MEMORY
2023, VOL. 31, NO. 7, 905–917
https://doi.org/10.1080/09658211.2023.2207804

Routledge
Taylor & Francis Group

Recognition, remember-know, and confidence judgments: no evidence of cross-contamination here!

Helen L. Williams[a], Glen E. Bodner [b] and D. Stephen Lindsay [c]

[a]School of Psychology, Keele University, Keele, UK; [b]Flinders University, Adelaide, Australia; [c]University of Victoria, Victoria, Canada

ABSTRACT

We report three experiments designed to reveal the mechanisms that underlie subjective experiences of recognition by examining effects of how those experiences are measured. Prior research has explored the potential influences of collecting metacognitive measures on memory performance. Building on this work, here we systematically evaluated whether cross-measure contamination occurs when remember-know (RK) and/or confidence (C) judgments are made after old/new recognition decisions. In Experiment 1, making either RK or C judgments did not significantly influence recognition relative to a standard no-judgment condition. In Experiment 2, making RK judgments in addition to C judgments did not significantly affect recognition or confidence. In Experiment 3, making C judgments in addition to RK judgments did not significantly affect recognition or patterns of RK responses. Cross-contamination was not apparent regardless of whether items were studied using a shallow or deep levels-of-processing task – a manipulation that yielded robust effects on recognition, RK judgments, and C. Our results indicate that under some conditions, participants can independently evaluate their recognition, subjective recognition experience, and confidence. Though contamination across measures of metamemory and memory is always possible, it may not be inevitable. This has implications for the mechanisms that underlie subjective experiences that accompany recognition judgments.

ARTICLE HISTORY
Received 13 December 2022
Accepted 19 April 2023

KEYWORDS
Remember-know; confidence judgments; recognition memory; subjective experience; metamemory

The remember-know (RK) task is commonly used to gauge the subjective experiences associated with recognition decisions (Tulving, 1985; Yonelinas, 2002). In a typical RK experiment, after studying a set of items participants complete a recognition test. In the common two-step procedure, they first decide whether a test item is "old" (studied) or "new" (not studied). For each item deemed old, they then further categorise their recognition experience as remember or know. *Remembering* is the experience of recollecting some details about the item's presentation during the study phase (e.g., thoughts, images, perceptual features). *Knowing* is often defined as recognition that is accompanied by a feeling of familiarity but unaccompanied by recollection of any details about the study experience. A third option, *guess*, is sometimes provided as well, to avoid know judgments being used for low-familiarity or strategic responding (Gardiner et al., 1996; Mäntylä, 1993).

Another approach to measuring how recognition is experienced is to ask for confidence judgments or ratings. The similarities between RK and confidence judgments have been widely debated (cf., Gardiner & Java, 1990; Haaf et al., 2021; Ingram et al., 2012; McCabe et al., 2009; Parks et al., 2011; Parks & Yonelinas, 2007; Smith et al., 2011; Wixted, 2007; Wixted & Mickes, 2010). Our study does not focus on whether RK and confidence judgments capture the same underlying processes. Rather, we sought to answer two straightforward methodological questions. The first was: Does asking participants to assess their subjective experience either through RK or confidence judgments affect recognition? To tackle this question, Experiment 1 compared old/new recognition across three groups: the *standard group* made no post-recognition judgments, the *RK group* made a remember/know/guess judgment for every item judged old, and the *C group* made a confidence judgment for each item judged old.

Why might the mere inclusion of a RK or confidence judgment influence recognition? We reasoned that asking people to assess their subjective experience or confidence might encourage them to consider their old/new recognition decisions more thoroughly, perhaps

CONTACT Helen L. Williams h.l.williams@keele.ac.uk School of Psychology, Keele University, Dorothy Hodgkin Building, Keele, Staffordshire ST5 5BG, UK

affecting their accuracy and/or response bias. The idea that task performance can be changed by inclusion of a self-report measure is referred to as "reactivity," and this has often been studied in the metacognitive domain (e.g., Double & Birney, 2017, 2018; Fox et al., 2011; Mitchum et al., 2016). Self-reported metacognitive judgments can be influenced by both information-based processes based on beliefs about one's own abilities and competencies, and experience-based processes such as cues resulting from subjective feelings that occur during a cognitive experience (Koriat et al., 2008). If confidence judgments or RK judgments direct a participant's attention toward cues that are more diagnostic of prior study, metacognitive monitoring should be enhanced (see Bodner & Lindsay, 2003). Consistent with this possibility, Double and Birney (2017, 2018) reported evidence that participants who made confidence ratings after providing their solution to problem-solving or reasoning tasks outperformed control participants who did not. They suggested that reactivity to the word "confidence" throughout the task facilitated metacognitive monitoring and thereby enhanced performance (Double & Birney, 2019).

Two studies that set out to gauge the effect of making RK judgments on recognition performance across one-step and two-step procedures have yielded different findings. Hicks and Marsh (1999) compared standard old/new recognition against both a two-step RK recognition task ("O/N then RK", where an old/new judgment is followed by a RK judgment for all "old" items) and a one-step RK recognition task (where a single-step *remember/know/new* judgment is made). The one-step task resulted in higher hits and higher false alarms (i.e., liberal response bias) compared to the standard old/new or two-step (O/N then RK) recognition conditions (for which bias was comparable and slightly conservative). In contrast, Mulligan et al. (2010) found that, compared to standard old/new recognition, inclusion of an RK or source judgment at test resulted in enhanced recognition for items shown in the same modality at study and test, but this occurred whether a one-step or two-step procedure was used.

Different outcomes have also been reported in two studies that examined the effect of making RK judgments on recognition in other tasks. Using the DRM paradigm (Roediger & McDermott, 1995), Smith et al. (2008) found that false recognition of critical items was reduced after visual presentation compared to auditory presentation in a one-step remember/know/guess/new task but was similar across presentation modalities in a standard old/new recognition task; correct recognition did not differ across task variants. However, in their one-step condition, Smith et al. did not count guesses as "old" responses when calculating hit and false alarm rates; had they done so, the difference in false alarms across tasks would have been modest at best. Naveh-Benjamin and Kilb (2012) asked younger and older participant groups to complete item (single word) and associative (paired words) recognition tests either with or without RK judgments. Typically, older adults display an associative deficit in comparison to younger adults. Inclusion of RK judgments eliminated this deficit. That is, associative task accuracy was boosted for older adults but not for younger adults when RK judgments were made.

In related work, Rotello et al. (2005) and Geraci et al. (2009) reported that differences in how RK response options are defined can also affect recognition. Rotello et al. (2005) defined remembering in a standard versus conservative way across groups and found that conservative instructions resulted in fewer hits being classified as remembered. Although differences for hits (.73 vs. .80) and false alarms (.39 vs. .48) across these conditions were not reliable, given their sample size of N = 24 per condition these differences would yield medium effect sizes, $d = .57$, 95% CI *Mdiff* [-.14, .003] and $d = .56$, 95% CI *Mdiff* [-.19, .006], based on Lakens (2013). Hence, their study may have been underpowered for detecting these effects. Thus, it remains possible that conservative remember instructions also led participants to be more conservative in their recognition decisions. Geraci et al. (2009) found a higher hit rate in a confidence judgment condition (sure vs. unsure) compared to an RK condition. However, they did not include a standard recognition task, thus it remains unclear whether the presence of RK judgments impaired recognition and/or whether the presence of confidence judgments improved recognition (cf. the problem-solving studies of Double & Birney, 2017, 2018, 2019). In addition, retrieval condition was varied within-subjects, with the RK testing session taking place one week before the confidence judgment session. Therefore it is also unclear whether the obtained pattern was due to practice with the experimental procedures and/or the inclusion of confidence judgments. We have also found that asking people to assess subjective experience can reduce hits and false alarms (Williams & Lindsay, 2016, 2022). However, these findings involved cross-experiment comparisons; inclusion of RK judgments was not the focus. Nonetheless, the patterns observed in these studies prompted us to revisit this issue in the present work.

In Experiment 1, we aimed to resolve the question of whether post-recognition judgments influence recognition. To this end, we compared old/new recognition across three groups: the *standard group* made no post-recognition judgments, the *RK group* made a remember/know/guess judgment for every item judged old, and the *C group* made a confidence judgment for each item judged old.

Our second research question was whether there is "cross contamination" in situations where both RK and confidence judgments are collected. This possibility has been noted by several researchers; for example: " … asking multiple orthogonal questions in sequence is likely to cause confusion and allow participants to blur the questions together so that decisions on confidence and RK are not independent" (Migo et al., 2012, p. 1442; see also Bruno & Rutherford, 2010; Holmes et al., 1998; Humphreys et al., 2003; Yonelinas, 2001). Researchers

have employed specific designs to ensure that the two judgment types could not influence one another. For example, comparing subjective experience and confidence across separate groups of participants, separate experiments, or with a week delay between judgment conditions (e.g., Gardiner & Java, 1990; Rajaram, 1993; Rajaram et al., 2002; Yonelinas, 2001). However, one experiment has examined the influence of RK and confidence judgments on each other. Sommer et al. (2021) compared recognition and subjective experiences across five response conditions: Confidence (C; 1 "very sure new" to 6 "very sure old"), RK-1-step (RKN), RK-2-step (O/N then RK), C + RK, and RK + C. False alarms were higher when RK judgments preceded confidence ratings, due to more liberal responding for items given know judgments in that condition compared to the others. However, in both combined-judgment conditions (RK + C and C + RK), the initial judgment was a 1-step judgment that *combined* the RK judgment or confidence rating with the recognition decision; thus their design did not include a "pure" comparison of how RK and confidence judgments influence each other when made *following* a recognition decision.

When might cross-contamination of RK and confidence ratings occur? One possibility is that RK judgments may be influenced by the presence of confidence ratings, but not the converse. This pattern might arise if confidence judgment instructions are better understood or adhered to than RK instructions (Geraci et al., 2009). Another possibility is that confidence ratings may be influenced by the presence of RK judgments, but not the converse. This pattern might arise given that participants are more conservative in making remember responses compared to high/strong/sure confidence responses (Dunn, 2004; Gardiner & Java, 1990; Geraci et al., 2009; Haaf et al., 2021). Thus, the presence of RK instructions might lead participants to adjust their confidence to bring it in line with their reported subjective experience. This possibility follows from Gardiner's (2001) claim that "it is surely the subjective state of awareness that gives rise to confidence in memory, not confidence that gives rise to the state of awareness" (p. 1356). We explored these two possibilities in Experiments 2 and 3. In Experiment 2, we compared confidence judgments across a *C group* (as per Experiment 1) and an *RK + C group* who made a remember/know judgment and a confidence judgment for each recognised item. The design of Experiment 3 was the reverse of Experiment 2; here we compared remember/know judgments across an *RK group* (as per Experiment 1) and an *RK + C group* (as per Experiment 2).

In all three experiments we also manipulated levels of processing (LOP) at encoding; LOP was varied within-participants in Experiment 1 and between-participants in Experiments 2 and 3. LOP instructions were primarily included at encoding so that participants had to make a judgment for each target item and therefore pay attention to the experiment. We did not have any strong theoretical predictions for how LOP might affect the impact of measurement variants on recognition but some previous research suggests that it could do. The deeper the LOP at encoding, the more semantic and contextual information is likely available for retrieved items at test (e.g., Gardiner, 1988). Indeed, deeper LOP increases both overall recognition and rate of remember responses (e.g., Gardiner, 1988; Gardiner et al., 1996; Perfect et al., 1995; Rajaram, 1993). Research suggests that LOP can affect how strict or lenient a participant is when assigning remember judgments to recognised items (Bodner & Lindsay, 2003; Tousignant et al., 2015; Tousignant & Bodner, 2012; Williams & Lindsay, 2019). If test-list context influences how participants define the subjective experience response options for themselves during the task (Bodner & Lindsay, 2003), then the mere presence of judgments in the task context might also influence participants' recognition decisions, and perhaps differentially for deep versus shallow LOP items. On the one hand, making post-recognition judgments could enhance recognition of deeply encoded items preferentially because participants would know what kinds of cues they should access from memory to support a remember response. On the other hand, making post-recognition judgments could improve recognition of shallowly encoded items by inducing participants them to consider aspects of each item that are not otherwise considered for shallowly encoded items (e.g., thoughts arising during encoding).

Experiment 1: Does making RK or C judgments affect recognition?

Experiment 1 tested our first research question: Does asking participants to assess their metacognitive recognition experience either through RK or C judgments affect recognition? We compared old/new recognition across three groups: The *standard group* made no post-recognition judgments, the *RK group* made a remember/know/guess judgment for every item judged old, and the *C group* made a confidence judgment for each item judged old. Half the items were studied under shallow encoding instructions and half under deep encoding instructions; deep and shallow items were intermixed with new items at test.

Method

Design and participants

A mixed design was used, with encoding condition (shallow vs. deep) as the within-subjects factor and test group (standard vs. RK vs. C) as the between-subjects factor. Table 2 provides the *N*s for each condition. Participants were excluded if their hit or false alarm rate suggested they misunderstood the instructions or were guessing (*z*-scores of >±3; criteria set prior to data collection; $n = 7$). This left 151 participants for analysis (72 female; mean age = 25.51 years, SD = 9.86). This sample gave us a priori power of .92 to detect a medium effect of test group (Cohen's f = .25; G*Power 3.1.5; Faul et al., 2007).

Stimuli

Stimuli were medium-frequency 5- to 7-letter words from the MRC Psycholinguistic database (mean familiarity rating of 427, range 400-480); 24 words were randomly allocated to each of four lists. Each participant studied two lists, one under shallow encoding instructions and one under deep encoding instructions. The other two lists served as lure items on the recognition test. Use of lists as target or lure stimuli was counterbalanced across participants. Two filler items were shown at the start (primacy buffers) and end (recency buffers) of each study list; thus, in total 48 targets were studied plus 8 fillers. To acquaint participants with the test procedure, 4 studied fillers were intermixed with 4 lure fillers at the start of the recognition test; these were not analyzed. The stimuli are available online (https://osf.io/hf38m/).

Procedure

The experiment was approved by the Keele University Ethics Review Panel (Ref: ERP384). The data were collected online using Qualtrics. The instructions informed participants that they would study two lists of words, each using a different task, for a later memory test. The instructions then explained and provided examples of the shallow encoding task ("does this word contain the letter a?"; response: *yes* or *no*) and deep encoding task ("how pleasant is this word?"; response: a rating between 1 "*not pleasant*" and 6 "*very pleasant*"). Task order was counterbalanced across participants. Participants were reminded of the encoding task prior to each list. Item presentation order was randomised for each participant. Each item was preceded by a fixation point "+" for 1 s. On-screen buttons appeared with each item and participants used their mouse to respond. Responses were self-paced.

After the second list, participants completed a 12-trial distractor task (mental rotation). Recognition test instructions were then presented. Participants were informed that half the words on the test had appeared on one of the two study lists and the rest were new words that had not been presented for study. Their task was to decide whether each word was "old" (studied) or "new" (not studied); examples were provided. Participants in the RK and C groups were further instructed that if they thought the word was an "old" word they would make a second judgment. Instructions and an example screen appropriate to their group were presented.

The RK group was instructed to categorise their recognition as remember, know, or guess, based on the definitions shown in Table 1. They were told that reminders of the definitions would be shown at the bottom of each page, but that they should try to learn them so that they could make their judgments quickly and easily. The C group was instructed to rate their confidence and were shown an example item with a confidence scale (1 = "not at all confident" to 7 = "extremely confident"). Recognition test items were randomised, presented, and responded to as per the study phase items. For all phases of the experiment participants were instructed to respond as quickly as possible while remaining accurate.

Table 1. Response options and definitions in the RK group in Experiment 1.

Response	Definition
Remember	You have an experience of recollection for the word. This could include being consciously aware of some aspect or aspects of what was experienced at the time the word was presented in the learning phase (e.g., aspects of the physical appearance of the item, or of something that happened in the room, or of what you were thinking or doing at the time). In other words, you should choose "Remember" if you have a sense of yourself in the past and/or the word brings back to mind a particular association, image, or thought, from the time of study. *For example, if you see someone on the street you may think "Who is that? Oh yes, it's the person I saw in line in the book store, I remember thinking what a funny hat they had on … "*
Know	You feel that you just know that the word was a word you saw in the learning phase, or you have a feeling of familiarity for the word, but you cannot consciously recollect anything about its actual occurrence or what was experienced at the time of its occurrence. In other words, you should choose "Know" if the word feels familiar or if you know the item was one you studied but you cannot recollect any details associated with seeing it before. *For example, if you see someone on the street you may think "Who is that? I know I've seen that person before, but I don't recall where that would have been … "* or you may think *"They look very familiar … I don't know where I know them from but they seem familiar … "*
Guess	You do not have any memories or feelings associated with the word and you are simply guessing that the word was one of the words you saw in the learning phase.

Results and discussion

Our Supplementary Analyses (see https://osf.io/hf38m/) provide the proportion of hits and false alarms by subjective experience response (RK group) and confidence level (C group). Analyses confirmed that our participants used these post-recognition responses appropriately. In the RK group, remember judgments were more frequent for deep than shallow hits (.74 vs. .50), $t(52) = 5.79$, $p < .001$, $d = 0.80$, whereas the reverse was true for know judgments (.22 vs. .33), $t(52) = 3.11$, $p = .003$, $d = 0.43$, consistent with prior findings (Gardiner et al., 1996; Perfect et al., 1995; Rajaram, 1993). In the C group, confidence was higher for deep than shallow hits (6.39 vs. 5.01), $t(45) = 10.79$, $p < .001$, $d = 1.59$.

Our main analyses examined whether making RK or C judgments influenced recognition in terms of hits, false alarms, discrimination (d'), and/or response bias (c); means shown in Table 2. Because there was only one false alarm rate per participant, d' and c were calculated across the whole set of shallow and deep items. The Snodgrass and Corwin (1988) *1/2N* correction was employed for false alarm rates of 0 or hit rates of 1 in a given condition. Eta-squared (η^2) is reported as a measure of effect size.

Hit rates were analysed in a 2 (encoding condition: shallow vs. deep) x 3 (test group: standard vs. RK vs. C) mixed-factor ANOVA. An LOP effect reflected more hits

Table 2. Means [between-subjects 95% CIs] of recognition performance measures by test group, Experiment 1.

Test group	N	Deep proportion hit	Shallow proportion hit	Proportion false alarms	Discrimination (*d'*)	Response bias (*c*)
Standard	51	.92 [.89, .95]	.61 [.55, .68]	.15 [.12, .18]	1.87 [1.68, 2.05]	.14 [.03, .25]
C	46	.90 [.88, .93]	.63 [.57, .70]	.13 [.09, .16]	2.04 [1.84, 2.23]	.18 [.07, .30]
RK	54	.91 [.88, .93]	.61 [.55, .68]	.14 [.11, .18]	1.91 [1.73, 2.09]	.18 [.07, .28]

following deep than shallow encoding, $F(1, 148) = 309.01$, $MSE = 0.020$, $p < .001$, $\eta^2 = .68$. In contrast, the hit rate did not differ significantly across test groups, $F(2, 148) = 0.053$, $MSE = 0.041$, $p = .95$, $\eta^2 = .001$, and the interaction with encoding condition was not significant, $F(2, 148) = 0.40$, $MSE = 0.020$, $p = .67$, $\eta^2 = .002$. A one-way ANOVA indicated that the false alarm rate also did not differ significantly test groups, $F(2, 148) = 0.68$, $MSE = 0.013$, $p = .51$, $\eta^2 = .009$. For the signal-detection measures, analogous one-way ANOVAs indicated that neither discrimination (*d'*) nor response bias (c) differed significantly across groups, $F(2, 148) = 0.79$, $MSE = 0.46$, $p = .45$, $\eta^2 = .011$, and $F(2, 148) = 0.17$, $MSE = 0.15$, $p = .85$, $\eta^2 = .002$, respectively.

Bayes factors (BFs) were used to assess the strength of evidence for these results. For hits, a Bayesian ANOVA (using JASP version 0.11.1; JASP Team, 2019) compared the strength of evidence for models assuming the following effect(s) against a model assuming only null effects: 1) encoding-only, 2) test-only, 3) encoding + test, 4) encoding + test + interaction. Each model produces a Bayes factor, which quantifies the relative strength of evidence for that model in comparison to the null model. The ratio of the BFs from the best-fitting model vs. next-best model allows us to quantify the degree of support for the best-fitting model. The encoding-only model best predicted the data ($BF_{10} = 4.73 \times 10^{39}$). The next best model was the encoding + test model ($BF_{10} = 3.04 \times 10^{38}$). However, the encoding-only model was preferred over the encoding + test model by a Bayes factor of 15.59, providing *strong* evidence that hits were influenced by encoding condition but not by test group (classification specified by Wagenmakers et al., 2018). For false alarms, *d'*, and *c*, a model assuming an effect of test group was compared against the null model, and there was *strong* to *moderate* evidence that test group did not influence false alarms ($BF_{01} = 8.61$), *d'* ($BF_{01} = 7.79$), or *c* ($BF_{01} = 13.20$). In sum, assessing confidence or subjective experience after each "old" recognition decision did not alter recognition for items studied at either a shallow or deep level of encoding.

Experiment 2: Does making RK judgments affect recognition and/or C judgments?

Experiments 2 and 3 tested our second research question: When both RK and C judgments are collected, does one type of judgment influence the other? Experiment 2 compared confidence judgments across a *C group* (as per Experiment 1) and an *RK + C group* who made a remember/know judgment and a confidence judgment for each recognised item. This design enabled us to evaluate whether making both RK and C judgments affects recognition relative to when only C judgments are made, and also whether making RK judgments influences C judgments.

Method

Design and participants

Different from Experiment 1, Experiment 2 used a fully between-subjects design with encoding condition (shallow vs. deep) and test group (RK + C vs. C) as the factors. University of Victoria undergraduates participated for bonus credit. Participants were excluded if their hit or false alarm rates suggested they misunderstood the instructions or were guessing (*z*-scores of >±3; $n = 4$). This left 195 participants for analysis (148 female; mean age = 20.63 years, $SD = 3.68$, range = 18-40). Assignment to groups was randomised, resulting in the Ns shown in Table 4. This sample gave us a priori power of .94 to detect a medium effect of test group (Cohen's f = .25; G*Power 3.1.5; Faul et al., 2007).

Stimuli

Stimuli were medium-frequency 5–8 letter words from the MRC Psycholinguistic database (mean familiarity rating of 424, range = 350-480); 56 words were randomly allocated to two lists. Use of lists as target or lure stimuli was counterbalanced across participants. Filler items buffered the study lists, as per Experiment 1. To acquaint participants with the test procedure the 4 studied fillers were intermixed with 4 lure fillers at the start of the recognition test; fillers were not analyzed. Stimuli are available online (https://osf.io/hf38m/)

Procedure

Experiments 2 and 3 received ethical approval from the University of Victoria Human Research Ethics Office (Ref:12-503). Participants were tested individually, and the experiment was run on E-Prime version 2.0. Instructions were provided on screen for either the shallow encoding task ("does the word contain the letter 'a'?"; response: *yes* or *no*) or the deep encoding task ("is the word pleasant?"; response: *yes* or *no*). During the study phase, target words were presented individually, preceded by a fixation cross "+" for 750 ms. Responses were made using number keys (1 = *yes*, 2 = *no*), and participants had 2 s to respond. Item presentation order was randomised for each participant.

After the study phase, participants completed two brief distractor tasks (speed of processing). The old/new recognition judgment instructions were then presented, as per Experiment 1. Participants in the RK + C group were then

Table 3. Response options and definitions in the RK + C groups in Experiments 2 and 3.

Response	Definition
R	You have an experience of **Remembering** the word. This could include seeing the word in your mind's eye, recollecting something that you thought or pictured when you saw the word on the original list, and/or having a sense of yourself in the past. For example, if you see someone on the street you may think *"Who is that? Oh yes, it's the person I saw in line in the book store, I remember thinking what a funny hat they had on … "*
K	You feel that you just **Know** that the word was on the previous list without any of the other feelings associated with vividly remembering that you have seen the word before. *For example, if you see someone on the street you may think "Who is that? Oh yes, it's my friend Rob".*
F	You have a feeling of **Familiarity** with the word and because of that you think that the word was one you saw on the study list. *For example, if you see someone on the street you may think "Who is that? They look very familiar … I don't know where I know them from but they are definitely familiar … "*
G	You had no feelings of familiarity or any other memories associated with the word and simply **Guessed** that the word had been on the previous list.

instructed to make RK and confidence judgments for each word deemed "old". For the RK judgment, they were asked 'What is your EXPERIENCE of recognizing this word?' and they classified their experience as R (for "remember"), K (for "know"), F (for "familiar"), or G (for "guess") as described in Table 3; these definitions were provided on paper for reference during the test.[1] For the confidence judgment, they were asked to rate their confidence on a scale of 1 "not confident at all" to 7 "extremely confident." Judgment order was counterbalanced across items, and this assignment was counterbalanced across participants.[2] The C group only judged confidence.

Each test trial began with a fixation cross "+" for 750 ms. A word was then presented above the cues "new (press "1" key)" and "old (press "2" key)". After an "old" response, participants made their RK and/or C judgments by clicking a response box on the screen using the mouse. In between responses, brief variable blank intervals were inserted to vary the lag between judgments (for purposes relevant to a separate research project). Participants were instructed to respond as quickly as possible while remaining accurate.

Results and discussion

We were interested in whether recognition measures differed across groups, and more so in whether assessing one's recognition experience alongside confidence judgments changes confidence ratings. Following Williams et al. (2022), we excluded judgments (old/new, RK, or C) made faster than 300 ms or slower than 8 s (> 0.5% per judgment).

Recognition

Analyses followed Experiment 1 except where noted. We first examined whether making remember-know judgments influenced recognition in terms of hits, false alarms, discrimination (d'), and/or response bias (c); the means are shown in Table 4. For each measure we conducted a 2 (encoding condition: shallow vs. deep) x 2 (test group: C vs. RK + C) between-subjects ANOVA.

The ANOVAs yielded robust main effects of encoding on hits and d' (deep > shallow), false alarms (shallow > deep), and c (more conservative after deep than shallow encoding), respectively: $F(1, 191) = 213.09$, $MSE = 0.016$, $p < .001$, $\eta^2 = .53$; $F(1, 191) = 259.39$, $MSE = 0.280$, $p < .001$, $\eta^2 = .57$; $F(1, 191) = 15.35$, $MSE = 0.011$, $p < .001$, $\eta^2 = .074$; and $F(1, 191) = 27.58$, $MSE = 0.155$, $p < .001$, $\eta^2 = .13$. In contrast, the main effect of test group was not significant for hits, $F(1, 191) = 0.35$, $p = .56$, d', $F(1, 191) = 1.81$, $p = .18$, false alarms, $F(1, 191) = 0.59$, $p = .45$, or c, $F(1, 191) = 0.126$, $p = .72$. Similarly, test group did not interact with encoding for hits, $F(1, 191) = 1.32$, $p = .25$, d', $F(1, 191) = 0.22$, $p = .64$, false alarms, $F(1, 191) = 1.64$, $p = .20$, or c, $F(1, 191) = 1.79$; $p = .18$. The corresponding Bayesian ANOVA model comparisons are provided in Table 5. These analyses provided *moderate* support for the conclusion that only encoding influenced these recognition measures. In sum, recognition was similar whether participants made both remember-know and confidence judgments or only confidence judgments after "old" decisions.

Confidence judgments

We next examined whether making RK judgments influenced confidence judgments. To this end, we conducted separate 2 (encoding condition: shallow vs. deep) x 2 (test group: C vs. RK+C) between-subjects ANOVAs on mean confidence for both hits and false alarms[3]; the means are shown in Figure 1. Confidence judgments were made on a scale of 1-7.

For hits, there was a significant main effect of encoding, reflecting higher confidence after deep than shallow encoding, $F(1, 191) = 178.14$, $MSE = 0.523$, $p < .001$, η^2

Table 4. Means [between-subjects 95% CIs] of recognition performance measures by encoding and test groups, Experiments 2 and 3.

Experiment	Encoding condition	Test group	N	Proportion hit	Proportion false alarms	Discrimination (d')	Response bias (c)
Exp. 2	Shallow	C	50	.57 [.53, .61]	.16 [.13, .19]	1.24 [1.10, 1.39]	.43 [.32, .54]
		RK + C	48	.61 [.57, .64]	.17 [.14, .20]	1.31 [1.16, 1.46]	.38 [.27, .49]
	Deep	C	48	.86 [.82, .90]	.12 [.09, .15]	2.43 [2.28, 2.60]	.06 [-.05, .17]
		RK + C	49	.85 [.81, .89]	.09 [.06, .12]	2.57 [2.42, 2.72]	.16 [.05, .27]
Exp. 3	Shallow	RK	50	.63 [.60, .66]	.22 [.19, .25]	1.18 [1.06, 1.31]	.25 [.15, .35]
		RK + C	48	.61 [.57, .64]	.17 [.14, .20]	1.35 [1.22, 1.48]	.39 [.29, .49]
	Deep	RK	48	.83 [.80, .87]	.10 [.07, .13]	2.37 [2.25, 2.50]	.17 [.07, .27]
		RK + C	47	.85 [.81, .88]	.10 [.07, .13]	2.37 [2.25, 2.50]	.13 [.02, .23]

Table 5. Results of 2 (encoding condition) x 2 (test group) Bayesian ANOVAs on each recognition measure in Experiment 2.

Measure	Best model	BF_{10} = for best model	Next best model	BF_{10} = for next best model	BF_{10} = for model comparison	Evidence classification
Hits	Encoding-only	4.47×10^{29}	Encoding + test	8.32×10^{28}	5.38	Moderate evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
False alarms	Encoding-only	170.01	Encoding + test	33.63	5.06	Moderate evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
d'	Encoding-only	1.23×10^{34}	Encoding + test	4.41×10^{33}	2.80	Anecdotal evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
c	Encoding-only	33,262.25	Encoding + test	5,753.68	5.78	Moderate evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
Mean confidence Hits	Encoding-only	1.04×10^{26}	Encoding + test	1.68×10^{25}	6.20	Moderate evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
Mean confidence FAs	Test-only	0.300	Encoding-only	0.238	1.26	Moderate evidence for null model

Note: Evidence classification specified by Wagenmakers et al. (2018).

= .48. The main effect of test group was not significant, $F(1, 191) = 0.15$, $p = .70$, nor was the interaction, $F(1, 191) = 0.08$, $p = .78$. For false alarms, confidence did not differ significantly across encoding conditions, $F(1, 186) = 0.82$, $MSE = 1.18$, $p = .37$, $\eta^2 = .004$, or test groups, $F(1, 186) = 1.56$, $p = .21$, but here the interaction was significant, $F(1, 186) = 4.42$, $p = .037$, $\eta^2 = .023$, see Figure 1. For the shallow encoding condition, confidence ratings for false alarms were equivalent across the C and RK + C groups, $t(96) = 0.63$, $p = .53$, but for the deep encoding condition, confidence ratings for false alarm were higher in the C group than in the RK + C group, $t(90) = 2.28$, $p = .025$, $d = .48$. The corresponding Bayesian analyses produced *moderate* evidence that only encoding influenced confidence for hits, and *moderate* evidence that the null model was the best fit for confidence in false alarms (see Table 5). In sum, the effects of encoding were robust but effects of test group were generally absent for both recognition and confidence, and making RK judgments did not significantly affect participants' confidence.

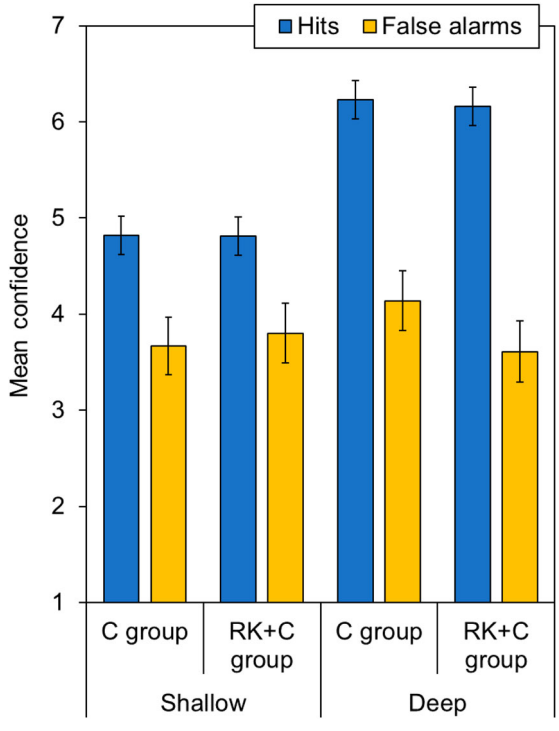

Figure 1. Mean confidence for hits and false alarms by encoding condition and test group in Experiment 2. Error bars show 95% confidence intervals.

Experiment 3: Does making C judgments affect recognition and/or RK judgments?

In Experiment 2, RK judgments appeared not to influence recognition or confidence. In Experiment 3 we tested the converse, namely whether (1) making confidence judgments influences RK judgments, and (2) whether making both confidence and RK judgments affects recognition relative to when only RK judgments are made. After each "old" recognition response, participants either made both RK and C judgments (RK + C group) or made only RK judgments (RK group). Their recognition and patterns of recognition experiences were compared. To establish generality, we again varied LOP at encoding.

Method

Design and participants

As in Experiment 2, we used a between-subjects design with encoding condition (shallow vs. deep) and test

Table 6. Results of 2 (encoding condition) x 2 (test group) Bayesian ANOVAs on each DV in Experiment 3.

DV	Best model	BF_{10} for best model	Next best model	BF_{10} for next best model	BF_{10} for model comparison	Evidence classification for model comparison and best model
Hits	Encoding-only	5.90×10^{24}	Encoding + test	9.40×10^{23}	6.28	Moderate evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
FAs	Encoding-only	253,106.24	Encoding + test	83,863.49	3.02	Moderate evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
d′	Encoding-only	2.56×10^{37}	Encoding + test	9.13×10^{36}	2.80	Anecdotal evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
c	Encoding-only	21.36	Encoding + test	4.94	4.32	Moderate evidence for encoding-only model being the best model. Encoding-only = strong support for H_1
Remember Hits	Encoding-only	32,142.65	Encoding + test	13,451.46	2.39	Anecdotal evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
Know Hits	Encoding-only	3.91	Encoding + test	0.77	5.07	Moderate evidence for encoding-only model being the best model. Encoding-only = moderate support for H_1
Familiar Hits	Encoding-only	1.36×10^{19}	Encoding + test	3.87×10^{18}	3.50	Moderate evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
Guess Hits	Encoding-only	2,454.56	Encoding + test	434.64	5.65	Moderate evidence for encoding-only model being the best model. Encoding-only = extremely strong support for H_1
Remember FAs	Null	1	Test-only	0.18	5.56	Moderate evidence for null model
Know FAs	Null	1	Test-only	0.23	4.39	Moderate evidence for null model
Familiar FAs	Encoding-only	1.01	Encoding + test	0.17	6.01	Moderate evidence for encoding-only model being the best model. Encoding-only = anecdotal support for H_1
Guess FAs	Encoding-only	1.29	Encoding + test	0.24	5.38	Moderate evidence for encoding-only model being the best model. Encoding-only = anecdotal support for H_1

Note: Evidence classification specified by Wagenmakers et al. (2018).

group (RK + C vs. RK) as the factors. Participants were University of Victoria undergraduates who participated for bonus credit. Data sets were excluded from analysis if proportion of hits or false alarms (FAs) suggested participants had not understood the instructions or had been guessing (*z*-scores of >±3; $n = 6$). This left 193 participants for analysis (140 female; mean age = 19.88 years, $SD = 3.62$, range = 18-40). Assignment to encoding and test groups was randomised. Ns per group are shown in Table 4.

Stimuli and procedure

The Experiment 2 stimuli and procedure were used. The RK + C group was identical to the RK + C group in Experiment 2. In the RK group, participants made only RK judgments after their "old" judgments. The only other difference was that here the distractor task was computerised.

Results and discussion

Below, in turn we examined whether recognition differed across the groups, and whether asking participants to assess confidence alongside their recognition experience altered their RK judgments.

Recognition

As per Experiment 2, we first examined whether making confidence judgments influenced recognition in terms of hits, *d′*, false alarms, and *c*; means shown in Table 4. The ANOVAs yielded robust main effects of encoding on hits and *d′* (deep > shallow): $F(1, 189) = 169.02$, $MSE = 0.014$, $p < .001$, $\eta^2 = .47$, and $F(1, 189) = 300.90$, $MSE = 0.197$, $p < .001$, $\eta^2 = .61$ respectively; false alarms (shallow > deep): $F(1, 189) = 38.62$, $MSE = 0.011$, $p < .001$, $\eta^2 = .166$; and *c* (more conservative for deep than shallow): $F(1, 189) = 11.06$, $MSE = 0.128$, $p = .001$, $\eta^2 = .054$. Test group did not have a significant effect on any of these measures: hits, $F(1, 189) = 0.046$, $p = .83$; *d′*, $F(1, 189) = 1.77$, $p = .19$; false alarms, $F(1, 189) = 1.84$, $p = .18$; or *c*, $F(1, 189) = 0.83$, $p = .36$. Nor did test group significantly interact with encoding on hits, $F(1, 189) = 1.19$, $p = .28$; *d′*, $F(1, 189) = 1.75$, $p = .19$; false alarms, $F(1, 189) = 2.77$, $p = .10$; or *c*, $F(1, 189) = 3.29$, $p = .071$. The corresponding Bayesian ANOVA model comparisons and BFs produced *anecdotal* to *moderate* support for the conclusion that only encoding condition influenced these recognition measures (see Table 6). In sum, recognition was similar whether participants made both confidence and RK judgments or just RK judgments after each "old" decision.

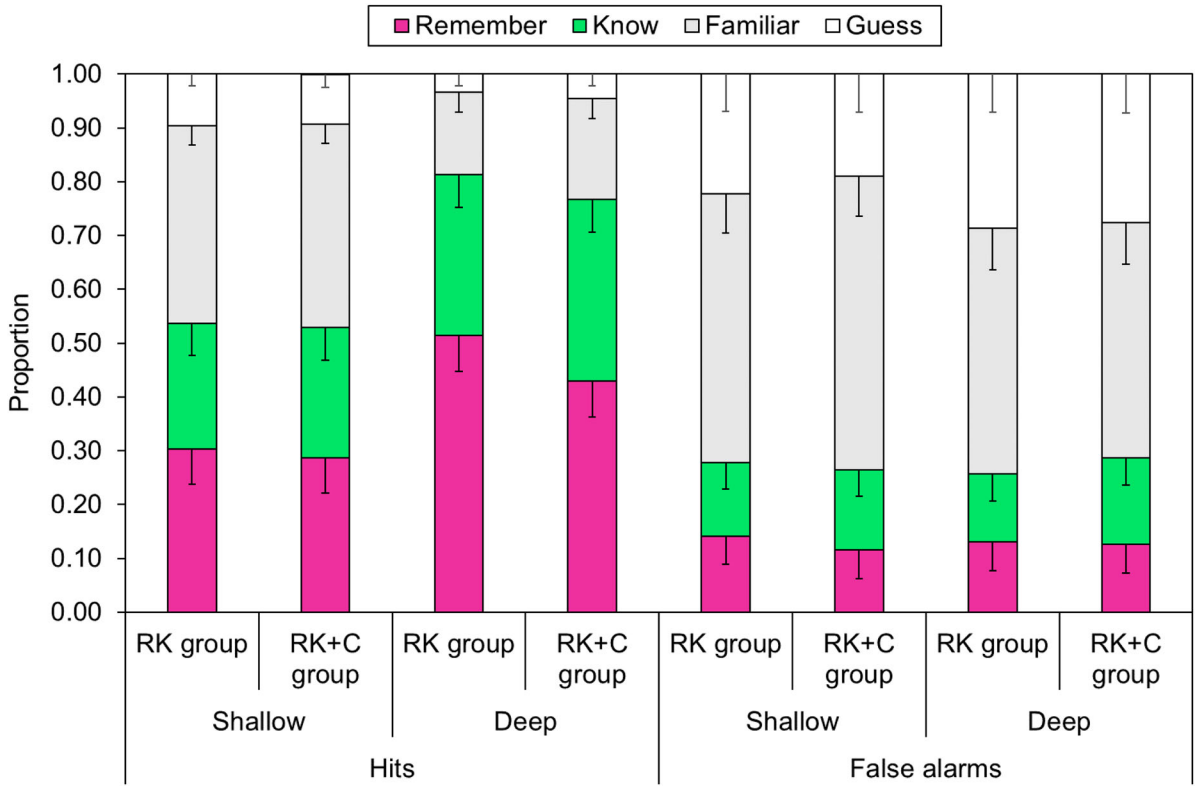

Figure 2. Mean proportion of hits and false alarms assigned to subjective experience response options (remember, know, familiar, guess) by encoding condition and test group in Experiment 3. Error bars show only lower 95% confidence intervals to avoid overlap.

Remember-Know judgments

We next examined whether making confidence judgments influenced RK judgments.[4] First we compared hits for each RK response using a 2 (encoding condition: shallow vs. deep) x 2 (test group: RK vs. RK+C) between-subjects ANOVA (see Figure 2). For R judgments, there was a significant effect of encoding condition, $F(1, 189) = 27.64$, $MSE = 0.054$, $p < .001$, $\eta^2 = .126$, but no significant effect of test group, $F(1, 189) = 2.24$, $p = .14$, or interaction, $F(1, 189) = 1.08$, $p = .30$. The same pattern occurred for K judgments, respectively, encoding: $F(1, 189) = 6.95$, $MSE = 0.045$, $p = .009$, $\eta^2 = .035$; test group: $F(1, 189) = 0.61$, $p = .44$, and interaction: $F(1, 189) = 0.25$, $p = .62$. For F judgments there was a significant reverse-LOP effect, with fewer F judgments after deep than shallow encoding, $F(1, 189) = 122.55$, $MSE = 0.016$, $p < .001$, $\eta^2 = .39$, but again the main effect of test group and interaction were not significant, $F(1, 189) = 1.41$, $p = .24$, and $F(1, 189) = 0.46$, $p = .50$, respectively. The same pattern occurred for G judgments, respectively, encoding condition showed significant reverse-LOP effect: $F(1, 189) = 21.26$, $MSE = 0.007$, $p < .001$, $\eta^2 = .101$; no significant effect of test group: $F(1, 189) = 0.12$, $p = .73$, and no significant interaction: $F(1, 189) = 0.56$, $p = .46$. The corresponding Bayesian analyses produced *anecdotal* to *moderate* support for the conclusion that only encoding influenced recognition judgments for hits (see Table 6).[5]

Equivalent ANOVAs were conducted for false alarms and although a few of the effects of LOP were significant, none of the effects of test group or their interaction were significant, respectively; for R judgments, encoding $F(1, 187) < 0.007$, $MSE = 0.034$, $p = .98$, test group, $F(1, 187) = 0.28$, $p = .60$, and interaction, $F(1, 187) = 0.16$, $p = .69$; for K judgments, encoding, $F(1, 187) < 0.007$, $MSE = 0.030$, $p = .98$, test group, $F(1, 187) = 0.81$, $p = .37$, and interaction, $F(1, 187) = 0.20$, $p = .65$; for F judgments, encoding, $F(1, 187) = 4.05$, $MSE = 0.070$, $p = .046$, $\eta^2 = .021$, test group, $F(1, 187) = 0.13$, $MSE = 0.070$, $p = .72$, and interaction, $F(1, 187) = 0.76$, $MSE = 0.070$, $p = .39$; for G judgments, encoding, $F(1, 187) = 4.53$, $MSE = 0.062$, $p = .035$, $\eta^2 = .024$, test group, $F(1, 187) = 0.37$, $p = .54$, and interaction, $F(1, 187) = 0.098$, $p = .76$. For R and K false alarms, the corresponding Bayesian ANOVAs provided most support for the null model; for F and G false alarms there was evidence that LOP influenced participants' use of these response options, but this support was *anecdotal* and effect sizes were small (see Table 6). In sum, effects of test group were absent for both recognition and subjective recognition judgments, and making C judgments did not affect participants' recognition judgments.

General discussion

Our first research question was whether asking participants to assess their subjective experience, either

through RK or confidence judgments, affects old/new recognition for unrelated words. In Experiment 1, the addition of RK or C judgments did not alter recognition relative to a no-judgment condition in terms of hits, false alarms, or signal-detection indices. The similarity of recognition with or without *RK judgments* replicates Hicks and Marsh (1999) – here we showed that this similarity holds whether items were encoded in a deep or shallow LOP task. The similarity of recognition with or without *C judgments*, on the other hand, contradicts Geraci et al. (2009), and suggests that their finding of an influence of C judgments on recognition may have been an artefact of their use of an RK-then-C testing order, which confounded recognition task practice with judgment order. Reassuringly for memory researchers, asking participants to make RK or C judgments alongside their old/new judgments appeared not to affect how they experienced or output their recognition responses.

Our findings raise the question of why making subjective judgments results in metacognitive reactivity in some situations but not in others. In contrast to our null findings, Mulligan et al. (2010) found including RK judgments improved recognition, but only in their modality-match condition. They reasoned that this influence arose because retrieval of perceptual information was particularly pertinent in that condition. Similarly, Naveh-Benjamin and Kilb (2012) found that older adults' associative memory was improved by the presence of a remember-know judgment, whereas younger adults' associative memory and item recognition were not affected. They suggested that requiring RK judgments provides older adults with a trigger to adopt associative strategies during encoding and retrieval that they do not otherwise employ. Moreover, Double and Birney (2017, 2018, 2019) found that requiring a subjective judgment influenced performance in another cognitive domain – problem solving. In a problem-solving task such as a Latin Square or Raven's Matrix, how processing is metacognitively monitored is self-initiated and intentional. In a recognition task, in contrast, one's processing is more stimulus-driven and unfolds at least in part in an automatic manner (e.g., Eich, 1980). Perhaps reactivity to self-report measures is more likely when a task prompts additional processing in the primary task that otherwise would not be performed. In our recognition task, the requirement to make RK or C judgments may not have prompted extra processing, suggesting that participants evaluated their recognition decisions similarly regardless of whether they were also asked to make RK or C judgments.

Turning to our second research question, Experiments 2 and 3 tested the common (but hitherto untested) assumption that confidence and RK judgments influence each other (Holmes et al., 1998; Humphreys et al., 2003; Migo et al., 2012; Yonelinas, 2001). Reassuringly, we found no evidence that making RK judgments influences C judgments (Experiment 2), or that making C judgments influences RK judgments (Experiment 3). These null results were unexpected. We had thought that being asked to assess retrieval of contextual associated info (remembering) might reduce participant's use of the highest confidence response. That expectation follows from the finding that participants are typically more lenient in making high confidence ratings than in making remember judgments (e.g., Dunn, 2004; McCabe et al., 2011). Our experiments replicated this pattern, 59-68% of deep LOP items were assigned to the highest level of confidence across our experiments, whereas only 43-51% were assigned to remember. But, crucially, patterns of RK or C judgments were not affected by the requirement to make the other type of judgment.

Geraci et al. (2009) suggested that C judgment instructions may be more easily understood by participants than RK instructions. Moreover, the wording of RK instructions, and the response options participants are offered in RK tasks, can impact how a given option is used (Geraci et al., 2009; Rotello et al., 2005; Williams & Lindsay, 2019; see also Umanath & Coane, 2020, for evidence and discussion of differences between how psychologists and lay people differentiate remembering from knowing). Based on such findings, Haaf et al. (2021) argued that C judgments should be preferred over RK judgments. However, as noted by Migo et al. (2012), use of C judgments can also be problematic. For example, participants may find it difficult to make fine discriminations between highly confident memories (Mickes et al., 2011). In addition, when given detailed confidence scales, participants tend to reduce the scale and respond in only fixed increments (Mickes et al., 2007). Kantner and Dobbins (2019) showed that confidence in recognition can be influenced by individual differences even more than by the accuracy of the recognition decision. And McCabe et al. (2011), in an experiment employing think-aloud protocols at both study and test, found that C judgments showed weaker correspondence with levels of contextual retrieval than did RK judgments. From these findings, McCabe et al. noted that it is hard to know what type of information participants use as the basis for making their C judgments.

Regardless of which type of post-recognition judgment is deemed "best," our findings indicate that either RK or C judgments can be used without contaminating basic recognition measures. We were unable to find compelling evidence that the mere presence of RK or C judgments influences recognition. Nor did we find that these two different post-recognition judgments cross-contaminate each other. Nonetheless, some caution is warranted regarding the generalizability of our findings. No research has so far examined whether there is reactivity in tests using other materials (e.g., categorised words, pictures), or for other types of recognition test (e.g., associative recognition, multiple-choice tests). Additionally, our participants made C judgments using a 1–7 scale, but confidence has been measured many other ways across studies. For example, perhaps calibration accuracy for

confidence might be affected by the presence of RK judgments when confidence is measured on a 0-100% scale. Further research examining reactivity and the potential for contamination with other materials, other confidence scales, and other types of recognition test would therefore be beneficial.

Typical confidence instructions provide participants with limited or no guidance regarding the type of information they should base their C ratings on. Therefore, C judgments may provide less precise information regarding the subjective experiences associated with retrieval than do RK judgments (McCabe et al., 2011). As long as RK definitions and instructions are published to enable fully-informed comparisons across experiments (Migo et al., 2012; Williams & Lindsay, 2019), we suggest that RK judgments add value beyond that provided by C judgments. But our findings suggest that there is no harm in collecting both types of measures on each trial of an old/new word recognition experiment.

Notes

1. In Experiment 1 participants used the response options R/K/G while in Experiments 2 and 3 participants used the response options R/K/F/G (cf. Table 1 and Table 3). This was due to changes in lab protocols across the period in which these experiments were conducted. The separation of non-recollection into Know and Familiar is discussed in Williams and Lindsay (2019).
2. Order of judgments was only applicable to the RK+C group so it could not be included as a factor in the main ANOVAs. Our supplementary Analyses confirm that there were no effects of the order of post-recognition judgments in Experiments 2 and 3 (see https://osf.io/hf38m/).
3. Our Supplementary Analyses (see https://osf.io/hf38m/) provide the proportion of hits and false alarms by recognition experience (RK+C group) and confidence level (C group). As in Experiment 1, participants used the responses appropriately.
4. Our Supplementary Analyses (see https://osf.io/hf38m/) provide mean confidence and the proportion of hits and false alarms by confidence level (1-7).
5. In typical R/F/G or R/K/G tasks with three response options, the proportion of K (or F) responses tends to be lower after deep compared to shallow encoding, because of the increase in R responses (cf. Supplementary Analyses Figure 1 for Experiment 1 means: https://osf.io/hf38m/). The rather unusual result of more know responses after deep encoding is due to the separation of know and familiar responses. We are exploring patterns of K and F responses in 4-option R/K/F/G tasks in other work (Williams et al., 2022). Critically, here there was no significant difference in patterns of subjective experience across test groups, suggesting that making confidence judgments did not affect how participants used the four response options.

Acknowledgements

Data and materials for all experiments are available online: https://osf.io/hf38m/. None of the experiments were pre-registered.

Experiment 1 derived from discussions with GEB at the University of Calgary funded by Economic and Social Research Council (ESRC) Future Research Leaders award ES/N001753/1 to HLW. Experiments 2 and 3 were supported by a Commonwealth Postdoctoral Fellowship to HLW funded by the Government of Canada and a Natural Sciences and Engineering Research Council (NSERC) Discovery Grant to DSL. We thank David Drohan, Scott Richardson, Sierra Hall, and Michael Davies for help with data collection.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by Economic and Social Research Council: [grant no ES/N001753/1]; Foreign Affairs and International Trade Canada; Natural Sciences and Engineering Research Council of Canada.

ORCID

Glen E. Bodner http://orcid.org/0000-0001-7837-4879
D. Stephen Lindsay http://orcid.org/0000-0002-6439-987X

References

Bodner, G. E., & Lindsay, D. S. (2003). Remembering and knowing in context. *Journal of Memory and Language*, *48*(3), 563–580. https://doi.org/10.1016/S0749-596X(02)00502-8

Bruno, D., & Rutherford, A. (2010). How many response options? A study of remember–know testing procedures. *Acta Psychologica*, *134*(2), 125–129. https://doi.org/10.1016/j.actpsy.2010.01.002

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

Double, K. S., & Birney, D. P. (2018). Reactivity to confidence ratings in older individuals performing the Latin Square task. *Metacognition and Learning*, *13*(3), 309–326. https://doi.org/10.1007/s11409-018-9186-5

Double, K. S., & Birney, D. P. (2019). Do confidence ratings prime confidence? *Psychonomic Bulletin & Review*, *26*(3), 1035–1042. https://doi.org/10.3758/s13423-018-1553-3

Dunn, J. C. (2004). Remember-know: A matter of confidence. *Psychological Review*, *111*(2), 524–542. https://doi.org/10.1037/0033-295X.111.2.524

Eich, J. E. (1980). The cue-dependent nature of state-dependent retrieval. *Memory & Cognition*, *8*(2), 157–173. https://doi.org/10.3758/BF03213419

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. https://doi.org/10.1037/a0021663

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, *16*(4), 309–313. https://doi.org/10.3758/BF03197041

Gardiner, J. M. (2001). Episodic memory and autonoetic consciousness: A first-person approach. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *356*(1413), 1351–1361. https://doi.org/10.1098/rstb.2001.0955

Gardiner, J. M., & Java, R. I. (1990). Recollective experience in word and nonword recognition. *Memory & Cognition*, *18*(1), 23–30. https://doi.org/10.3758/BF03202642

Gardiner, J. M., Java, R. I., & Richardson-Klavehn, A. (1996). How level of processing really influences awareness in recognition memory. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *50*(1), 114–122. https://doi.org/10.1037/1196-1961.50.1.114

Geraci, L., McCabe, D. P., & Guillory, J. J. (2009). On interpreting the relationship between remember-know judgments and confidence: The role of instructions. *Consciousness and Cognition*, *18*(3), 701–709. https://doi.org/10.1016/j.concog.2009.04.010

Haaf, J. M., Rhodes, S., Naveh-Benjamin, M., Sun, T., Snyder, H. K., & Rouder, J. N. (2021). Revisiting the remember–know task: Replications of Gardiner and Java (1990). *Memory & Cognition*, *49*(1), 46–66. https://doi.org/10.3758/s13421-020-01073-x

Hicks, J. L., & Marsh, R. L. (1999). Remember-know judgments can depend on how memory is tested. *Psychonomic Bulletin & Review*, *6*(1), 117–122. https://doi.org/10.3758/BF03210818

Holmes, J. B., Waters, H. S., & Rajaram, S. (1998). The phenomenology of false memories: Episodic content and confidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(4), 1026–1040. https://doi.org/10.1037/0278-7393.24.4.1026

Humphreys, M. S., Dennis, S., Maguire, A. M., Reynolds, K., Bolland, S. W., & Hughes, J. D. (2003). What you get out of memory depends on the question you ask. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(5), 797–812. https://doi.org/10.1037/0278-7393.29.5.797

Ingram, K. M., Mickes, L., & Wixted, J. T. (2012). Recollection can be weak and familiarity can be strong. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(2), 325–339. https://doi.org/10.1037/a0025483

JASP Team. (2019). JASP (Version 0.9) [Computer software].

Kantner, J., & Dobbins, I. G. (2019). Partitioning the sources of recognition confidence: The role of individual differences. *Psychonomic Bulletin & Review*, 1317–1324. https://doi.org/10.3758/s13423-019-01586-w

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.), *Handbook of memory and metamemory* (pp. 117–136). New York: Psychology Press.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Mäntylä, T. (1993). Knowing but not remembering: Adult age differences in recollective experience. *Memory & Cognition*, *21*(3), 379–388. https://doi.org/10.3758/BF03208271

McCabe, D. P., Geraci, L., Boman, J. K., Sensenig, A. E., & Rhodes, M. G. (2011). On the validity of remember–know judgments: Evidence from think aloud protocols. *Consciousness and Cognition*, *20*(4), 1625–1633. https://doi.org/10.1016/j.concog.2011.08.012

McCabe, D. P., Roediger, H. L., McDaniel, M. A., & Balota, D. A. (2009). Aging reduces veridical remembering but increases false remembering: Neuropsychological test correlates of remember–know judgments. *Neuropsychologia*, *47*(11), 2164–2173. https://doi.org/10.1016/j.neuropsychologia.2008.11.025

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*(2), 239–257. https://doi.org/10.1037/a0023007

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*(5), 858–865. https://doi.org/10.3758/BF03194112

Migo, E. M., Mayes, A. R., & Montaldi, D. (2012). Measuring recollection and familiarity: Improving the remember/know procedure. *Consciousness and Cognition*, *21*(3), 1435–1455. https://doi.org/10.1016/j.concog.2012.04.014

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

Mulligan, N. W., Besken, M., & Peterson, D. (2010). Remember-Know and source memory instructions can qualitatively change old-new recognition accuracy: The modality-match effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(2), 558–566. https://doi.org/10.1037/a0018408

Naveh-Benjamin, M., & Kilb, A. (2012). How the measurement of memory processes can affect memory performance: The case of remember/know judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 194–203. https://doi.org/10.1037/a0025256

Parks, C. M., Murray, L. J., Elfman, K., & Yonelinas, A. P. (2011). Variations in recollection: The effects of complexity on source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 861–873. https://doi.org/10.1037/a0022798

Parks, C. M., & Yonelinas, A. P. (2007). Moving beyond pure signal-detection models: Comment on Wixted (2007). *Psychological Review*, *114*(1), 188–201. https://doi.org/10.1037/0033-295X.114.1.188

Perfect, T. J., Williams, R. B., & Anderton-Brown, C. (1995). Age differences in reported recollective experience are due to encoding effects, not response bias. *Memory*, *3*(2), 169–186. https://doi.org/10.1080/09658219508258964

Rajaram, S. (1993). Remembering and knowing: Two means of access to the personal past. *Memory & Cognition*, *21*(1), 89–102. https://doi.org/10.3758/BF03211168

Rajaram, S., Hamilton, M., & Bolton, A. (2002). Distinguishing states of awareness from confidence during retrieval: Evidence from amnesia. *Cognitive Affective & Behavioral Neuroscience*, *2*(3), 227–235. https://doi.org/10.3758/CABN.2.3.227

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Rotello, C. M., Macmillan, N. A., Reeder, J. A., & Wong, M. (2005). Theremember response: Subject to bias, graded, and not a process-pure indicator of recollection. *Psychonomic Bulletin & Review*, *12*(5), 865–873. https://doi.org/10.3758/BF03196778

Smith, C. N., Wixted, J. T., & Squire, L. R. (2011). The hippocampus supports both recollection and familiarity when memories are strong. *The Journal of Neuroscience*, *31*(44), 15693–15702. https://doi.org/10.1523/JNEUROSCI.3438-11.2011

Smith, R. E., Hunt, R. R., & Gallagher, M. P. (2008). The effect of study modality on false recognition. *Memory & Cognition*, *36*(8), 1439–1449. https://doi.org/10.3758/MC.36.8.1439

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Sommer, T., Schröter, R., & Bayer, J. (2021). Probing emotional recognition memory: how different response formats affect response behaviour. *Memory*, *29*(9), 1216–1231. https://doi.org/10.1080/09658211.2021.1974049

Tousignant, C., & Bodner, G. E. (2012). Test context affects recollection and familiarity ratings: Implications for measuring recognition experiences. *Consciousness and Cognition*, *21*(2), 994–1000. https://doi.org/10.1016/j.concog.2012.01.009

Tousignant, C., Bodner, G. E., & Arnold, M. M. (2015). Effects of context on recollection and familiarity experiences are task dependent. *Consciousness and Cognition*, *33*, 78–89. https://doi.org/10.1016/j.concog.2014.11.011

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, *26*(1), 1–12. https://doi.org/10.1037/h0080017

Umanath, S., & Coane, J. H. (2020). Face validity of remembering and knowing: Empirical consensus and disagreement between participants and researchers. *Perspectives on Psychological Science*, *15*(6), 1400–1422. https://doi.org/10.1177/1745691620917672

Wagenmakers, E. J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Selker, R., Gronau, Q. F., Dropmann, D., Boutin, B., & Meerhoff, F. (2018). Bayesian inference for psychology. Part II: Example applications with JASP. *Psychonomic Bulletin & Review*, *25*(1), 58–76. https://doi.org/10.3758/s13423-017-1323-7

Williams, H. L., Conway, M. A., & Moulin, C. J. A. (2022). The relationship between source, confidence, and subjective experience when two judgments are made post-recognition. *Manuscript in preparation*.

Williams, H. L., & Lindsay, D. S. (2016). Recognition sensitivity, confidence, and bias in continuous versus study-test recognition procedures. *Proceedings of the Psychonomic Society*, *21*.

Williams, H. L., & Lindsay, D. S. (2019). Different definitions of the non-recollection-based response option(s) change how people use the "Remember" response in the Remember/Know paradigm. *Memory & Cognition*, *47*(7), 1359–1374. https://doi.org/10.3758/s13421-019-00938-0

Williams, H. L., & Lindsay, D. S. (2022). Continuous recognition versus study-test recognition. *Manuscript in preparation*.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152–176. https://doi.org/10.1037/0033-295X.114.1.152

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of Remember/Know judgments. *Psychological Review*, *117*(4), 1025–1054. https://doi.org/10.1037/a0020874

Yonelinas, A. P. (2001). Consciousness, control, and confidence: The 3 Cs of recognition memory. *Journal of Experimental Psychology: General*, *130*(3), 361–379. https://doi.org/10.1037/0096-3445.130.3.361

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*(3), 441–517. doi:10.1006/jmla.2002.2864