., a categorised list) benefited learning more than an item-specific processing task paired with stimuli also promoting item-specific processing. The JoLs and the established item-specific processing tasks resulted in equivalent positive reactivity for categorised lists and no reactivity for uncategorised lists. The parallel outcome between JoLs, the pleasantness rating and imagery tasks resulted in Senkova and Otani concluding that JoLs appeared to enhance item-specific processing.

Senkova and Otani (2021) highlighted that, as with much of the processing approach to memory, there are no measures of the type of processing underlying memory performance. The observation that JoLs performed equally to the established item-specific processing tasks does not necessarily mean JoLs and the established tasks use the same cognitive mechanisms. To address this problem, the authors propose that future research should use a repeated-measures design based on the work of Klein et al. (1989). Klein et al. adopted a repeated-measures design in which participants attempted the same recall test multiple times without additional study trials. Then, the performance gains (additional recall items on a subsequent test) were measured against the performance losses (items present on an initial test but not on later tests). The authors observed that item-specific processing resulted in item gains, whereas relational processing reduced the likelihood of item losses. A similar methodology for JoL reactivity research would allow an index of relational and item-specific processing for JoLs and provide a vital window into the underlying mechanism of JoL reactivity.

2.4.2.2.3 Elaborate Processing

The final hypothesis for JoL reactivity derives from Tekin and Roediger's (2020) study using a level of processing (LoP) research paradigm. The LoP framework, introduced by Craik and Lockhart (1972), posits that deep processing will produce longer lasting retention than shallow processing. Tekin and Roediger (2020) explored whether JoLs affected the LoP effect by integrating JoLs into Craik and Lockhart's (1972) LoP paradigm. They presented participants with a series of target words, along with orienting questions that were designed to elicit different LoP. To elicit orthographic processing (processing based on appearance, a shallow LoP), participants answered questions that were related to the appearance of the target word (e.g., is the word "cow" in lowercase?). To produce phonetic processing (intermediate LoP), participants answered phonemic questions related to the sound of the target word (e.g., does the word "cow" rhyme with "row"?). Finally, to encourage semantic processing (deep LoP), participants answered semantic questions that related to the meaning of the target word (e.g., is "cow" a type of animal?). The correct answer to half of the orienting questions was "yes" (e.g., is "chair" a type of furniture?"; congruent condition) and the other half "no" (e.g., is "chair" a type of fruit?; incongruent condition). Half of the participants made JoLs after studying each target word (JoL group), while the remaining participants did not (no-JoL group). After the encoding phase, the participants completed an old/new recognition test, where they had to discriminate the target words from new foils.

Tekin and Roediger (2020) observed a classic LoP effect, but the size of the effect was attenuated in the JoL group. The shallowest processing (orthographic) resulted in the greatest reactivity, whereas the deepest processing (semantic) resulted in the least reactivity. Tekin and Roediger suggested that JoLs may have improved performance by strengthening information that was not otherwise strengthened. When JoLs were added to the orthographic questions, they provided additional elaborate processing that was otherwise missing and bolstered the encoding process. In contrast, adding JoLs to the semantic encoding condition provided little benefit because the semantic condition already encouraged elaborate processing. Hence, the authors argued that JoLs reduced the size of the LoP effect because JoLs provide elaborate processing when elaborate processing is otherwise absent.

Tekin and Roediger's (2020) observations suggest that JoLs provide elaborate processing. However, the study did not provide an insight into the nature of the elaborate processing. Tekin and Roediger (2020) suggest that JoLs may cause participants to consider the inherent memorability of the word, akin to performing semantic processing. Although, the authors highlight that this is the first JoL reactivity study using a LoP paradigm, and future research is required to begin to understand the JoL mechanisms that attenuate the LoP effect.

2.5 Summary

The presence of JoL reactivity has become less equivocal since the last JoL reactivity systematic review (Double et al., 2018). The present review has shown that JoLs can result in reactivity, with most studies reporting positive reactivity and few reporting negative reactivity. Furthermore, the review has shown reactivity is dependent upon influencing factors. At present, the most evidenced impacting factor is word relatedness, with related word pairs most often resulting in positive reactivity and unrelated pairs producing no reactivity. Other factors, such as study time, test delay, test type and population each have tentative evidence of impacting the strength and direction of JoL reactivity. However, any conclusions on JoL reactivity are almost exclusively restricted to word pair or word list learning paradigms.

What mechanisms drive reactivity? The present review explored the theoretical accounts of JoL reactivity. The cue-strengthening hypothesis (Soderstrom et al., 2015) posits that JoLs strengthen the cues that can later be used to guide recall, while the change-goal hypothesis (Mitchum et al., 2016) suggests JoLs change the participants' goal orientation away from mastering all items. Evidence for both theories is inconsistent, with support for cue-strengthening (Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015) contrasting with support for change-goal (Janes et al., 2018; Mitchum et al., 2016). Moreover, neither theory is supported by evidence of a direct change in cognition or memory (i.e., evidence of strengthened cues or a change of goal orientation, such as a change of strategy) and the theories fail to explain reactivity in non-relational contexts, such as word lists. JoL reactivity in non-relational contexts may be best understood by elaborate processing theories, namely that JoLs provide item-specific processing (Senkova & Otani, 2021) or semantic processing akin to deep processing within the LoP framework (Tekin & Roediger, 2020). However, as with the relational theories, these accounts lack evidence of underlying cognitive mechanisms.

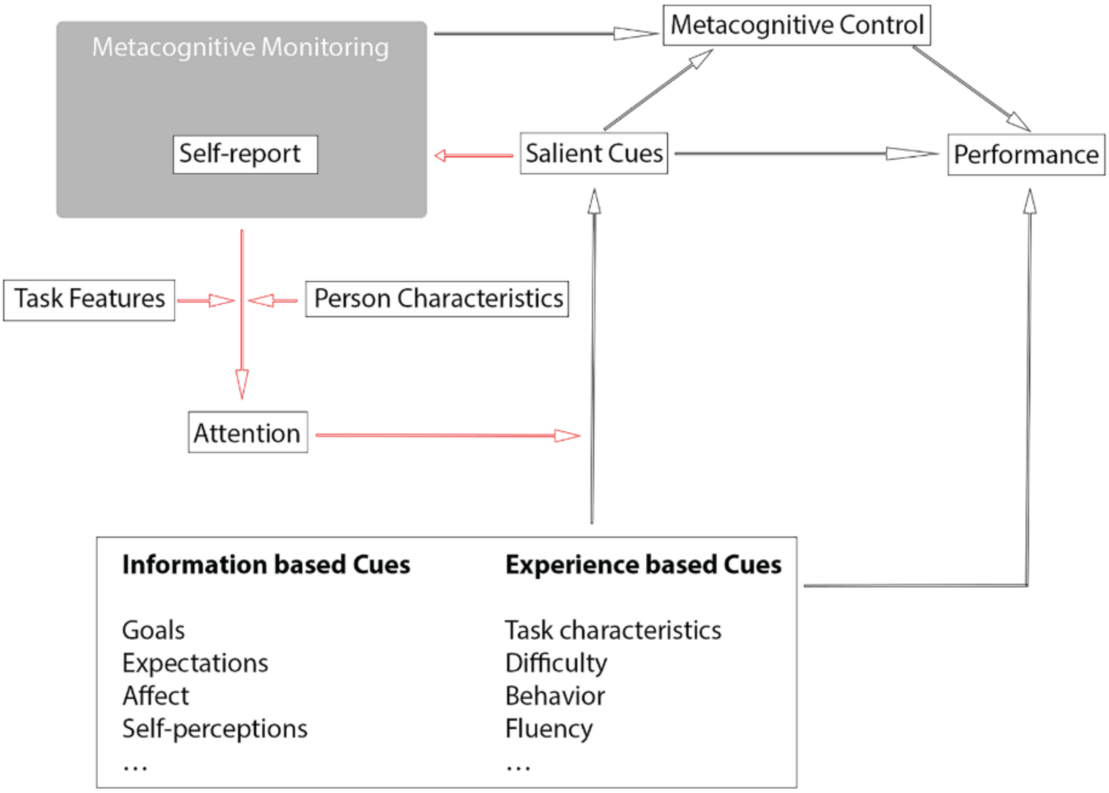
2.6 Future Directions

2.6.1 Cue-driven Metacognitive Framework of Reactivity.

The present review has discussed reactivity through the themes and findings identified within the systematic literature review. In a recent metacognition reactivity paper, Double and Birney (2019) similarly reviewed the theoretical frameworks of reactivity and synthesised different accounts into a single cue-driven metacognitive framework of reactivity (Figure 2). The framework incorporates cue processing, goal orientation, and attentional theories to explain metacognitive judgment reactivity. The framework was developed in relation to different metacognitive measures and drew upon research informing the multiple measures. JoL reactivity research, including many of the studies included in this review, helped inform the creation of the framework.

Figure 2

Double and Birney's (2019) Cue-driven Metacognitive Framework of Reactivity



Note. Red paths indicate the additional demands that form the process of reactivity. The framework was reproduced with the author's permission.

The framework is built upon the cue-utilisation theory, namely, that participants draw upon information-based cues and experience-based cues to inform a metacognitive judgement (Koriat et al., 2008). Information-based cues are from pre-existing beliefs (e.g., "Am I a quick learner?"), whereas experience-based cues are from the learning experience (e.g., "Did I solve the problem quickly?"). From this theory, the authors propose the cue-driven metacognitive framework.

In the framework, participants have access to information- and experience-based cues. When the participant attends to these cues, the cues become salient and inform the judgment. However, the judgment changes the participant's attention, altering the quantity and quality of the salient cues. For example, the cue-strengthening hypothesis suggests attention from JoLs strengthens the associative strength of word pairs (an experience-based cue), while the changed-goal hypothesis suggests attention from JoLs changes the participant's goal-orientation (an information-based cue). Consequently, the altered salient cues create reactivity effects, impacting metacognitive monitoring, control, and cognitive performance.

Double and Birney (2019) explain that a judgment modifies attention according to task features (e.g., word pairs), person characteristics (e.g., age) and judgment features (e.g., scale). Consequently, the direction and strength of judgment reactivity is determined by how attention is modified and the impact of the altered attention upon the available cues. For example, Double et al., (2018) showed how related word pairs reliably produce greater positive reactivity than unrelated pairs, and Tauber and Witherby (2019) observed positive JoL reactivity in younger adults as opposed to older adults. Furthermore, the nature of the reactivity will depend on whether the performance measure is sensitive to salient cues and if cues with a motivational effect impact the participants' goals and approach (i.e., goal orientation), impacting control decisions.

Double and Birney (2019) emphasise that the framework is a tentative attempt to incorporate existing metacognitive measure reactivity theories and not an attempt to provide a complete overview. Nonetheless, the framework incorporates the cue-strengthening hypothesis (altered cues drive reactivity), the changed-goal hypothesis (motivational cues alter metacognitive control) and attentional accounts (judgments alter attention). It can allow for reactivity in different contexts, such as relational and non-relational contexts, and different strengths of reactivity due to different factors informing attention (task features and person characteristics) and different cues becoming salient.

A limitation of the framework is the potentially reductive integration of attention. Research conducted after the publication of the framework suggests JoLs may promote a form of item-specific elaborate processing (Senkova & Otani, 2021) or deep processing within the LoP framework (Tekin & Roediger, 2020). The framework's inclusion of altered attention may understate or misrepresent the potential enhanced processing that occurs during a metacognitive judgment. The framework may instead replace attention with item-specific elaborate processing. This alteration would allow the framework to incorporate the item-specific-relational account of encoding strategies (Mulligan & Peterson, 2015), as the present review highlighted JoLs potential to strengthen relational information (Rivers et al., 2021) and item-specific information (Senkova & Otani, 2021).

2.6.2 Impact on JoL Reactivity Research

JoL reactivity is a memory phenomenon that requires a thorough exploration akin to other memory experiments (Myers et al., 2020). This includes exploring the impact of variations in stimuli, retrieval, participants and encoding (Jenkins, 1979). JoL reactivity research has focused extensively on three types of stimuli (unrelated and related word pairs and word lists) and three types of retrieval (cued-recall, free-recall and recognition). Little is known about the JoL reactivity of other stimuli (e.g., pictures and faces) and tests (e.g., vocal recall). Similarly, only one study in this review studied the role of participants (Tauber & Witherby, 2019) or the role of encoding as a primary research question (Tekin & Roediger, 2020). Future research should continue to explore the impact of different stimuli (e.g., educational texts, Ariel et al., 2021), retrieval (e.g., competing tests, Myers et al., 2020), participants (e.g., children, Zhao et al., 2021) and encoding opportunities (e.g., encoding questions, Tekin & Roediger, 2020) on JoL reactivity.

2.6.3 Impact on Metacognition Research

The presence of JoL reactivity could inform the understanding of metacognitive processes (Double et al., 2018). JoL reactivity suggests that metacognitive monitoring is not necessarily spontaneous (Mitchum et al., 2016; Soderstrom et al., 2015). JoL reactivity research can deepen the understanding of metacognitive decision-making, including which contexts lead to judgments and which lead to memory benefits (Double et al., 2018).

Metacognitive measures, such as JoLs, have provided an understanding of metacognitive monitoring in experimental designs (Rhodes, 2016) and an insight into how such monitoring judgments impact behaviour (Rhodes & Tauber, 2011). However, many studies have used JoLs without accounting for reactivity, introducing an unaddressed confounding variable (Double & Birney, 2019). It is vital for future research to be mindful of JoL reactivity when designing and interpreting results. If memory researchers want to use a measure of metacognitive monitoring, they may explore the use of potentially less reactive measures, such as speak-aloud protocols (Fox et al., 2011) or the exploration of alternative meta-memory judgments such as judgments of forgetting (JoFs). Although, recent findings suggest JoFs and JoLs produce equivalent degrees of reactivity (Li et al., 2021), highlighting the importance of a non-judgment control group to observe potential reactivity.

2.6.4 Impact on Education

JoLs have the potential to be a tool for education given the trend towards positive JoL reactivity in certain contexts. JoLs have very few barriers as they are made quick and require few additional resources or extra study time. Indeed, the studies included in this review commonly showed how a JoL group can receive less dedicated study time than a control group but achieve greater performance (Li et al., 2021; Myers et al., 2020; Rivers et al., 2021). Furthermore, preliminary evidence suggests there is no significant difference to reactivity when JoLs are learnt with simple or complex instructions (Tauber & Witherby, 2019; Witherby & Tauber, 2017), positive reactivity can last over time (Tekin & Roediger, 2020; Witherby & Tauber, 2017) and JoLs can enhances children's learning (Zhao et al., 2021).

JoLs may be theoretically suitable for the classroom, but there is limited evidence of JoL reactivity with educational material. Ariel et al., (2021) provided the first exploration of JoL reactivity with educational material by asking participants to read a text paragraph and provide an immediate JoL. No reactivity was reported when participants gave an aggregate JoL (an average rating of overall confidence) or JoLs for specific concepts within the paragraph. However, positive reactivity did occur when participants engaged in retrieval practice for specific concepts prior to the JoL. The authors concluded that the participants appeared to insufficiently retrieve the information required for the JoL to produce reactivity without a specific retrieval practice. This spotlights the importance of exploration into JoL reactivity in an educational context before JoLs are recommended to educators.

JoLs for education is an exciting prospect, but JoL reactivity research is in its infancy, and as such, there must be caution in assuming JoLs will be suitable for all participants. Previous work has shown participants with reduced working memory find metacognitive monitoring challenging while performing a comprehension task (Griffin et al., 2008). This dual processing cost is the foundation of the dual-task hypothesis (see Mitchum et al., 2016), although there is currently little evidence that the hypothesis explains reactivity within the current JoL reactivity research (Janes et al., 2018). However, future research exploring different populations, such as those with reduced working memory or other learning differences, may require different hypotheses, such as the dual-task theory, to further understand JoL reactivity.

2.7 Limitations

The present review provided summary statistics and a thematic review, but an updated meta-analysis would be a worthwhile endeavour for future research. In the last review of the JoL literature, Double et al.'s (2018) inclusion of a meta-analysis identified overall results not otherwise available from the individual studies. This included providing effect sizes for subgroups, with a moderate positive effect for related word pairs and word lists and no reactivity to unrelated pairs. A meta-analysis may have provided new results to contribute to the field of JoL reactivity.

Another limitation is that shared by Double et al. (2018), in that the nature of reactivity research can make a systematic search extremely challenging. Reactivity is often not the primary aim of a study and, as a result, reactivity may not be mentioned in an abstract. The present review attempted to capture all relevant studies (this included having a second researcher cross-checking applicable papers) but it is possible the review missed experiments meeting the inclusion criteria. Similarly, this review is beholden to the published literature. While the review attempted to combat publishing bias by including thesis papers, it is possible that the overall trend of reactivity is partly indicative of the unavailability of null results.

Lastly, the inclusion criteria for the systematic search required each study to have a no-JoL control group. While this allowed for the synthesis of JoL reactivity evidence, the search may have excluded papers that could contribute to the reactivity mechanism discussion (e.g., a study attempting to evidence a change of underlying cognition between a JoL and an alternative intervention). Future systematic reviews should broaden the inclusion criteria if exploring underlying mechanisms.

2.8 Conclusion

To conclude, the present systematic literature review provided a synthesis of the immediate JoL reactivity evidence base. JoL reactivity can occur within certain contexts, particularly when using related-word pair stimuli, which typically produces positive reactivity and infrequently negative reactivity. Theoretical accounts of JoL reactivity provide tentative explanations for reactivity, but there is yet to be an evidenced, prominent theory that draws upon evidence of underlying cognitive mechanisms. These findings suggest that future research should attempt to find such evidence, as well as explore JoL reactivity outside of word pair or word list research paradigms to build a more comprehensive understanding of JoL reactivity. The impact of said research could deepen the understanding of metacognitive decision-making and potentially begin towards JoLs as an evidence-based tool for education.

Chapter 3 Investigating Judgment of Learning Reactivity in a Transfer Appropriate Processing Paradigm

Abstract

Research has shown that immediate judgments of learning (JoLs) can affect memory in paired-associate learning paradigms. Such *JoL reactivity* is often attributed to JoLs providing elaborate processing for related word pairs. This is an equivalent line of reasoning to what is suggested to underpin the levels of processing (LoP) effect: tasks that foster deep, elaborate processing should produce longer-lasting retention than tasks that encourage shallow processing. Morris et al. (1977) highlighted that memory is not solely about the depth of encoding but also how memory is tested (the transfer appropriate processing (TAP) effect). Subsequently, we examined the role of encoding processes and test format on JoL reactivity, in a TAP paradigm (Morris et al., 1977) with word pair associates. In an initial encoding phase, we presented the participants with related, rhyming, or unrelated word pairs to induce different LoP. Half of the participants made a JoL after studying each word pair, while the remaining participants simply studied each word pair for an equivalent duration. Afterwards, all participants completed either standard or rhyme old/new recognition tests, where they had to discriminate the targets from the encoding phase (or rhymes of the targets) from novel foils. The performance scores failed to demonstrate the LoP effect. In the standard recognition test, there was no significant difference between the related and rhyming pairs. However, in the rhyme recognition test, the participants successfully recognised more rhymes of targets from the rhyming pairs than the related and unrelated pairs. Thus, we successfully replicated the TAP effect, albeit that the LOP component of the effect was not significant. However, we did not observe significant evidence of JoL reactivity, regardless of encoding or test condition. The study is the first to investigate JoL reactivity using a TAP paradigm with word pair associates and provides a foundation for future work to examine the role of the test on JoL reactivity and JoL reactivity in alternative paradigms.

3.1 Introduction

Judgements of learning (JoLs) are a metamemory measure that requires participants to predict the likelihood of recalling learnt material (Arbuckle & Cuddy, 1969). Memory research employs JoLs with different types of stimulus materials (e.g., word lists, Tekin & Roediger, 2020; or educational texts, Ariel et al., 2021) but are most frequently used with word pairs (Rhodes, 2016). For example, a participant presented with the word pair *sky-dog* may be asked to judge the likelihood that they will recall the target word *dog* in an upcoming test on a scale from 0 (not at all confident) to 100 (extremely confident). The JoLs are considered accurate if they reflect subsequent performance (Schwartz & Metcalfe, 2017). JoLs can be made immediately after learning or made following a delay, with delayed JoLs shown to have greater accuracy than immediate JoLs (Rhodes & Tauber, 2011).

Traditionally, JoLs have been used as a tool to measure metamemory (Rhodes, 2016). Only recently have JoLs been found to affect memory in and of themselves: an effect termed *JoL reactivity* (Double et al., 2018; Rhodes & Tauber, 2011). Reactivity refers to an intentional or unintentional change in behaviour or performance in response to a measurement or observation (Double & Birney, 2019). The first systematic review and meta-analysis of JoL reactivity (Rhodes & Tauber, 2011) compared experiments in which participants studied simple stimuli (e.g., word pairs, such as *sun-moon*) and then provided an immediate or delayed JOL. The authors observed that delayed JoLs resulted in greater performance than immediate JoLs in a later memory test and concluded that delayed JoLs provide a larger positive reactive effect compared to immediate JoLs.

More recently, researchers found that immediate JOLs can be reactive compared to a no-JoL control group (a group experiencing the same encoding and test phases but without making JoLs at encoding). Such JoL reactivity has been most often observed in single-word learning (Li et al., 2021; Senkova & Otani, 2021; Zechmeister & Shaughnessy, 1980) and paired-associate (word pairs) learning paradigms (Janes et al., 2018; Myers et al., 2020; Rivers et al., 2021). In the first study directly examining the effect of immediate JoLs on memory, Soderstrom et al. (2015) asked participants to learn cue-target word pairs comprising of strongly related (e.g., *blunt-sharp*), weakly related (e.g., *boxer-terrible*) and unrelated words (e.g., *sack-flag*). Half of the participants studied each word pair for 4 seconds before providing a JoL for a further 4 seconds (the JoL group). The remaining participants (the no-JoL group) simply studied each word pair for 8 seconds to match the study duration of the JoL group. After a three minute distraction task, all participants completed a cued-recall test (e.g., *blunt-???*). The JoL group successfully recalled more targets from the strongly related pairs than the no-JoL group. However, this *positive reactivity effect* did not extend to the weakly related and unrelated word pairs. Thus, immediate JoLs improved subsequent cued recall, but only for word pairs with a strong semantic association.

While Soderstrom et al.’s (2015) findings appear robust, it is important to note that discrepant results have also been observed. For example, Mitchum et al.'s (2016) participants completed a similar paired-associate learning task, but no-JoL reactivity was seen for strongly related word pairs, and *negative* reactivity (impaired recall) was seen for unrelated pairs. Janes et al. (2018) suggested this divergent result may be due to methodological differences. Soderstrom et al. controlled for an equal study duration between groups (self-paced), whereas Mitchum et al.'s participants were able to control the time they spent studying each word pair (self-paced). Janes et al. (2018) attempted to understand the impact of these methodological differences by replicating Soderstrom et al. and Mitchum et al. in a single study. Janes et al. observed positive reactivity for related pairs in both research designs, with a slight drop in performance for the self-paced condition. Although, there was no difference in allocated study time between related and unrelated pairs. Therefore, Janes et al.'s results contradicted Mitchum et al.'s findings but supported the suggestion that using a self-paced study phase methodology can result in JoL reactivity. Thus, the contrary results remain unresolved (Rivers et al., 2021).

In recent years, JoL reactivity research has continued to observe positive (Maxwell & Huff, 2022), negative (DeYoung & Serra, 2021) and no reactivity (Dougherty et al., 2018). However, there is increasing evidence that reactivity depends on the relatedness between word pairs. Double et al. (2018) provided the first meta-analysis of JoL reactivity and observed reactivity for related word pairs (e.g*., beer-pub*) or word lists (e.g., *man, shop)*, but not unrelated pairs (e.g., *light-fish*) or mixed lists of related and unrelated pairs. Subsequent studies comparing related and unrelated word pairs have consistently reported positive reactivity only for related pairs (Janes et al., 2018; Myers et al., 2020; Rivers et al., 2021), an effect termed the *increased relatedness effect* (Janes et al., 2018).

3.1.1 Theoretical Accounts of JoL Reactivity

Several theories have been put forward to explain the increased relatedness effect. Mitchum et al. (2016) hypothesised that JoLs affect participants' goal orientations. Here, JoLs help a participant attend to the most memorable aspects of a word pair, creating greater study efforts towards some word pairs over others. Mitchum et al. suggested unrelated word pairs are de-emphasised because they have no salient relationship, thereby producing negative reactivity. However, other authors have argued the changed-goal hypothesis may incorporate a different change in goal orientation. For example, DeYoung and Serra (2021) suggested the changed goal hypothesis would predict negative reactivity for related pairs as participants place less emphasis on easier-to-learn pairs (i.e., participants put less effort towards easier, related pairs, inadvertently reducing subsequent recall). By contrast, Janes et al. (2018) argued that the changed goal theory would predict positive reactivity for related pairs because participants would prioritise the related, easier to learn pairs, thereby boosting performance (relative to the no-JoL group). There is no consensus on the direction of reactivity within the changed-goal account of JoL reactivity. Still, there is a shared understanding that JoLs may affect goal orientation.

In contrast to the changed-goal hypothesis, Soderstrom et al. (2015) proposed a cue-strengthening hypothesis informed by two encoding theories. The first theory is Koriat's (1997) cue-utilisation approach, which posits that JoLs are based on the intrinsic cues of a word pair (e.g., the cue and target association). For example, the highly associative word pair *women-men* would likely receive a highly rated JoL. The second is de Winstanley's (1996) account of generation effects, which suggests that generation (creating a mental outcome rather than receiving a defined stimulus) improves subsequent recall. For example, generating a word pair (e.g., *cold-???*) results in greater performance than reading a word pair (e.g., *cold-hot*; Bertsch et al., 2007). Soderstrom et al. combined the two theories and formed a cue-strengthening hypothesis. JoLs strengthen the cues that inform the JoL and provide a generative effect, enhancing subsequent test performance. Therefore, JOLs will boost memory for related pairs (relative to no-JoLs) more than unrelated pairs. Although, Soderstrom et al. stipulated that JoL reactivity will only occur if the final test is sensitive to the same cues used to inform the JoLs. For example, if a participant generated a JoL for the word pair *light-dark*, a cued-recall test (*light-???*) should facilitate JoL reactivity because it contains the same cue (*light*) used to inform the JoL and requires knowledge of the association between the cue and the target. In contrast, a free-recall test (a test without cues or prompts) would be less likely to facilitate reactivity for there is no cue and no requirment to recall the cue target association.

Mixed support has emerged for the two theories – the changed-goal and cue-strengthening accounts - in subsequent years. Janes et al. (2018) replicated Soderstrom et al.'s (2015) and Mitchum et al.’s (2016) experiments and observed reactivity for mixed lists of related and unrelated pairs but not for *pure* lists of only related or unrelated word pairs. They suggested that this finding was best explained by the changed-goal hypothesis because reactivity was only present when there was an available comparison between the word pairs. In contrast, Rivers et al. (2021) provided support for the cue-strengthening hypothesis with a novel within-subjects experiment. They asked participants to give JoLs for some word pairs, but not others, with mixed lists of related and unrelated word pairs and, in a subsequent experiment, with unmixed *blocked* lists of related or unrelated pairs. JoLs only improved subsequent cued-recall of the related word pairs, regardless of whether the word pairs were presented in mixed or unmixed lists. Rivers et al. argued that, according to the changed-goal hypothesis, JoLs should have produced a global change in attention towards the related word pairs. That is, JoLs should have enhanced recall of *all* related pairs (even those without a JoL), because the presence of JoLs on half the trials would result in a prioritisation of all related word pairs. However, the authors acknowledged that this proposition relies on a participant making a global change in their learning goals (i.e., participant priorities change for all word pairs). If a participant makes local changes in their learning goals (i.e., a participant prioritises change for only those with a JoL), then the changed-goal hypothesis could still explain the results.

At present, there is no consensus supporting either the changed-goal or cue-strengthening hypothesis. However, recent findings suggest that JoLs provide an elaborate processing opportunity akin to the cue-strengthening hypothesis. Senkova and Otani (2021) compared JoLs to tasks that enhance item-specific processing (i.e., enhance the item's distinctiveness in memory, such as an imagery task) to investigate if JoLs provide a similar mechanism. The JoLs and the established tasks resulted in equivalent positive reactivity for word lists, suggesting that JoLs provide item-specific processing. While this account differs from the cue-strengthening hypothesis (which suggests that JoLs enhance relational information), they share an understanding that JoLs force participants to encode certain types of information more than they would otherwise.

3.1.2 Levels of Processing

If JoLs provide an elaborate processing opportunity, it will make JoLs analogous to the encoding tasks from the levels of processing (LoP) tradition. The LoP framework, introduced by Craik and Lockhart (1972), posits that deep processing will produce longer lasting retention than shallow processing. Support for the LoP framework was first evidenced by Craik and Tulving (1975). They presented participants with a series of target words, along with orienting questions that were designed to elicit different LoP. To elicit orthographic processing (processing based on appearance, a shallow LoP), participants answered questions that were related to the appearance of the target word (e.g., is the word “cow” in lowercase?). To produce phonetic processing (intermediate LoP), participants answered phonemic questions related to the sound of the target word (e.g., does the word “cow” rhyme with “row”?). Finally, to encourage semantic processing (deep LoP), participants answered semantic questions that related to the meaning of the target word (e.g., is “cow” a type of animal?). After the encoding phase, the participants completed an old/new recognition test, where they had to discriminate the target words from new foils. Across ten experiments, the authors found that targets from the semantic condition were recognised best, followed by targets from the phonemic condition, and finally the targets from the orthographic condition. Thus, tasks that required deep processing of the target words produced the best subsequent recognition, a pattern known as the LoP effect.

Tekin and Roediger (2020) conducted the first study to investigate JoL reactivity in a LoP paradigm. The participants were presented with target words from a word list, along with semantic, phonemic, and orthographic orienting questions, as in Craik and Tulving's (1975) seminal experiments. The correct answer to half of the orienting questions was “yes” (e.g., is “chair” a type of furniture?”; congruent condition), while the correct answer to the remaining questions was “no” (e.g., is “chair” a type of fruit?; incongruent condition). In addition, half of the participants made JoLs after studying each target word (JoL group), while the remaining participants did not (no-JoL group). Finally, all participants completed an old/new target recognition test. The authors observed a classic LoP effect, but the size of the effect was attenuated in the JoL group. The shallowest processing (orthographic) resulted in the greatest reactivity, whereas the deepest processing (semantic) resulted in the least reactivity. Tekin and Roediger suggested that JoLs may have improved performance by strengthening information that was not otherwise strengthened. When JoLs were added to the orthographic questions, they provided additional elaborate processing that was otherwise missing and bolstered the encoding process. In contrast, adding JoLs to the semantic encoding condition provided little additional benefit because the semantic condition already encouraged elaborate processing. Hence, the authors argued that JoLs reduced the size of the LoP effect because JoLs provide elaborate processing when elaborate processing is otherwise absent.

Tekin and Roediger (2020) provided an initial insight into the relationship between JoLs and LoP. However, Morris et al. (1977) demonstrated many years ago that memory is not solely about the depth of encoding but also how memory is tested. Shortly after Craik and Tulving's (1975) LoP findings, Morris et al. (1977) adapted the LoP paradigm such that half of the participants received a standard old/new recognition test, while the remaining participants received a rhyme recognition test (e.g., “was there a target presented during the study that rhymes with *regal*?”). Consistent with the LoP framework, Morris et al. found that the participants recognised more targets from the semantic condition than the phonemic condition. However, in the rhyme recognition test, the phonemic encoding condition outperformed the semantic encoding. Morris et al. therefore concluded that memory is not just about how deeply the items are encoded. Memory also depends on whether the encoding and retrieval processes match, a process dubbed *transfer appropriate processing* (TAP).

The JoL reactivity literature has discussed reactivity with related word pairs in the same way that LoP was discussed before TAP was observed. Reactivity occurs when semantic encoding is enhanced by related pairs, but not unrelated pairs. In other words, reactivity is about how deeply the pairs are encoded. The present study aims to investigate if reactivity is more than deep versus shallow encoding: could JoL reactivity, like LoP, depend on whether the encoding and retrieval processes match?

3.1.3 The Present Experiment

The paired-associate learning paradigm has provided observations of JoL reactivity (Janes et al., 2018; Soderstrom et al., 2015). We amended this paradigm to further examine the role of TAP in JoL reactivity. The TAP paradigm, as explored by Morris et al. (1970), elicited three forms of processing (semantic, processing based on semantic association; orthographic, processing based on appearance; and phonemic, processing based on phonic processing, such as rhyme) with two types of tests (a standard and rhyme recognition test). Morris et al.'s approach enabled the observation that memory performance is affected by the similarity in the cognitive processes that are required at encoding and retrieval (i.e., rhyme encoding produces better memory on the rhyme recognition test than semantic encoding). Subsequently, the present experiment adopts a novel paradigm that combines the paired-associate learning paradigm with Morris et al.'s TAP paradigm. We introduced rhyming pairs (to elicit phonemic processing) and a rhyme recognition test while retaining the method and procedures of the paired-associated paradigm (including related and unrelated pairs, the standard recognition test, and the no-JoL control group). This allowed for an investigation into whether JoL reactivity depends on if the encoding and retrieval processes match.

The initial encoding phase of the present experiment consisted of the participants being presented with related, rhyming and unrelated associate word pairs. Half of the participants made a JoL after studying each word pair. After a short filled retention interval, all participants completed either a standard old/new or a rhyme recognition test on the target words.

We expected the related word pairs to foster the deepest processing because participants would process the semantic relationship between the words. We therefore hypothesised:

(1) The participants would recognise the most targets from the related word pairs than the rhyme and unrelated word pairs in the standard recognition test, reflecting the LoP effect.

(2) The targets from the rhyming word pairs would be processed phonemically and would therefore be best recognised on the rhyme recognition test, reflecting TAP.

(3) Following Tekin and Roediger's (2020) observations, we anticipated that JoLs would attenuate the LoP effect in the standard recognition test. The unrelated word pairs would generate the greatest reactivity, whereas the related word pairs would generate the least.

The novel, competing hypotheses for this experiment pertain to the rhyme recognition test:

(4) If JoLs enhance semantic processing *only*, they should undermine the *phonemic* encoding (rhyming pairs) to produce an attenuated TAP effect. Under this account, the JOL group

should show a significantly smaller benefit of encoding word pairs phonemically (vs. semantically) in the rhyme recognition test, compared to the no-JoL group..

If, however, JoLs enhance the processing of *any* salient relationship (not just semantic relationships), then JoLs should also enhance phonemic processing for the rhyming pairs. There should be no improvement to phonemic processing for the related and unrelated pairs, however, since these pairs do not contain any phonemic relationship. If this is the case, then JoLs should exaggerate the TAP effect.

3.2 Method

3.2.1 Participants

A power analysis using G*Power (Faul et al., 2007) indicated a sample size of 140 participants to detect an effect size of Cohen's *d* = 0.48 (the effect size for the difference between JOL and no-JOL conditions in Myers et al., 2020, Experiment 3), assuming α =.05, power of .80, and a two-tailed test. Six participants were recruited from the University of Southampton and 134 participants from Prolific (www.prolific.co/). The university participants were undergraduate Psychology students who took part for partial course credit, and the Prolific participants received financial compensation of £7.50/hour. The Prolific participants were not restricted based on geography (i.e., an international sample). The experiment received approval by the Psychology Ethics Committee at the University of Southampton. All participants provided informed consent before completing the study.

The participants were required to self-report English as their first language. This was to ensure the participants had the language proficiency to understand the study instructions and the word pairs. This decision replicates the inclusion criteria of Tekin and Roediger (2020).

One participant was excluded for answering that English was not their first language, another was excluded for answering "yes" to a cheat check, and two participants failed a bot check (see below). The data from three Prolific participants was lost due to a technical error. The final sample consisted of 133 participants (89 female, 44 male) aged 18 and 59 years (*M* = 30.57 years, *SD* = 10.25 years). The participants were randomly allocated to a JoL condition with a standard recognition test (*n* = 34) or rhyme test (*n* = 34), or a no-JoL condition with a standard recognition test (*n* = 33) or rhyme test (*n* = 32). We did not inform the participants of their condition or of the differences between conditions.

3.2.2 Materials

Thirty related word pairs (mean forward associative strength = .33; e.g., *walk-run*) were selected from the University of South Florida Free Associations Norms Database (Nelson et al., 2004). Each target word was paired with an unrelated cue word (e.g., *yield-run*), a rhyme cue (e.g., *fun-run*), and a further rhyming word to be presented in the rhyme recognition test (e.g., *gun-run*), each without any associative strength. The full list of materials is available in Appendix B.

Each participant received a random allocation of 15 target words during the encoding phase. Five target words were presented with a semantic cue (e.g., *cat-dog*), five targets were presented with a rhyming cue (e.g., *hand-grand*), and five targets were presented with an unrelated cue (e.g., *men-hide*). The additional rhyming word for each target was presented in the rhyme recognition test instead of the target. Four additional word pairs, consisting of cities and countries (e.g., *Paris-France*), were created for practice trials. The experiment was programmed in JavaScript using the jsPsych library (de Leeuw, 2015) and was hosted on JATOS (Lange et al., 2015).

3.2.3 Design

The experiment had a 3 (Encoding task: semantic, rhyme, or unrelated) × 2 (Test type: standard target recognition or rhyme recognition) × 2 (Judgment type: JoL or No-JoL) mixed-factorial design, with the encoding condition manipulated within-subjects, and the test and JOL conditions manipulated between-subjects.

3.2.4 Procedure

The participants completed the experiment on a full-screen web browser on a remote laptop or desktop PC. The participants first provided information about their web browser, age, gender, and English fluency. Afterwards, the participants completed a “bot check” by selecting a red letter from a grid of black letters. Unlike an in-person study, online studies can be manipulated by users employing programmes (“bots”) to satisfice study completion (Proflific, 2022). The letter task is a simple quality control measure to ensure that real participants complete the study, while also ensuring that their screen and keyboard worked. Afterwards, the participants were instructed to turn off distractions, complete the experiment in a single session, and not to close the webpage during the experiment.

The instructions explained to the participants that they would study a series of word pairs. Each pair consisted of a lower case cue and an uppercase target (e.g., *walk-RUN*). The instructions explained that a later test would assess the participants' memory of the target words. The participants agreed to use only their memory during the study (i.e., not recording the word pairs).

All participants completed four practice encoding trials using the practice word pairs. Then, the participants began the encoding phase. Each participant was shown 15 word pairs (five semantic, five rhyme, and five unrelated word pairs) in a random order. Participants in the no-JoL condition studied each pair for 8 seconds. Participants in the JoL condition were shown each pair for 4 seconds before being asked to provide a JoL (the word pair remained visible). The JoL screen presented a JoL instruction, "Please rate the likelihood that you will be able to recall the capitalised target word if shown the lowercase word on a later test", and a 0-100 scale from "not at all likely" to "certain". The participants used the mouse to move the scale slider. Participants had 4 seconds to provide their JoL to equate the study time with the no-JoL group. If a participant failed to make a JoL, a new screen asked them to "please click on the confidence rating scale within the allocated time" before moving on to the next word pair. A 1 second interval separated each word pair.

After the encoding phase, the participants completed a 3 minute distractor task. The participants were instructed to answer whether arithmetic questions (e.g. *(4 x 4) + 1 = 17*) were correct or incorrect. Next, all participants completed the test phase. All participants received 30 randomly ordered test trials: 15 with targets (or rhymes of the targets) from the encoding phase and 15 with novel foils. Participants in the standard target recognition test condition answered “yes” or “no” to, "Is this one of the TARGET words presented earlier?" alongside the presentation of an uppercase target or foil word. Participants in the rhyme recognition condition, by contrast, answered “yes” or “no” to, "Does this word rhyme with one of the TARGET words presented earlier?" alongside the presentation of an uppercase word that rhymed with one of the targets or foils. For example, the test word *MOUSE* rhymes with the target word *HOUSE*. If *HOUSE* had been presented during the encoding phase, the correct answer would be *yes*. The rhyming word was always novel to the participant because it was not the rhyming cue used in the rhyme condition during the encoding phase. All participants had to respond before moving to the next trial (responding was self-paced). After the test phase, all participants confirmed whether they recorded the stimuli and received a written debrief.

3.2.5 Analysis

We first report the performance scores and then analyse the JoL ratings. We conducted ANOVAs and post-hoc t *tests* using R Version 4.1 (R Core Team, 2014). The alpha level for all analyses was set to .05. All pairwise comparisons reported as significant used a Bonferroni correction (α = .01667) to control for familywise error. For the performance scores, we calculated the false alarm rate (i.e., an incorrect "yes" recognition response to a foil), the hit rate (i.e., a correct "yes" recognition response to a target) and the corrected hit rate (i.e., the hit rate subtract the false alarm rate) to understand if there were differences in the data. Supplemental materials reporting data and analysis scripts are available on the Open Science Framework (https://osf.io/ak57u/).

3.3 Results

3.3.1 Test Performance

To understand the test performance scores, we calculated the false alarm rate, the hit rate and the corrected hit rate. Analysis of the hit rate and corrected hit rate did not reveal different outcomes. Consequently, we focused our analysis below on the uncorrected hit and false alarm rates, but the equivalent analysis for the corrected hits is provided in Appendix C.

3.3.1.1 False Alarm Rate

A 2 (Test type: standard target recognition or rhyme recognition) x 2 (Judgment type: JoL or No-JoL) between-subjects ANOVA on the false alarms scores revealed that the main effect of judgment type was not significant, $F(1, 129) = 0.14$, $p = .71$, $\eta_g{}^{2} = .001$, nor was the judgment type x test type interaction $F(1, 129) = 1.06$, $p = .36$, $\eta_g{}^{2} = .008$. The main effect of test group was significant $F(1, 129) = 14.11$, $p < .001$, $\eta_g{}^{2} = .10$. The rhyme recognition group had significantly higher false alarms (*M* = 12.53%, *SD* = 15.18%) than the standard recognition group (*M* = 4.78%, *SD* = 6.94%).

3.3.1.2 Hit Rate

We performed a 3 (Encoding task: semantic, rhyme, or unrelated) × 2 (Test type: standard target recognition or rhyme recognition) × 2 (Judgment type: JoL or no-JoL) mixed ANOVA on hit rates. Overall, hits were significantly more likely for items in the standard test (*M* = 82.89%, *SD* = 18.91%) than the rhyme test (*M* = 40.40%, *SD* = 28.21%), $F(1, 129) = 205.36$, p < .001, $\eta_g{}^{2} = .45$. The ANOVA further revealed a significant main effect of test type for rhyme (*M* = 66.17%, *SD* = 30.77%), semantic (*M* = 61.20%, *SD* = 33.64%) and unrelated encoding conditions (*M* = 58.05%, *SD*

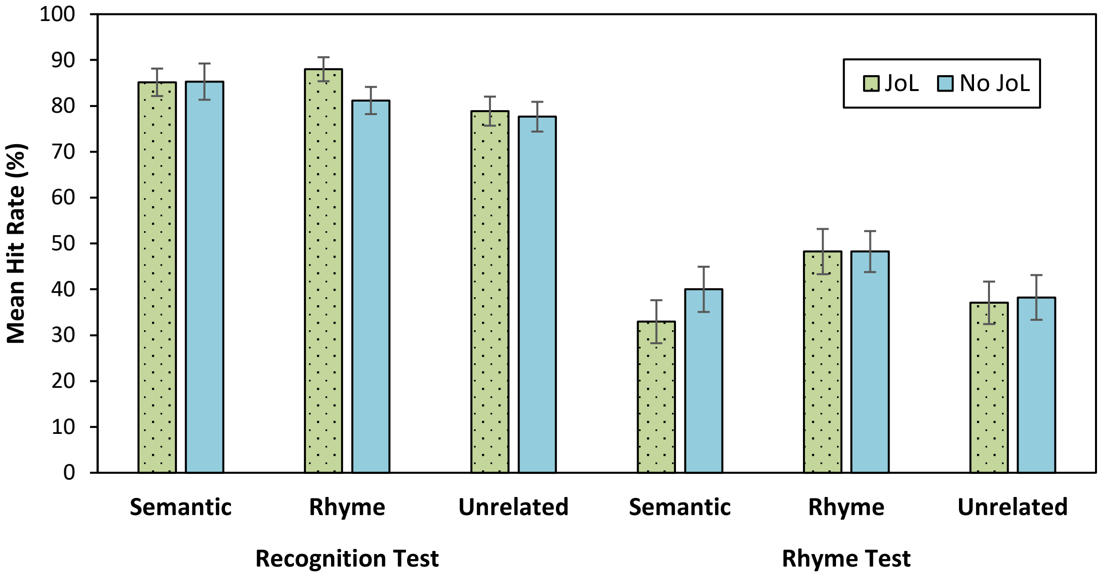
= 31.32%, *F*(2, 258) = 5.36, *p* = .005, η_g^2 = .02, and a significant test type x encoding task interaction, *F*(2, 258) = 3.58, *p* = .03, η_g^2 = .01. The other main effects and interactions were not significant (see appendix D for the full hit rate analysis). Consequently, the results reject our hypotheses (3 and 4) on JoL reactivity, given that we did not observe reactive effects.

To further examine the significant interaction between test type and encoding task, paired samples *t*-tests were performed for the encoding conditions in each test type. For the standard target recognition test, the percentage of hits was significantly higher for the semantic condition (*M* = 85.97%, *SD* = 17.06%) than the unrelated condition (*M* = 78.21%, *SD* = 18.98%), *t*(66) = 2.82, *p* = .006. There was not a significant difference between the rhyme (*M* = 84.48%, *SD* = 19.95%) and semantic conditions, *t*(66) = 0.53, *p* = .60, or the rhyme and unrelated conditions, *t*(66) = 2.08, *p* = .04 (non-significant in accordance with the Bonferroni correction, α = .01667). The results were contrary to our hypothesis (1): the performance scores failed to demonstrate the LoP effect as there was no significant difference between the semantic and rhyme condition.

For the rhyme recognition test, the percentage of hits was significantly higher for the rhyme condition (*M* = 47.58%, *SD* = 28.67%) than semantic condition (*M* = 36.06%, *SD* = 27.00%), *t*(65) = 2.94, *p* = .004, and the unrelated condition (*M* = 37.58%, *SD* = 27.96%), *t*(65) = 2.51, *p* = .015. There was no significant difference between semantic and unrelated conditions, *t*(65) = 0.35, *p* = .73. The results support our hypothesis (2) that the participants would successfully recognise more rhymes of targets from the rhyming pairs than the related and unrelated pairs (replicating the TAP effect). However, the LoP component of the effect was not significant, given the no significant difference between the semantic and unrelated pairs.

Figure 3

Mean Percentage of Hits Per Encoding Condition for Each Recognition Test



Note. The error bars represent the mean standard error.

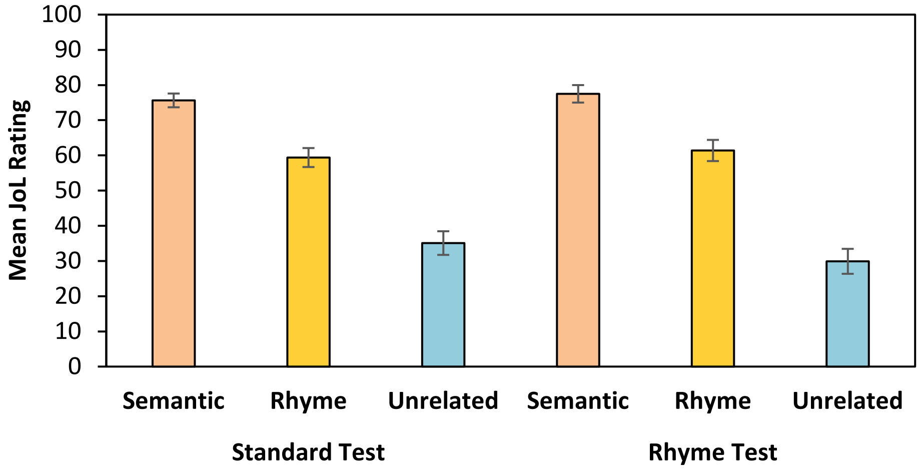
3.3.2 JoL Ratings

The participants in the JoL condition failed to provide JoLs on an average of 5.1% of trials (*SD* = 12.73%). To compare the remaining JoL ratings across the conditions, we conducted a 3 (Encoding task: semantic, rhyme, or unrelated) x 2 (Test type: standard target recognition or rhyme recognition) mixed analysis of variance (ANOVA) for participants in the JoL groups only (given that the no-JoL groups did not give JoL ratings). One participant was excluded for not providing any JoL ratings, resulting in 67 applicable participants. The results indicated that the main effect of encoding task was significant, *F*(2, 130) = 244.70, *p* < .001. Subsequent paired samples t-tests compared JoL ratings in the rhyme, semantic and unrelated encoding conditions.

The JoL ratings for the semantic pairs (*M* = 76.55, *SD* = 12.83) were significantly higher than the rhyming pairs (*M* = 60.38, *SD* = 16.42), *t*(66) = 10.18, *p* < .001, and the unrelated pairs (*M* = 32.54, *SD* = 19.99), *t*(66) = 18.64, *p* < .001. The rhyming pairs were significantly higher than the unrelated pairs, *t*(66) = 13.45, *p* < .001. The main effect of test type was not significant *F*(1, 65) = 0.08, *p* = .90, η_g^2 < .001, nor was the encoding task × test group interaction, *F*(2, 130) = 2.08, *p* = .14, η_g^2 = .01. Greenhouse-Geisser corrected *p* values are reported where appropriate to correct for violations of sphericity.

Figure 4

Mean JoL Ratings per Encoding Condition for Each Recognition Test



Note. The error bars represent the mean standard error.

3.4 Discussion

In this experiment, we examined the effects of JoLs in a TAP paradigm with word paired associates. There was a significant difference between the JoL ratings per encoding condition. Semantic word pairs (e.g., *leap-jump*) received the mean highest rating, followed by rhyme pairs (e.g., *dump-jump*), followed by unrelated pairs (e.g., *lie-jump*). This pattern suggests the participants successfully predicted the LoP effect. In contrast, the performance scores failed to demonstrate the LoP effect. In the standard recognition test, there was no significant difference between the semantic condition (related pairs, the deepest form of encoding) and the rhyme condition (rhyming pairs, the intermediary form of encoding). However, in the rhyme recognition test, the participants successfully recognised more rhymes of targets from the rhyming pairs than the related and unrelated pairs. Thus, we successfully replicated the TAP effect, albeit that the LoP component of the effect was not significant. Surprisingly, however, we saw no significant evidence of JoL reactivity, irrespective of the encoding condition (related, rhyme and unrelated pairs) or the type of recognition test (standard or rhyme).

3.4.1 JoL Reactivity

The present experiment failed to show JoL reactivity. We observed no significant difference in recognition performance between the JoL and No-JoL groups. This observation rejects our prediction that JoLs would attenuate the LoP effect in the standard recognition group and attenuate or enhance the TAP effect in the rhyme recognition group. This is a particularly surprising finding given the present experiment adopted a similar methodology to Myers et al., (2020: experiments 3 and 4). They reported positive reactivity from presenting related and unrelated word pairs before an item-recognition test. Other studies have also reported JoL reactivity with similar methodologies, such as word pairs with cued-recall tests (Janes et al., 2018; Mitchum et al., 2016; Myers et al., 2020; Rivers et al., 2021; Soderstrom et al., 2015; Tauber & Witherby, 2019) or word lists with old/new recognition tests (Li et al., 2021; Yang et al., 2015; Zechmeister & Shaughnessy, 1980). Therefore, it is surprising that the standard recognition test did not result in reactivity given the evidence of JoL reactivity within the literature. The same cannot be stated for the rhyme recognition test given that it was a novel addition to JoL reactivity research.

There was no clear methodological cause for the lack of reactivity in the standard recognition test. As discussed above, the procedure for the JoL participants in the standard recognition test was close to Myers et al.'s (2020) procedure, which reported positive reactivity. There were some methodological differences, such as the retention delay (5 minutes in Myers et al. vs 3 minutes in the present study), study time (12 seconds in Myers et al. vs 8 seconds in the present study), and the use of a within-subject design (Experiment 4 only). However, it is unlikely these differences would impact reactivity, given reactivity was reported in studies with less study time (e.g., 5 seconds: Senkova & Otani, 2021), a longer retention delay (e.g., 2 days: Witherby & Tauber, 2017), and a within-subject design (Rivers et al., 2021). Furthermore, the other results of the present experiment suggest sound methodology. The JoL ratings matched previous work showing greater JoL ratings for related word pairs (Arbuckle & Cuddy, 1969; McCabe & Soderstrom, 2011) and the rhyme recognition scores demonstrated the TAP effect. In sum, the lack of reactivity does not appear the result of methodological error.

The lack of reactivity may have been due to the participants' age. Tauber and Witherby (2019) consistently showed across five word pair experiments that JoL reactivity was present for young adults but not for older adults. Many studies which reported JoL reactivity used an undergraduate population (Janes et al., 2018; Mitchum et al., 2016; Witherby & Tauber, 2017), whereas the participants' age in the present study ranged from 18 and 59 years (*M* = 30.57 years, *SD* = 10.25 years). Our participants' older ages may have reduced the size of any putative JoL reactivity effect. Other metacognitive monitoring tasks are impacted by individual characteristics, such as confidence (Double & Birney, 2017) and working memory (Griffin et al., 2008). It may be that age, or other unrecorded characteristics, can explain our failure to see evidence of JoL reactivity.

The present experiment may have detected reactivity if a cued-recall test was used instead of the old/new recognition tests. Myers et al. (2020) compared JoL reactivity from word pairs on cued-recall, free-recall, and recognition tests. They predicted that JoLs would only boost cued-recall performance as this was the only test that required relational knowledge (knowledge of the cue-target *association*). Myers et al. reported positive reactivity in the cued-recall test and no reactivity in the free-recall test, conforming to their predictions. However, recognition tests produced a small positive reactivity effect. The authors suggested that recognition tests may allow a participant to retrieve other elements of the encoding context via the target word, such as the strengthened word pair relationship. This explanation explains the lack of reactivity for the free-recall test, for it does not provide the target or cue for retrieval of the encoding context. Given Myers et al.'s results, it is possible that the rhyme recognition test in the present experiment lacked the strength afforded by the original cue or target to result in reactivity (the rhyming word in the test was different to that used in the encoding context). However, it is surprising that there was no reactivity for the standard recognition test, especially given that other studies report positive JoL reactivity using recognition tests (Li et al., 2021; Myers et al., 2020; Tekin & Roediger, 2020; Yang et al., 2015). As discussed above, we do not have a ready explanation for these discrepancies. However, future research may benefit from changing the type of test within the TAP paradigm to cued-recall, given such tests reportedly result in greater reactivity (Myers et al., 2020).

Overall, we have no clear reason for why we did not observe any JoL reactivity in our experiment. Consequently, the data and analysis in the present experiment provide little evidence that JoLs result in reactivity. Since the advent of studies designed to primarily investigate JoL reactivity (Soderstrom et al., 2015), only four papers include at least one experiment failing to observe JoL reactivity (Dougherty et al., 2018; Mitchum et al., 2016; Myers et al., 2020; Robey et al., 2017), two of which used a novel paradigm with a recall attempt pre-JoL, limiting the generalisability of any conclusions on JoL reactivity (Dougherty et al., 2018; Robey et al., 2017). Thus, the present experiment contributes to the limited number of previous studies that have failed to find evidence of JoL reactivity in an experiment primarily designed to explore the reactive effects of JoLs.

3.4.2 Levels of Processing

We failed to observe a consistent LoP effect, rejecting our prediction in the introduction. In the standard recognition test, there was no significant difference in the related and rhyming word pairs. This result is surprising given that we expected the related pairs to elicit deep, semantic processing, resulting in greater recall compared to the rhyming pairs, which elicit shallower, phonemic processing. This result fails to conform to the established LoP literature. Craik and Lockhart (1972) presented participants with a series of target words, along with orienting questions that were designed to elicit different LoP. Across ten experiments, the authors found that targets from the semantic condition (deep processing) were recognised the best, followed by targets from the phonemic condition (intermediate processing) and, finally, the targets from the orthographic condition (appearance; shallow processing). More recently, Tekin and Roediger (2020) replicated Craik and Lockhart's LoP paradigm, but with half of the participants providing JoLs. The authors observed the classic LoP effect with JoL and no-JoL participants.

The present experiment, unlike Craik and Lockhart (1972) and Tekin and Roediger (2020), adopted word pairs to elicit the LoP effect rather than adopting orientating questions. This could lead to speculation that word pairs may fail to elicit the LoP effect. However, the present experiment observed greater recall for related word pairs than unrelated pairs in the standard recognition test; a finding common in the literature (Janes et al., 2018; McCabe & Soderstrom, 2011; Rivers et al., 2021). Thus, the present experiment and the word pair literature show that word pairs can produce deeper semantic processing, akin to the LoP effect. However, it is unclear why we did not observe semantic processing (related pairs) result in greater recall performance than phonemic processing (rhyming pairs). This is the first experiment to introduce rhyming pairs while attempting to elicit the LoP effect. Subsequently, future research could replicate the use of rhyming pairs to explore if rhyming pairs consistently achieve non-significantly different recall scores than related pairs.

3.4.3 Transfer Appropriate Processing

The present experiment showed the TAP effect, confirming our prediction in the introduction. In the rhyme recognition test, the participants successfully recognised more rhymes of targets from the rhyming pairs than the related and unrelated pairs. Our results mostly conform to the findings of Morris et al.'s (1977) seminal study. Morris et al. presented participants with orthographic, rhyme, and semantic encoding tasks via old/new questions for target words to induce LoP. The deepest processing (semantic) performed best on the standard old/new recognition test, but the intermediary processing (rhyme) performed best on the old/new rhyme recognition test. However, unlike Morris et al., the LoP component of the effect was not significant in the present experiment. Furthermore, we induced TAP using word pair relatedness rather than via encoding questions. To the best of the author's knowledge, the present experiment is the first to observe TAP in a word pair associate methodology.

3.4.4 Metacognitive Awareness of the LoP Effect

In the present experiment, we also explored if the participants could predict the LoP effect. We first confirmed that there was no significant difference between the JoL ratings of the standard and rhyme recognition test groups. This was expected as we did not tell participants in advance which test they would receive. Subsequently, we collapsed the results across test groups. We observed a significant difference between the JoL ratings per encoding condition. The participants gave the highest JoL ratings to semantic pairs, followed by rhyme pairs, followed by unrelated pairs. The result indicates that participants could predict the LoP effect. However, the participants did not know which test to expect. Knowledge of the type of test may influence JoLs (e.g., a participant may give a higher JoL to a rhyming pair if they know they will later receive a rhyme test). Future research may inform participants of the test format before the encoding phase if examining metacognitive awareness of LOP and TAP.

The JoL ratings in the present experiment are in line with Koriat's (1997) cue utilisation framework. The framework suggests that JoLs are more sensitive to intrinsic cues (e.g., the association between the cue and target) than extrinsic cues (e.g., presentation of the encoding task). Previous research has shown that participants base their JoLs on the intrinsic cue of word pair relatedness (Arbuckle & Cuddy, 1969; McCabe & Soderstrom, 2011). It therefore follows that related word pairs in the present experiment received the greatest JoL ratings.

Tekin and Roediger's (2020) exploration of JoLs in an LoP paradigm provided a different pattern of JoL ratings compared to the present experiment. Tekin and Roediger adopted an LoP paradigm, which included an encoding phase to induce semantic, phonemic, and orthographic processing with old/new congruent and incongruent orienting questions. The authors reported that participant JoL ratings did not differ between the conditions in two of the three experiments. In the remaining experiment, the JoL ratings were greatest for the deep processing task (semantic) but only for congruent tasks. Tekin and Roediger therefore concluded that participants were only somewhat aware of the LoP effect. Given that the present experiment reported JoL ratings consistent with the LoP effect, it can appear that Tekin and Roediger reported opposing findings. However, the difference may be more apparent than real. Tekin and Roediger propose that their participants’ JoL ratings also align with Koriat's (1997) cue utilisation framework. Tekin and Roediger suggest that the encoding questions are an extrinsic cue, changing the presentation of the encoding task. The participants may have been less aware that the encoding question impacted their learning. Consequently, the differences between the two studies are likely due to the intrinsic versus extrinsic cues informing the JoL.

3.4.5 Limitations

A limitation of the present experiment was that the participants did not self-report if they identified as having a specific language or reading difficulty (albeit they did report English as their first language). This is particularly pertinent due to the inclusion of rhyming stimuli. For example, dyslexia is a developmental reading disorder defined by a phonological deficit (Roelle et al., 2017). Subsequently, one may expect that the potentially dyslexic participant had greater difficulty with the phonemic processing elicited by the rhyme pairs compared to their non-dyslexic counterpart, introducing an unaddressed confounding variable. However, the random allocation of the participants should have mitigated this concern (i.e. the potential participants with a language or reading difficulty would have been distributed across the groups). Moreover, the observed TAP effect in the present study suggests that if such language or reading difficulties were present, they did not have such an effect as to prevent TAP. Nonetheless, future studies that collect participants' characteristics relating to language and reading difficulties may provide a more detailed insight into JoL reactivity.

A further limitation is the use of an international sample. The majority of the JoL research is based in the United States (e.g., Myers et al., 2020; Rivers et al., 2021; Senkova & Otani, 2021) and, at present, there is yet to be an exploration into JoL reactivity between cultures and nations. Consequently, it is possible that the present findings may have been influenced by the participants' geography. Future online studies need to collect participants' location of study and may look to restrict the sample to a specific geography until there is a better understanding of JoL reactivity across cultures.

3.4.6 Implications

It is possible that our results highlight a *publication bias* in the JoL reactivity literature (that is, a paper with significant results is more likely published than those with null results: Easterbrook et al., 1991). There is no direct evidence for this claim, given the nature of bias. However, the present study is one of few to report null findings from an experiment designed to primarily investigate JoL reactivity. This is in a cultural context in which publication bias remains an ongoing risk to the reproducibility of scientific findings (Marks-Anglin & Chen, 2020). Future researchers exploring JoL reactivity will need to take the necessary steps to mitigate against potential publication bias. This could include seeking publication based on a preregistered report (a submission of planned and unplanned research; see Nosek et al., 2019) to support the publication of potential null findings.

It is also possible that future research may allow for a new perspective on the present findings. This could be achieved via replication of the study. First, a replication could address that the present study was potentially underpowered. The power calculation for the experiment identified a need for 140 participants, but only 133 were included in the analysis. Second, replication is crucial in supporting the validity of outcomes, as already observed in the JoL reactivity literature. Dougherty et al. (2005) reported that JoLs provided memory benefits beyond that of a recall opportunity, although Dougherty et al. (2018) and Robey et al. (2017) observed no such effect from four replication experiments. Future replications of the present study may result in JoL reactivity, suggesting the present findings are a Type II error. Alternatively, a replication may achieve similar null results, thus requiring further experimentation to understand the factors that create the unique findings.

Lastly, JoL reactivity research is underdeveloped compared to other memory phenomena, and more research is required to understand when reactivity occurs (Myers et al. 2020). We speculated that the non-significant reactive effects in the present study could be due to the sample age range, or the type of test, but such speculation can only draw upon the limited research exploring JoL reactivity. Most JoL reactivity research draws upon a single paradigm (paired-associates: e.g., Janes et al., 2018; Myers et al., 2020; Rivers et al., 2021; Tauber & Witherby, 2019). There is no research into how JoL reactivity is impacted by participant characteristics other than age (Tauber & Witherby, 2019; Zhao et al., 2021), and only a single study has explored the impact of different types of tests (Myers et al., 2020). Future research should address these gaps and look to new paradigms, such as those used in the present experiment.

3.4.7 Conclusion

In conclusion, we observed the TAP effect in a TAP paradigm (Morris et al., 1977), with word paired associates. However, the experiment failed to produce significant JoL reactivity. The results contradict previous findings of JoL reactivity in paired-associate learning paradigms. Future research should continue to explore the interaction between TAP and JoL reactivity while broadening the understanding of the factors that determine reactive effects.

Open practices statement

The *R* code used for data screening and analyses in addition to applicable stimuli and data files are available on the Open Science Framework (OSF) at https://osf.io/ak57u/. The experiment was not preregistered.

Chapter 4 Thesis Conclusion

As an educational psychologist, the present thesis has been a fascinating exploration of the field of JoL reactivity. The systematic literature review and empirical study provided insight into the early stages of this cognitive phenomenon, and I hope that the present research will contribute to the growing literature. However, during the first three thesis chapters, I did not have the opportunity to explore the implications of the thesis for my field and area of interest: educational psychology. This brief concluding chapter reflects on my thesis's implications for education, learning, and educational psychology practice.

4.1 Education and Learning

During the thesis introduction (chapter 1), I shared that my motive for investigating JoLs was to inform good classroom practice. I explained that while JoLs for education is an exciting prospect, JoL reactivity research is in its infancy, and there is a need to develop a foundational understanding before exploring JoLs in the classroom. Following my systematic review and empirical research, I believe my findings make the need for a foundational understanding even more evident. The systematic review highlighted that very little is understood outside the increased relatedness effect (related words generate greater reactivity), and there is little consensus on the underlying mechanisms driving JoL reactivity. Furthermore, the empirical study did not observe reactive effects. To bring JoLs into the classroom (outside of a research capacity) would be to implement an approach without sufficient evidence or theoretical understanding.

The understanding of JoL reactivity is yet to reach the threshold before being recommended to educators. However, given that JoLs require minimal time and resources, could educators still implement JoLs on the chance to improve learning without risk? I argue against this position. The empirical study failed to observe JoL reactivity, including those participants using the standard word-relatedness paradigm adopted in studies resulting in reactivity (e.g., Myers et al., 2020; Soderstrom et al., 2015). In effect, the empirical study included a failed replication of the methods thought to produce reactive effects. If future research begins to challenge the presence of JoL reactivity, educators may inadvertently create ineffective time in the classroom. This could come at an opportunity cost for effective alternative strategies (such as testing). Furthermore, the systematic literature review highlighted that very little is understood about individual differences in the JoL reactivity literature. Previous work has shown participants with reduced working memory find metacognitive monitoring challenging while performing a comprehension task (Griffin et al., 2008). It is possible that implementing JoLs in the classroom may inadvertently privilege particular students. A JoL rating may be simple and require minimal resources, but viewing them as risk-free is a misconception.

Throughout the review and empirical chapters, I drew upon the JoL reactivity literature with adult participants. Given my motive to investigate if JoLs could inform classroom practice, I may have been best placed to explore JoL reactivity literature with children and young people. Such literature may have helped increase the utility of the thesis for education and the classroom. However, there is almost no such literature. While there is research that uses JoLs with children and young people (Baars et al., 2017; Bayard et al., 2021; Grainger et al., 2016; Halamish et al., 2018; Roebers et al., 2020) there is only one JoL reactivity paper (Zhao et al., 2021). Therefore, the systematic review by necessity had to focus on an adult population. Similarly, while I may have explored child and young participants for the empirical paper, we identified a gap in the literature (the TAP effect) which required building upon existing methods in JoL reactivity research. Consequently, we included adult participants (like in the previous JoL reactivity literature) to allow for a comparison between papers without a significant variation in the sample (child versus adult). The present research sought to inform the foundational understanding of JoL reactivity which future research can develop. Such research may provide the pre-requisite to an applied approach (such as bringing JoLs in the classroom) or research with a different sample (such as children).

4.2 Educational Psychology

The educational psychologist's role often includes assessing and recommending tools for learning. As discussed in the introduction (chapter 1), cognitive and educational psychology has provided empirical support for some classroom practices (such as testing for memory performance) while supporting the argument against others (such as highlighting). JoLs for education is still in its infancy. However, it provides an example of the necessary steps to explore a new educational tool and cautions against bold, unsubstantiated claims regularly made by those promoting educational resources. The JoL reactivity literature serves to highlight the depth and systematic progression to become evidence-based. An educational psychologist will benefit from understanding the lengths it takes to thoroughly explore a tool or cognitive phenomena, such as JoL reactivity, to better critique tools for learning.

The empirical paper in the present thesis (chapter 3) speculated that the observed findings highlight a publication bias in the JoL reactivity literature (that is, a paper with significant results is more likely published than those with null results: Easterbrook et al., 1991). There is no direct evidence for this claim, given the nature of bias. However, the study was one of few to report null findings from an experiment designed to primarily investigate JoL reactivity. This is in a cultural context in which publication bias remains an ongoing risk to the reproducibility of scientific findings (Marks-Anglin & Chen, 2020). Educational psychologists are in a position to influence future research and help mitigate potential publication bias. This could include seeking publication based on a preregistered report (a submission of planned and unplanned research; see Nosek et al., 2019) to support the publication of potential null findings. Similarly, future research may adopt the same practice in the present thesis and commit to sharing their materials and analysis on a publically available platform (such as the Open Science Framework: OSF).

4.3 Final Conclusion

The present thesis contributed to the JoL reactivity literature with a systematic literature review and empirical study. The review observed a growing consensus that JoLs produce positive reactivity with semantically related word pairs. We also observed that relational accounts of reactivity are most common in the literature but have inconsistent evidence. There are emerging non-relational accounts, but these are tentative frameworks. The empirical study investigated if JoL reactivity depends on whether the encoding and retrieval processes match (transfer appropriate processing). We failed to observe JoL reactivity, regardless of encoding or test condition. The study was the first to investigate JoL reactivity using a TAP paradigm with word pairs and provides a foundation for future work to examine the role of the test on JoL reactivity and JoL reactivity in alternative paradigms. In sum, despite the upsurge of JoL reactivity research since 2015, the field is still in its infancy. JoLs can not yet be recommended as a tool for education but remains a fascinating phenomenon deserving of future research.

Appendix A JBI Quality Appraisal Checklist

Note. Y = Yes; N = No; UC = Unclear; See question wording below.

First author	Year	1	2	3	4	5	6	7	8	9	10	11	12	13	Overall
DeYoung	2021	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Li	2021	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Rivers	2021	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Senkova	2021	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Myers	2020	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Tekin	2020	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Tauber	2019	Y	Y	Y	UC	N	UC	Y	Y	Y	Y	Y	Y	Y	Include
Dougherty	2018	Y	Y	UC	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Janes	2018	Y	Y	UC	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Robey	2017	Y	Y	UC	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Witherby	2017	Y	Y	UC	UC	N	UC	Y	Y	Y	Y	Y	Y	Y	Include
Mitchum	2016	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Soderstorm	2015	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Yang	2012	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Tauber	2012	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Dougherty	2005	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Kelemen	1997	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include
Zechmeister	1980	Y	Y	Y	UC	UC	UC	Y	Y	Y	Y	Y	Y	Y	Include

JBI critical appraisal questions:

1. Was true randomization used for assignment of participants to treatment groups?

2. Was allocation to treatment groups concealed?

3. Were treatment groups similar at the baseline?

4. Were participants blind to treatment assignment?

5. Were those delivering treatment blind to treatment assignment?

6. Were outcomes assessors blind to treatment assignment?

7. Were treatment groups treated identically other than the intervention of interest?

8. Was follow up complete and if not, were differences between groups in terms of their follow up adequately described and analysed?

9. Were participants analysed in the groups to which they were randomized?

10. Were outcomes measured in the same way for treatment groups?

11. Were outcomes measured in a reliable way?

12. Was appropriate statistical analysis used?

13. Was the trial design appropriate, and any deviations from the standard RCT design (individual randomization, parallel groups) accounted for in the conduct and analysis of the trial?

Appendix B Word List Stimuli

Target	Related	Rhyme	Unrelated	Test rhyme	Forward strength	Backward strength
BOSS	employer	toss	prune	LOSS	0.387	0.122
CAR	engine	tar	loose	FAR	0.359	0.011
CARD	poker	lard	moon	HARD	0.414	0.061
CAT	dog	fat	purse	BAT	0.667	0.513
CHART	graph	start	paste	TART	0.15	0.275
CLOTH	fabric	moth	please	SLOTH	0.381	0.115
CRASH	accident	mash	wire	DASH	0.128	0.132
CUT	trim	gut	noise	HUT	0.497	0
DAD	mom	bad	throat	LAD	0.759	0.71
DIRT	filth	flirt	inch	SKIRT	0.693	0.037
GLUE	sticky	clue	wrote	BLUE	0.185	0.371
HAND	glove	grand	warmth	BAND	0.552	0.048
HOUSE	home	spouse	muscle	MOUSE	0.333	0.582
JUMP	leap	dump	juice	PUMP	0.522	0.067
KNIFE	fork	life	tire	WIFE	0.37	0.327
LIE	fib	die	switch	PIE	0.816	0.066
LIGHT	dark	night	brush	MIGHT	0.428	0.371
MEN	women	ten	hide	PEN	0.45	0.614
MIX	blend	fix	rice	SIX	0.565	0.114
PAIN	agony	rain	wind	LANE	0.649	0.032
PUB	beer	rub	wing	TUB	0.518	0
RARE	scarce	care	waist	PAIR	0.134	0.021
RUG	carpet	jug	tool	MUG	0.248	0.468
RUN	walk	fun	yield	GUN	0.465	0.493
SHY	timid	fly	match	TRY	0.689	0.104
SMALL	shrink	tall	wool	BALL	0.486	0
SMOKE	cigarette	broke	height	CLOAK	0.449	0.323
SOCK	shoe	lock	voice	DOCK	0.212	0.617

STOP	pause	pop	point	DROP	0.427	0
TREE	bush	knee	shed	SKI	0.395	0.034

Appendix C Analysis of Corrected Hit Rate

We conducted 3 (Encoding task: semantic, rhyme, or unrelated) x 2 (Test type: standard target recognition or rhyme recognition) x 2 (Judgment type: JoL or no-JoL) mixed ANOVA on corrected hit rates.

The main effect of encoding task was not significant *F*(1, 129) = 0.13, *p* = .91, η_g^2 < .001. The test type x encoding task interaction was not significant *F*(1, 129) = 2.35, *p* = .18, η_g^2 = .009. The encoding task x judgment type interaction was not significant *F*(2, 258) = 1.23, *p* = .29, η_g^2 = .005. The test type x encoding task x judgment type interaction was not significant *F*(2, 258) = 0.10, *p* = .90, η_g^2 < .001.

The main effect of test group significant was significant *F*(1, 129) = 281.88, *p* < .001, η_g^2 = .53. The standard recognition group had a significantly corrected higher hit rate (*M* = 78.11, *SD* = 20.75) than the rhyme recognition group (*M* = 27.88, *SD* = 27.18).

The main effect of encoding task was significant, *F*(2, 258) = 5.36, *p* = .005, η_g^2 = .02. Paired samples *t*-tests revealed there was a significant difference in performance scores between the rhyme (*M* = 57.54, *SD* = 33.16) and unrelated condition (*M* = 49.42, *SD* = 31.32), *t*(132) = 3.26, *p* = .001. There was not a significant difference between the rhyme and semantic condition (*M* = 52.58, *SD* = 36.55), *t*(132) = 2.01, *p* = .046, or the semantic and unrelated condition, *t*(132) = 1.23, p = .22.

The test group x condition interaction was significant, *F*(2, 258) = 3.58, *p* = .03, η_g^2 = .01. Paired samples *t*-tests were performed to compare performance scores of the encoding conditions in each test type. For the rhyme recognition test, there was a significant difference between the rhyme (*M* = 35.05, *SD* = 27.36) and semantic condition (*M* = 23.54, *SD* = 25.33), *t*(65) = 2.94, *p* = .004, and the rhyme and unrelated condition (*M* = 25.05, *SD* = 27.75), *t*(65) = 2.51, *p* = .015. There was not a significant difference between the semantic and unrelated conditions: *t*(65) = 0.35, *p* = .73. For the standard target recognition test, there was a significant difference between the semantic and unrelated condition, *t*(66) = 2.82, *p* = .006. There was not a significant difference between the rhyme (*M* = 79.70, *SD* = 21.29) and semantic condition (*M* = 81.19, *SD* = 19.10), *t*(66) = 0.53, *p* = .60, and rhyme and unrelated condition (*M* = 73.43, *SD* = 21.26) within the standard test group: *t*(66) = 2.08, *p* = .04.

Appendix D Analysis of Hit Rate

We conducted a 3 (Encoding task: semantic, rhyme, or unrelated) x 2 (Test type: standard target recognition or rhyme recognition) × 2 (Judgment type: JoL or no-JoL) mixed ANOVA on the hit rates.

The main effect of encoding task was not significant *F*(1, 129) = 0.02, *p* = .89, η_g^2 < .001. The test type x encoding task interaction was not significant *F*(1, 129) = 0.70, *p* = .41, η_g^2 = .003. The encoding task x judgment type interaction was not significant *F*(2, 258) = 1.23, *p* = .29, η_g^2 = .005. The test type x encoding task x judgment type interaction was not significant *F*(2, 258) = 0.10, *p* = .90, η_g^2 < .001.

The main effect of test type significant was significant, *F*(1, 129) = 205.36, *p* < .001, η_g^2 = .45. The standard recognition group had a significantly higher hit rate (*M* = 82.86%, *SD* = 18.91%) than the rhyme recognition group (*M* = 40.40%, *SD* = 28.21%).

The main effect of encoding task was significant *F*(2, 258) = 5.36, *p* = .005, η_g^2 = .02. Paired samples *t*-tests revealed that there was a significant difference in performance scores between rhyming and unrelated pairs: *t*(132) = 3.26, *p* = .001. There was not a significant difference in performance scores between the rhyme (*M* = 66.17%, *SD* = 30.77%) and semantic condition (*M* = 61.20%, *SD* = 33.64%): *t*(132) = 2.01, p = .046, or the semantic and unrelated condition (*M* = 58.05%, *SD* = 31.32%): *t*(132) = 1.23, *p* = .22.

The test type x encoding task interaction was significant (2, 258) = 3.58, p = .03, η_g^2 = .01. Paired samples *t*-tests were performed to compare performance scores of the encoding conditions in each test type. For the rhyme recognition test, there was a significant difference between the rhyme (*M* = 47.58, *SD* = 28.67) and semantic condition (*M* = 36.06, *SD* = 27.00), *t*(65) = 2.94, *p* = .004, and the rhyme and unrelated condition (*M* = 37.58, *SD* = 27.96), *t*(65) = 2.51, *p* = .015. There was not a significant difference between semantic and unrelated conditions, *t*(65) = 0.35, *p* = .73. For the standard target recognition test, there was a significant difference between the semantic (*M* = 85.97, *SD* = 17.06) and unrelated condition (*M* = 78.21, *SD* = 18.98), *t*(66) = 2.82, *p* = .006. There was not a significant difference between the rhyme (*M* = 84.48, *SD* = 19.95) and semantic condition, *t*(66) = 0.53, *p* = .60, and the rhyme and unrelated condition, *t*(66) = 2.08, *p* = .04.

List of References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*(1), 126–131.

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review*, *33*(2), 693–712. https://doi.org/10.1007/s10648-020-09556-8

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2017). Effects of problem solving after worked example study on secondary school children's monitoring accuracy. *Educational Psychology*, *37*(7), 810–834. https://doi.org/10.1080/01443410.2016.1150419

Baars, M., van Gog, T., de Bruin, A., & Paas, F. (2018). Accuracy of primary school children's immediate and delayed judgments of learning about problem-solving tasks. *Studies in Educational Evaluation*, *58*(February), 51–59. https://doi.org/10.1016/j.stueduc.2018.05.010

Bayard, N. S., van Loon, M. H., Steiner, M., & Roebers, C. M. (2021). Developmental Improvements and Persisting Difficulties in Children's Metacognitive Monitoring and Control Skills: Cross-Sectional and Longitudinal Perspectives. *Child Development*, *92*(3), 1118–1136. https://doi.org/10.1111/cdev.13486

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory and Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

Bhaskar, R. (2008). *A realist theory of science*. Routledge.

Bhaskar, R., Danermark, B., & Price, L. (2017). *Interdisciplinarity and wellbeing: A critical realist general theory of interdisciplinarity*. Routledge.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Boland, A., Cherry, G. M., & Dickson, R. (2017). *Doing a systematic review* (2nd ed.). SAGE Publications Ltd.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684.

https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology*, *104*(3), 268–294. https://doi.org/10.4324/9781315440446

Creswell, J. . W. (2009). Mapping the field of mixed methods research. *Journal of Mixed Methods Research*, *3*(2), 95–108. http://journals.sagepub.com/doi/pdf/10.1177/1558689808330883

Çubukcu, F. (2008). How to enhance reading comprehension through metacognitive strategies. *Journal of International Social Research*, *1*(2), 83–93.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. https://doi.org/10.3758/s13428-014-0458-y

de Winstanley, P. A. (1996). Generation effects and the lack thereof: The role of transfer-appropriate processing. *Memory*, *4*(1), 31–48. 10.1080/741940667

DeYoung, C. M., & Serra, M. J. (2021). Judgments of learning reflect the Animacy advantage for memory, but not beliefs about the effect. *Metacognition and Learning*, *16*(3), 711–747. https://doi.org/10.1007/s11409-021-09264-w

Double, K. S., & Birney, D. P. (2017). Are you sure about that? Eliciting confidence ratings may influence performance on Raven’s progressive matrices. *Thinking and Reasoning*, *23*(2), 190–206. https://doi.org/10.1080/13546783.2017.1289121

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology*, *10*, 1–12. https://doi.org/10.3389/fpsyg.2019.02755

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory*, *26*(6), 741–750. https://doi.org/10.1080/09658211.2017.1404111

Dougherty, M. R., Robey, A. M., & Buttaccio, D. (2018). Do metacognitive judgments alter memory performance beyond the benefits of retrieval practice? A comment on and replication attempt of Dougherty, Scheck, Nelson, and Narens (2005). *Memory and Cognition*, *46*(4), 558–565. https://doi.org/10.3758/s13421-018-0791-y

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory and Cognition*, *33*(6), 1096–1115. https://doi.org/10.3758/BF03193216

Downs, S. H., & Black, N. (1998). The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *Journal of Epidemiology and Community Health*, *52*(6), 377–384. https://doi.org/10.1136/jech.52.6.377

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013a). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, Supplement*, *14*(1), 4–58. https://doi.org/10.1177/1529100612453266

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013b). What works, what doesn't. *Scientific American Mind*, *24*(4), 46–53.

Easterbrook, P. J., Gopalan, R., Berlin, J. A., & Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, *337*(8746), 867–872. https://doi.org/10.1016/0140-6736(91)90201-Y

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/a0021663

Grainger, C., Williams, D. M., & Lind, S. E. (2016). Judgment of learning accuracy in high-functioning adolescents and adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *46*(11), 3570–3582. https://doi.org/10.1007/s10803-016-2895-1

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 and Cognition*, *36*(1), 93–103. https://doi.org/10.3758/MC.36.1.93

Halamish, V., Nachman, H., & Katzir, T. (2018). The effect of font size on children's memory and metamemory. *Frontiers in Psychology*, *9*(AUG), 1–9. https://doi.org/10.3389/fpsyg.2018.01577

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 497–514. https://doi.org/10.1016/S0022-5371(81)90138-9

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin and Review*, *25*(6), 2356–2364. https://doi.org/10.3758/s13423-018-1463-4

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429–446). Erlbaum.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772–775. https://doi.org/10.1126/science.1204035

Karpicke, J. D., & Roediger, H. L. (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances Long-term retention. *Journal of Experimental Psychology: Learning Memory and Cognition*, *33*(4), 704–719. https://doi.org/10.1037/0278-7393.33.4.704

Karpicke, J. D., & Roediger, H. L. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151–162. https://doi.org/10.1016/j.jml.2006.09.004

Kelemen, W. L., & Weaver, C. A. (1997). Enhanced metamemory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning Memory and Cognition*, *23*(6), 1394–1409. https://doi.org/10.1037/0278-7393.23.6.1394

Klein, S. B., Loftus, J., Kihlstrom, J. F., & Aseron, R. (1989). Effects of item-specific and relational information on hypermnesic recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1192–1197. https://doi.org/10.1037/0278-7393.15.6.1192

Koriat, A., Nussinson, R., Bless, H., & Shaked, N. (2008). Information-based and experience-based metacognitive judgments: evidence from subjective confidence. In J. Dunlosky & R. A. Bjork (Eds.), *A handbook of memory and metamemory* (pp. 117–136). Psychology Press. https://doi.org/0278-7393/89/$00.7

Koriat, Asher. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, *126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: Management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407–411. https://doi.org/10.4097/kjae.2017.70.4.407

Lange, K., Kühn, S., & Filevich, E. (2015). “Just another tool for online studies” (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, *10*(6), 1–14. https://doi.org/10.1371/journal.pone.0130834

Li, B., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Liu, M., Yang, C., & Luo, L. (2021). Soliciting judgments of forgetting reactively enhances memory as well as making judgments of learning: Empirical and meta-analytic tests. *Memory and Cognition*, 1–17. https://doi.org/10.3758/s13421-021-01258-y

Lorch, R. F., Lorch, E. P., & Klusewitz, M. A. (1995). Effects of typographical cues on reading and recall of text. *Contemporary Educational Psychology*, *20*(1), 51–64. https://doi.org/10.1006/ceps.1995.1003

Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, *11*(6), 725–742. https://doi.org/10.1002/jrsm.1452

Marxen, D. E. (1996). Why reading and underlining a passage is a less effective study strategy than simply rereading the passages. *Reading Improvement*, *32*(2), 88–96.

Maxwell, N. P., & Huff, M. J. (2022). Reactivity from judgments of learning is not only due to memory forecasting: evidence from associative memory and frequency judgments. *Metacognition and Learning*, 1–37. https://doi.org/10.1007/s11409-022-09301-2

McCabe, D. P., & Soderstrom, N. C. (2011). Recollection-based prospective metamemory judgments are more accurate than those based on confidence: Judgments of remembering and knowing (JORKs). *Journal of Experimental Psychology: General*, *140*(4), 605–621. https://doi.org/10.1037/a0024014

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, *19*(4–5), 494–513. https://doi.org/10.1080/09541440701326154

Mitchum, A. L. (2011). *Reactive effects of memory performance predictions*. [Unpublished doctoral thesis]. Florida State University.

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General*, *145*(2), 200–219. https://doi.org/10.1037/a0039923

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533.

https://doi.org/10.1016/S0022-5371(77)80016-9

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(3), 859–871. https://doi.org/10.1037/xlm0000056

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory and Cognition*, *48*(5), 745–758. https://doi.org/10.3758/s13421-020-01025-5

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, and Computers*, *36*(3), 402–407. https://doi.org/10.3758/BF03195588

Nelson, T. O., & Dunlosky, J. (1991). When People's Judgments of Learning (JOLs) are Extremely Accurate at Predicting Subsequent Recall: The "delayed-JOL effect." *Psychological Science*, *2*(4), 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, Research and Evaluation*, *9*(6).

Paez, A. (2017). Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, *10*(3), 233–240. https://doi.org/10.1111/jebm.12266

Palmer, M. A., Sauer, J. D., Ling, A., & Riza, J. (2017). Caffeine cravings impair memory and metacognition. *Memory*, *25*(9), 1225–1234. https://doi.org/10.1080/09658211.2017.1282968

Peterson, S. E. (1991). The cognitive functions of underlining as a study technique. *Reading Research and Instruction*, *31*(2), 49–56.

Proflific. (2022). *What are the advantages and limitations of an online sample?* https://researcher-help.prolific.co/hc/en-gb/articles/360009501473-What-are-the-advantages-and-limitations-of-an-online-sample-

ProQuest. (2022). *Who We Are*. https://about.proquest.com/en/about/who-we-are/

R Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for

Statistical Computing. http://www.r-project.org/

Rhodes, M. G. (2016). Judgments of learning: Methods, data, and theory. In J. Dunlosky & J. Metcalfe (Eds.), *The Oxford Handbook of Metamemory* (pp. 90–117). Oxford University Press.

Rhodes, M. G., & Tauber, S. K. (2011). The Influence of Delaying Judgments of Learning on Metacognitive Accuracy: A Meta-Analytic Review. *Psychological Bulletin*, *137*(1), 131–148. https://doi.org/10.1037/a0021705

Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: support for the cue-strengthening hypothesis. *Memory*, *29*(10), 1342–1353. https://doi.org/10.1080/09658211.2021.1985143

Robey, A. M., Dougherty, M. R., & Buttaccio, D. R. (2017). Making retrospective confidence judgments improves learners' ability to decide what not to study. *Psychological Science*, *28*(11), 1683–1693. https://doi.org/10.1177/0956797617718800

Roebers, C. M., Kälin, S., & Aeschlimann, E. A. (2020). A comparison of non-verbal and verbal indicators of young children's metacognition. *Metacognition and Learning*, *15*(1), 31–49. https://doi.org/10.1007/s11409-019-09217-4

Roelle, J., Schmidt, E. M., Buchau, A., & Berthold, K. (2017). Effects of informing learners about the dangers of making overconfident judgments of learning. *Journal of Educational Psychology*, *109*(1), 99–117. https://doi.org/10.1037/edu0000132

Salomon, G. (1991). Transcending the qualitative-quantitative debate: The analytic and systemic approaches to educational research. *Educational Researcher*, *20*(6), 10–18. https://doi.org/10.3102/0013189X020006010

Schwartz, B. L., & Metcalfe, J. (2017). Metamemory: An update of critical findings. In J. H. Byrne (Ed.), *Learning and Memory: A Comprehensive Reference* (pp. 423–432). Elsevier.

Schwieren, J., Barenberg, J., & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning and Teaching*, *16*(2), 179–196. https://doi.org/10.1177/1475725717695149

Scotland, J. (2012). Exploring the philosophical underpinnings of research: Relating ontology and epistemology to the methodology and methods of the scientific, interpretive, and critical research paradigms. *English Language Teaching*, *5*(9), 9–16. https://doi.org/10.5539/elt.v5n9p9

Senkova, O., & Otani, H. (2021). Making judgments of learning enhances memory by inducing item-specific processing. *Memory and Cognition*, *49*(5), 955–967. https://doi.org/10.3758/s13421-020-01133-2

Siedlecka, M., Paulewicz, B., & Wierzchoń, M. (2016). But I was so sure! Metacognitive judgments are less accurate given prospectively than retrospectively. *Frontiers in Psychology*, *7*(218), 1–8. https://doi.org/10.3389/fpsyg.2016.00218

Soderstrom, N. C., & Bjork, R. A. (2014). Testing facilitates the regulation of subsequent study time. *Journal of Memory and Language*, *73*(1), 99–115. https://doi.org/10.1016/j.jml.2014.03.003

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning Memory and Cognition*, *41*(2), 553–558. https://doi.org/10.1037/a0038388

Son, L. K., & Metcalfe, J. (2000). Metacognitive and control strategies in study-time allocation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *26*(1), 204–221. https://doi.org/10.1037/0278-7393.26.1.204

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science*, *3*(5), 315–316. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Tauber, S. K., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Experimental Psychology*, *62*(4), 254–263. https://doi.org/10.1027/1618-3169/a000296

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging*, *34*(6), 836–847. https://doi.org/10.1037/pag0000376

Tekin, E., & Roediger, H. L. (2020). Reactivity of judgments of learning in a levels-of-processing paradigm. *Zeitschrift Fur Psychologie / Journal of Psychology*, *228*(4), 278–290. https://doi.org/10.1027/2151-2604/a000425

Trumbo, M. C. S., McDaniel, M. A., Hodge, G. K., Jones, A. P., Matzen, L. E., Kittinger, L. I., Kittinger, R. S., & Clark, V. P. (2021). Is the testing effect ready to be put to work? Evidence from the laboratory to the classroom. *Translational Issues in Psychological Science*, *7*(3), 332–355. https://doi.org/10.1037/tps0000292

Tufanaru, C., Munn, Z., Aromataris, E., Campbell, J., & Hopp, L. (2020). Systematic reviews of effectiveness. In E. Aromataris & Z. Munn (Eds.), *Joanna Briggs Institute Manual for Evidence Synthesis*. Joanna Briggs Institute. https://doi.org/10.46658/JBIMES-20-04

Welles, B. F. (2014). On minorities and outliers: The case for making big data small. *Big Data and Society*, *1*(1), 1–2. https://doi.org/10.1177/2053951714540613

Witherby, A. E., & Tauber, S. K. (2017). The influence of judgments of learning on Long-term learning and short-term performance. *Journal of Applied Research in Memory and Cognition*, *6*(4), 496–503. https://doi.org/10.1016/j.jarmac.2017.08.004

Witherby, Amber Elizabeth. (2016). *The direct effect of monitoring of learning on memory performance: How is it influenced by retention interval or judgment instructions?* [Unpublished doctoral thesis]. Colorado State University.

Yang, H., Cai, Y., Liu, Q., Zhao, X., Wang, Q., Chen, C., & Xue, G. (2015). Differential neural correlates underlie judgment of learning and subsequent memory performance. *Frontiers in Psychology*, *6*(1699), 1–12. https://doi.org/10.3389/fpsyg.2015.01699

Yeung, N., & Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1310–1321. https://doi.org/10.1098/rstb.2011.0416

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don’t. *Bulletin of the Psychonomic Society*, *15*(1), 41–44. https://doi.org/10.3758/BF03329756

Zhao, W., Li, B., Shanks, D. R., Zhao, W., Zheng, J., Hu, X., Su, N., Fan, T., Yin, Y., Luo, L., & Yang, C. (2021). When judging what you know changes what you really know: Soliciting metamemory judgments reactively enhances children’s learning. *Child Development*, 1–13. https://doi.org/10.1111/cdev.13689