Independent Recollection-Familiarity Ratings: Similar Effects of Levels-of-Processing Whether Amount or Confidence is Rated

Helen L. Williams1 and Glen E. Bodner2

1Keele University, 2Flinders University

BRIEF REPORT

Word count*:* 2,703 (main text)

Author Note

This research was supported by Economic and Social Research Council (ESRC) Future Research Leaders award ES/N001753/1 to HLW. We thank Bill Hockley and John Vokey for helpful reviews, Jamie Adams for programming and Mariya Singkh for data collection. Stimuli, instructions, and data are available online here: osf.io/8gyns. Correspondence concerning this article should be addressed to Helen L. Williams, School of Psychology, Dorothy Hodgkin Building, Keele University, Keele, Staffordshire, ST5 5BG. Email: h.l.williams@keele.ac.uk.

Abstract

Independent recollection-familiarity (RF) ratings are sometimes collected to measure subjective experiences of recollection and familiarity during recognition. Although the RF ratings task purports to measure the ‘degree’ to which each recognition state is experienced, the rating scale has been worded in terms of confidence rather than amount. Given prior evidence that wording influences recognition and remember/know judgments, we compared RF rating scales worded in terms of ‘amount’ versus ‘confidence’ across two groups. A robust levels-of-processing effect occurred on both recollection and familiarity ratings, and its magnitude was similar across scale wording. Scale wording did not influence recognition ratings, and, most importantly, it had little influence on ratings of recollection and familiarity. These findings suggest that participants may use confidence to rate amount, or vice versa. Regardless, researchers should align their task instructions and scale wording, and should publish them. Such alignment and transparency is crucial for interpreting measures of the memory states that arise during recognition memory.

Keywords: recollection; familiarity; recollection-familiarity ratings; recognition memory

*Public significance statement:* Recollection and familiarity are two ways we experience our recognition of stimuli in the world. One approach to measuring these two processes is to ask people to independently rate the extent of their recollection and familiarity experiences for each item on a memory test. We identified a disconnect between this method and the rating scale used in previous studies, so here we compared two different rating scales across groups. We found little difference how people used the measurement scales. We concluded that although these particular differences in instructions did not affect reports of recollection and familiarity it is still important that researchers publish their exact experimental instructions and definitions to facilitate cross-study comparisons of findings.

Independent Recollection-Familiarity Ratings: Similar Effects of Levels-of-Processing Whether Amount or Confidence is Rated

The remember/know (RK) task is commonly used to measure the subjective memory states associated with recognition memory processes. In this task, developed by Tulving (1985), participants are asked to classify a test item as ‘remember’ when memory is accompanied by specific recollection of associated/contextual information from study, or as ‘know’ when the item feels familiar but is not accompanied by recollection (for reviews, see Gardiner & Richardson-Klavehn, 2000; Migo, Mayes, & Montaldi, 2012). Despite its common use, the RK task has been controversial. For example, the binary nature of RK judgments might induce artefactual dissociative effects of manipulations on the two recognition states (Brown & Bodner, 2011).

A more basic issue is that the RK task disallows the possibility that a recognized item feels both recollected and familiar. In light of such issues, Higham and Vokey (2004) created the independent recollection-familiarity (RF) ratings task. In the RF ratings task, participants independently rate the degree to which they experience recollection and/or familiarity for each test item. In contrast to the dissociative patterns often found using the RK task, variables have typically produced parallel effects in the RF ratings task (Brown & Bodner, 2011; Higham & Vokey, 2004; Kurilla & Westerman, 2008; Tousignant & Bodner, 2012; but see Willems, Schroyen, & Bodner, in press).

Measures of recognition states have been used to adjudicate between dual-process and single-process models of recognition memory (see Dunn, 2008; Wixted & Mickes, 2010; Yonelinas, 2002). But it is premature to relate recognition states to underlying recognition processes if the validity of the tasks is in question. Consistent with that possibility, reports of recognition states can be influenced by the recognition task instructions (e.g., Hirshman & Henzler, 1995), the response options and wording (e.g., Tousignant, Bodner, & Arnold, 2015; Williams & Lindsay, 2018), and the test-list context (e.g., Bodner & Lindsay, 2003; Tousignant & Bodner, 2012). As examples, using Type A/B labels instead of R/K labels increases reports of remembering (McCabe & Geraci, 2009); allowing a guess response can alter reports of knowing (Gardiner, Richardson-Klavehn, & Ramponi, 1997; Bruno & Rutherford, 2010); the level of confidence emphasized in definitions of remembering (Rotello, Macmillan, Reeder, & Wong, 2005) or knowing (Geraci, McCabe, & Guillory, 2009; Williams & Lindsay, 2018) changes how response options are used; and using a two-step test (old/new, then remember/know) produces fewer remember responses than a one-step test (remember/know/new; Bruno & Rutherford, 2010; Hicks & Marsh, 1999).

Here we examined a potential disconnect between the intention of the RF ratings task and the wording of the ratings scale used in prior studies. The RF ratings task is intended to independently measure recollection and familiarity in terms of the “degree to which” or “how much” (Higham & Vokey, 2004), “extent to which” (Kurilla & Westerman, 2008), or the “level to which” (Brown & Bodner, 2011) each state is experienced for a given test item. Interpreted this way, the scales should ask participants to *quantify* the amount of each state they experience. However, the rating scale options in prior RF ratings studies were ‘definitely yes’, ‘probably yes’, ‘probably no’, and ‘definitely no’ – which instead ask participants to rate their *confidence* that they are or are not experiencing each state. The RF ratings task thus confounds memory contents (“how much”) with confidence (“how sure”).

# Consider a situation where, at test, a participant recollects a seemingly minor or irrelevant aspect the study experience with an item (e.g., that the monitor flickered during its presentation). If asked to rate the total “quantity” of recollection for this item, they might provide a low rating. But if asked to rate their “confidence”, they might feel that this recollection is distinctive and highly diagnostic of prior study, resulting in a high rating. In a similar vein, Gruppuso, Lindsay, and Kelley (1997) found that what participants deemed to be an experience of “remembering” depended on whether it allowed them to complete the memory task at hand. Speculatively, scale wording could even influence whether participants adopt an “all or none” threshold process or a “graded” signal-detection process. Alternatively, participants might simply base their confidence ratings on the amount of each state that they experience, or vice versa. If so, then ratings should not be affected by the choice of scale wording. To examine this issue, we measured whether recognition and memory states reported in the RF ratings task are influenced by rating scale wording. We manipulated levels-of-processing (LOP) at study, prior to a recognition test in which the RF ratings task was used. Based on prior findings (e.g., Brown & Bodner, 2011), we expected that a deeper LOP task (focused on the meaning of each word) would result in higher ratings of both recollection and familiarity than a shallower LOP task (focused on the structure of each word). Across groups, we manipulated whether the RF ratings scale was worded in terms of ‘amount’ or ‘confidence’. The amount scale aligns with the purported intention of the RF ratings task in prior studies, whereas the confidence scale aligns with the RF rating scale used in prior studies. The ‘amount’ and ‘confidence’ conceptualisations have also been used somewhat interchangeably, in recent studies of continuous models of recollection and familiarity (e.g., Ingram, Mickes, & Wixted, 2012; Wixted & Mickes, 2010). Of primary interest was whether scale wording would influence the size of the LOP effect on (1) recognition and (2) recollection and/or familiarity ratings. Such findings would indicate that RF ratings need to be interpreted in line with the dimension of recollection and familiarity experiences captured by the task instructions and rating scale. Method

## Participants

The Keele University Ethics Review Panel approved the study. Psychology undergraduates from Keele University participated for credit and were randomly assigned to the amount group (n = 32; female = 22; *mean age* *=* 19.0, range = 18-22, *SD* = 0.84) or confidence group (n = 31, excluding 1 participant with > 50% false alarms; female = 28; *mean age =* 19.1, range = 18-25, *SD* = 1.58). The calculated a priori power to detect a small effect (Cohen’s d = .30) with this size sample was .82 (G\*Power 3.1.5; Faul, Erdfelder, Lang, & Buchner, 2007).

## Stimuli

Stimuli were 128 nouns obtained from the SUBTLEX-UK frequency database, 4 to 7 letters in length, with a mean log Zipf frequency of 4.49. Four lists of 30 words were created. Each participant studied one list under shallow LOP instructions and one list under deep LOP instructions. The other two lists served as the new items (i.e., lures) on the recognition test. List assignment was counterbalanced across participants. Presentation order of stimuli was randomized by participant. Two primacy and two recency fillers were shown at the start and end of each study list; these were randomly mixed with four new fillers to form a practice test (not analyzed). The order of the 120 words on the main recognition test was randomized by participant.

## **Procedure**

Participants were tested individually using PsychoPy (Peirce, 2007). Stimuli remained onscreen until responses were collected via mouse click using onscreen response boxes. A 1-second fixation cross preceded each trial.

At study, participants were told they would study two lists of words for a recognition test. An instruction/example screen was then presented for either the shallow LOP task (“does this word contain the letter a?” yes/no) or deep LOP task (“do you think this word is pleasant?” yes/no). After completing the first study list, an instruction/example screen was presented for the second study list. The order of the LOP tasks was counterbalanced across participants. After studying both lists, participants completed a distractor task (WAIS-R digit symbol substitution; Wechsler, 1981).

At test, the instructions noted that half the words on the test were studied and half were new. Participants were told that their first task was to decide whether a word was ‘old’ (seen in the study phase) or ‘new’ (not seen in the study phase). They were shown an example old/new response screen. Participants were then instructed that if they responded ‘old’ they would then rate their experiences of recognizing that word. They were then shown the definitions of recollection, familiarity, and guessing as presented in Table 1. The next instruction screen, presented in Table 2, differentiated the two scale types. Both groups were then provided the following instruction:

“A recognition experience might be rated high on one scale and low on the other, high on both scales, or low on both scales. You should make the two ratings independently. To indicate that your OLD response was just a guess, select “0” for both ratings”.

Participants were provided the memory state definitions on paper as a reference, but they were also instructed to learn them to expedite their task. During the recognition test, when a participant indicated that a word was ‘old’, the appropriate rating scales for recollection and familiarity ratings were shown, in this order, below the word. Participants made their ratings by clicking a number (0 to 5) shown in response boxes beside each rating. When a participant indicated that a word was ‘new’ the test instead proceeded to the next item.

# Results

Effect sizes are reported for significant comparisons, using partial eta squared (ηp2) for Analyses of Variance (ANOVAs), and Cohen’s *d* for *t*-tests. Mean proportion of hits and false alarms calculated from old/new responses are shown in Table 3. For old items, a mixed 2 (LOP: deep vs. shallow) x 2 (scale: amount, confidence) ANOVA on hits showed a LOP effect, reflecting more hits in the deep (vs. shallow) LOP task, *F*(1, 61) = 266.95, *MSE* = .017, *p* < .001, ηp2 = .81. However, there was no main effect of scale, *F* < 1, and no interaction, *F* < 1. For new items, a one-way ANOVA on false alarms also showed no significant difference across scale, *F* < 1. In short, scale wording did not affect recognition performance.

Figure 1 provides the mean recollection and familiarity ratings by item type and scale. For old items, a 2(rating: recollection, familiarity) x 2(LOP: deep, shallow) x 2(scale: amount, confidence) mixed ANOVA[[1]](#footnote-1) was conducted on mean ratings. There was a significant LOP effect, reflecting higher ratings for deep (vs. shallow) items, *F*(1, 61) = 198.91, *MSE* = .538, *p* < .001, ηp2 = .77, and a significant rating effect, reflecting higher ratings of familiarity than of recollection, *F*(1, 61) = 19.54, *MSE* = .348, *p* < .001, ηp2 = .24. The main effect of scale was not significant, *F* < 1. The interaction between LOP and rating was significant, *F*(1, 61) = 41.85, *MSE* = .179, *p* < .001, ηp2 = .011. Follow-up tests showed that the LOP effect was significant both for recollection ratings, *t*(62)= 15.34, *p* < .001, *dz* = 1.93, and familiarity ratings, *t*(62)= 9.18, *p* < .001, *dz* = 1.16. The remaining interactions were not significant, *F*s < 1. Thus, for studied items, recollection and familiarity ratings were not influenced by scale wording.

For new items, a 2(rating: recollection, familiarity) x 2(scale: amount, confidence) ANOVA was conducted on mean ratings[[2]](#footnote-2). There was a significant main effect of rating, *F*(1, 60) = 69.27, *MSE* = .507, *p* < .001, ηp2 = .536, again reflecting higher ratings of familiarity than of recollection. Unlike for old items there was also a significant main effect of scale, *F*(1, 60) = 4.35, *MSE* = 1.83, *p* = .041, ηp2 = .068. Ratings for new items were generally lower when amount vs. confidence was rated. The interaction between rating and scale was not significant, *F* < 1.

We also assessed the strength of evidence supporting the experimental vs. null hypotheses using Bayes factors (BF; JASP Team, 2018). To simplify the analysis for old items, we collapsed ratings across LOP. Thus for both old and new items Bayesian ANOVAs compared four models against a null model: 1) a model that included only a main effect of rating, 2) a model that included only a main effect of scale, 3) a model that included both main effects, and 4) a model that included both main effects plus an interaction effect. The BF for each model quantifies the relative strength of evidence in support of that model in comparison to the null model. We took the model with the highest overall BF to be the best-fitting model. We then compared how well this model fit the data compared to the next-best model by taking the ratio of their BFs; this produces a new BF that quantifies the degree of superiority of our best-fitting model. For old items, the data were best predicted by a model that included only a main effect of rating (BF = 518.65). The next best model included both rating and scale (BF = 243.19), but the former model was preferred over the latter model by a BF of 2.13. This provides anecdotal evidence (based on classifications specified by Wagenmakers et al., 2018) that ratings for old items were influenced by rating type (recollection vs. familiarity) but not scale type (amount vs. confidence). Conversely, for new items, the best-fitting model included both rating and scale (BF = 1.157 x 109), and the next best-fitting model was the rating-only model (BF = 6.879 x 108). Here the former model was preferred over the latter model by a BF of 1.98, providing anecdotal evidence that scale wording influenced ratings for new items.

# Discussion

Recognition decisions and reports of recollection and familiarity in the RK task are often influenced by response options, definitions, and instructions (Bruno & Rutherford, 2010; Gardiner et al., 1997; Geraci et al., 2009; Hicks & Marsh, 1999; McCabe & Geraci, 2009; Rotello et al., 2005; Williams & Lindsay, 2018). To overcome a concern with the binary nature of the RK task, Higham and Vokey (2004) introduced the independent RF ratings task, which has since been used in several studies (e.g., Brown & Bodner, 2011; Kurilla & Westerman, 2008; Tousignant & Bodner, 2012). We identified a potential mismatch between the intention of using this task to measure “how much” of each state is experienced, and the use of rating scales that measure “how confident” a participant is that they experienced each state. After a LOP manipulation at study, recollection and familiarity ratings for old items on a recognition test were similar whether framed in terms of ‘amount’ or ‘confidence’. For new items, false alarms with the ‘amount’ wording yielded lower ratings, but here too there was no interaction with rating type. Thus, unlike previous studies using the traditional RK judgment task, variations of the RF ratings task did not impact the main findings. Instead, each group replicated the parallel effects of LOP on recollection and familiarity ratings reported by Brown and Bodner (2011).

Our experiment provides some assurance that both scale wordings tap into the underlying distinction between recollection and familiarity experiences first identified by Mandler (1980). The similarity in findings across scales could indicate that participants’ confidence ratings are based on the amount of each state they experience, or vice versa – or that the underlying dimension of the memory states measured by the RF ratings task is something else. This remains an important research question. Although the similarity across scales is consistent with a single-process state-trace account of recognition memory states (e.g., Dunn, 2008), other findings challenge that account (Ingram et al., 2012; Wixted & Mickes, 2010).

Even though we found little influence of scale wording, it is sensible for researchers to choose a scale wording that aligns with what they intend to use the RK ratings task to measure. Thus, researchers interested in ‘amount’ of recollection and familiarity states should use the ‘amount’ wording and those interested in ‘confidence’ in recollection and familiarity states should use the ‘confidence’ wording. This suggestion aligns with Gardiner’s (2008) advice regarding inclusion of a guess response option in the RK judgment task. Gardiner asserted that a guess option should be used if the focus is on recognition memory states, but can be left out if the focus is on overall recognition accuracy. Finally, echoing Migo et al. (2012), we recommend that researchers publish their verbatim task instructions, as we have done here. Following this recommendation should increase transparency and replicability, which in turn will foster our understanding of memory states and their relation to underlying memory processes.

References

Brown, A. A., & Bodner, G. E. (2011). Re-examining dissociations between remembering and knowing: Binary judgments vs. independent ratings. *Journal of Memory and Language, 65,* 98–108. http://dx.doi.org/10.1016/j.jml.2011.04.003.

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

Dunn, J. C. (2008). The dimensionality of the remember-know task: a state-trace analysis. *Psychological Review, 115*(2), 426-446. doi: 10.1037/0033-295X.115.2.426

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. doi: 10.3758/BF03193146

Gardiner, J. M. (2008). Remembering and knowing. In J. H. Byrne (Ed.), *Learning and memory: A comprehensive reference* (pp. 285-305). Oxford: Elsevier Ltd. doi: 10.1016/B978-012370509-9.00167-4

Gardiner, J. M., & Richardson-Klavehn, A. (2000). Remembering and knowing. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford Handbook of Memory* (pp. 229-244). New York, NY, US: Oxford University Press.

Gardiner, J. M., Richardson-Klavehn, A., & Ramponi, C. (1997). On reporting recollective experiences and ''direct access to memory systems''. *Psychological Science, 8*, 391-394. doi: 10.1111/j.1467-9280.1997.tb00431.x

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*, 701-709. doi: 10.1016/j.concog.2009.04.010

Gruppuso, V., Lindsay, D. S., & Kelley, C. M. (1997). The process-dissociation procedure and similarity: Defining and estimating recollection and familiarity in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(2), 259-278. doi: 10.1037/0278-7393.23.2.259

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

Higham, P. A., & Vokey, J. R. (2004). Illusory recollection and dual-process models of recognition memory. *Quarterly Journal of Experimental Psychology, 57A,* 714–744. http://dx.doi.org/10.1080/02724980343000468.

Hirshman, E., & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science, 9*(1), 61-65. doi: 10.1111/1467-9280.00011

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. doi: 10.1037/a0025483

Kurilla, B. P., & Westerman, D. L. (2008). Processing fluency affects subjective claims of recollection. *Memory and Cognition, 36,* 82–92. <http://dx.doi.org/10.3758/MC.36.1.82>.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, *87*(3), 252-271. doi: 10.1037/0033-295X.87.3.252.

McCabe, D. P., & Geraci, L. D. (2009). The influence of instructions and terminology on the accuracy of remember-know judgments. *Consciousness and Cognition, 18*, 401-413. doi: 10.1016/j.concog.2009.02.010

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

Peirce, J.W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8-13. doi: 10.1016/j.jneumeth.2006.11.017.

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

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. doi: 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. doi: 10.1016/j.concog.2014.11.011

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

Wechsler, D. (1981). *WAIS-R Manual: Wechsler Adult Intelligence Scale-Revised.* New York: Psychological Corporation.

Willems, S., Schroyen, S., & Bodner, G. E. (in press). Independent recollection/familiarity ratings can dissociate: Evidence from the effects of test context on recognition of event details. *Canadian Journal of Experimental Psychology*.

Williams, H. L. & Lindsay, D. S. (2018). *Different definitions of the non-recollection-based response option(s) change how people use the “Remember” response in the Remember/Know paradigm*. Under review.

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

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*, 441-517. doi: 10.1006/jmla.2002.2

Table 1

*Definitions of Recollection, Familiarity, and Guessing Provided to Participants.*

|  |  |
| --- | --- |
| Recognition Experience | Definition |
| Recollection | A recollection experience 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 how the word looked, or of something that happened in the room, or of what you were thinking or doing at the time). In other words, in an experience of recollection you may 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.  |
| Familiarity | A familiarity experience could include having a feeling of familiarity for the word, or the feeling that you just know that the word was one you studied earlier in the learning phase of the experiment.  |
| Guessing | You do not have a recollection or familiarity experience for the word but are simply guessing that the word was old. |

Table 2

*Instructions for Amount and Confidence Scales.*

|  |  |
| --- | --- |
| Scale | Instruction |
| Amount | For each word that you judge as OLD, you will then be asked to rate the strength of your recollection and familiarity experiences for that word, using the scales below. Please rate *the* *amount of recollection experienced* for this item, where: 0 = none, 1 = lowest, 5 = highest. Please rate *the amount of familiarity experienced* for this item, where: 0 = none, 1 = lowest, 5 = highest.  |
| Confidence | For each word that you judge as OLD, you will then be asked to rate the strength of your recollection and familiarity experiences for that word, using the scales below. Please rate *how confident you are that you experienced recollection* for this item, where: 0 = none, 1 = lowest, 5 = highest. Please rate *how confident you are that you experienced familiarity*for this item, where: 0 = none, 1 = lowest, 5 = highest.  |

Note. Italics added for emphasis (not italicized for participants).

Table 3

*Mean (SD) Proportion of “Old” claims by Item Type and Scale.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Scale  | N | Deep Items | Shallow Items | New Items |
| Amount | 32 | .87 (.15) | .49 (.21) | .13 (.08) |
| Confidence | 31 | .87 (.13) | .48 (.19) | .12 (.08) |



Figure 1. *Mean ratings of recollection and familiarity by item type and scale. Error bars show 95% CIs (between-subjects for each pair of ratings)*.

1. To handle the unequal Ns, Type III sums of squares were used in calculation of ANOVA. [↑](#footnote-ref-1)
2. The df is reduced for this analysis as one participant did not make any false alarms to new items. [↑](#footnote-ref-2)