Journal of Memory and Language 107 (2019) 152–168

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

I don't remember vs. I don't know: Phenomenological states associated with retrieval failures☆

Jennifer H. Coane[a,*], Sharda Umanath[b]

[a] Department of Psychology, Colby College, 5550 Mayflower Hill Drive, Waterville, ME 04901, USA
[b] Department of Psychological Science, Claremont McKenna College, 850 Columbia Ave, Claremont, CA 91711, USA

ARTICLE INFO

Keywords:
Accessibility
Availability
Knowledge
Phenomenology
Retrieval failures
Aging

ABSTRACT

When retrieval fails, what is the phenomenology of that experience? We explored different states of experience associated with retrieval failures that vary in intensity. Specifically, we examined the difference between not knowing and not remembering and the ways in which these states are described. Naïve and expert participants defined "I don't know" (DK) and "I don't remember" (DR). DR was associated with lack of accessibility and forgetting, whereas DK was associated with never having learned the information. To examine whether these states map onto distinct behavioral outcomes, in two experiments, younger and older adult participants were asked general knowledge questions with the option of responding DK or DR after a retrieval failure. On a final multiple-choice test (Exp. 2) or cued recall test following correct answer feedback (Exp. 3), when an initial DK response was given, performance was generally lower than following initial DR responses, suggesting that not remembering reflects a failure in accessibility, whereas not knowing reflects the experience of not having information in the knowledge base. The effect was large and robust across ages and tests.

Introduction

The ability to remember and use information at a later point in time is crucial for an organism's survival, and decades of memory research have examined the factors and processes involved in successful encoding and retrieval. Objective memory performance is generally defined in terms of accuracy: To what extent participants can correctly recall or reject items on a test, categorizing them as previously encountered or not. A closely related question, that has received much empirical attention as well, concerns the ability of participants to accurately predict or assess the contents of their memory. The latter, which relies on introspective, subjective judgments, is the focus of the present work.

Memory systems allow individuals to learn and store massive amounts of information, accessing that knowledge when needed or determining that the information needed is not known or is temporarily unable to be remembered. The systems are adaptive in that information that is not used or accessed frequently or recently is less accessible or is no longer available due to trace decay (Schacter, 2001). This distinction, between what is *available* (i.e., stored) in memory and what is *accessible* (i.e., can be retrieved at a specific point in time; Tulving & Pearlstone, 1966) underscores the dynamic and adaptive nature of memory. What is accessible is likely to be information that is common, recent (via encoding or retrieval), salient to the individual, or otherwise important. Furthermore, accessibility is highly dependent on cues or contextual information (e.g., Smith & Vela, 2001; Tulving & Thomson, 1973) and fluctuates across retrieval attempts, with information being more or less accessible due to a variety of factors (e.g., prior retrieval, presence of cues; Kornell, Klein, & Rawson, 2015; Light & Carter-Sobell, 1970; Thomson & Tulving, 1970). However, a far more substantial fount of information is available, even if some of this information is rarely or infrequently accessed. Successful use of memory systems likely depends on the individual's ability to determine whether or not the information is available in the first place.

According to the *doctrine of concordance* (see Tulving, 1989), cognitive processes, such as retrieval, behavior (i.e., the retrieval output), and the phenomenological experiences associated with these processes or behaviors are correlated. For example, according to concordance, examining one facet (e.g., behavior or what items are recalled on a test) can provide access to understanding the underlying process of retrieval and the phenomenological experience. Tulving further notes that much research in the field appears to rest on this assumption. However,

☆ JHC and SU contributed equally to the project and share first authorship. This work was supported by funding from a James McDonnell Foundation *Understanding Human Cognition* Grant awarded to JHC (#220020426). The funding agency had no input on the study design, data analysis, or writing.

* Corresponding author.
E-mail addresses: jhcoane@colby.edu (J.H. Coane), sumanath@cmc.edu (S. Umanath).

https://doi.org/10.1016/j.jml.2019.05.002
Received 17 August 2018; Received in revised form 10 May 2019
0749-596X/ © 2019 Published by Elsevier Inc.

cognitive processes, behavior, and phenomenology do not necessarily map directly on to one another but can be quite distinct. For instance, the cognitive process of retrieval (i.e., searching through memory for an item or event) is distinct from the experience of recollection, with the latter providing a qualitative description of the phenomenology associated with mental time travel, retrieval of specific details, and experience of re-living a specific episode or event (Tulving, 1985). Similarly, successful retrievals can be associated with distinct phenomenological states, as we discuss below. Thus, it is important to distinguish between the act of retrieval and the experience of retrieval. The question addressed in the present work is framed to parallel existing work examining the phenomenology of retrieval and concerns whether there are distinct phenomenological experiences associated with retrieval failures, whether individuals can describe these experiences in ways that discriminate them and importantly, whether they can use this information accurately to analyze the unavailability or inaccessibility of information.

Extensive research has examined the phenomenology associated with when an item is consciously retrieved from explicit/declarative memory (Squire, 1992). For example, information can be retrieved from semantic stores (i.e., general knowledge) or from event stores (Rubin & Umanath, 2015), more specifically episodic stores (i.e., specific events; Tulving, 1983).[1] Retrieval from these two stores reflects distinct types of remembered information and retrieval mode. According to Tulving (1985), retrieval from the former is associated with noetic consciousness and is reflected in a 'know' response, whereas retrieval from the latter is characterized by autonoetic consciousness and is reflected in a 'remember' response. Since Tulving (1985) original proposal, which associated distinct phenomenological experiences with retrieval from different memory systems, the use of the remember/know (R/K) paradigm (e.g., Gardiner & Java, 1990; Rajaram, 1993) has been a widely used tool for directly measuring the phenomenology of retrieval, in particular within episodic recognition tasks.

The key aspect of the R/K paradigm is that participants are reporting on the *feeling* or experience associated with the retrieval event. When used in the context of episodic retrieval tasks, when a *remember* response is given, they are said to be experiencing a recollective event, and the associated "phenomenal flavor" (Tulving, 1985, pg. 3) of mental time travel. Conversely, when a *know* response is given, participants are said to be experiencing a sense of familiarity, often in the absence of the recollective details associated with *remembering*. Although there have been challenges to how participants use and understand the terms *remember* and *know* (Migo, Mayes, & Montaldi, 2012; Umanath & Coane, submitted for publication), a core assumption of the R/K paradigm is that it can provide insight into the feelings associated with the cognitive process of retrieval. We note that, in the R/K literature, the use of *remember* and *know* responses has often been used to provide direct estimates of recollection and familiarity, respectively, thereby conflating phenomenology and cognitive processes (but see Wais, 2008; Wixted & Squire, 2011, for discussions). As suggested by the R/K paradigm, retrieval can be associated with distinct phenomenological experiences, and the terms used to describe retrieval success or retrieval failures can index the specific phenomenological qualities (what Tulving, 1985, referred to as types of consciousness).

In contrast, models of metacognition, such as Nelson and Narens' (1990, 1994), assume that cognitive operations and processes are distinct from the phenomenological ones. The latter monitor the output of the former and guide behavior. Much of the research on metacognition has examined when cognitive processes and phenomenological experiences do and do not align with one another. Often, the focus has been on the processes involved in how participants quantitatively evaluate their performance, either at the moment or in a predictive manner. During encoding, participants might be asked to make a judgment of learning (JOL; e.g., Jersakova, Allan, Booth, Souchay, & O'Connor, 2017) and assess on a numerical scale the extent to which they think they will be able to retrieve the learned information on a subsequent test. During a test, participants might be asked to report on a scale their feeling of knowing (FOK; e.g., Hart, 1965; Koriat, 1993, 1995), or the "feeling that one will be able to recognize – from a list of items – an item that is currently inaccessible" (Schwartz, 2006, pg. 153).

These monitoring and predictive tasks provide a wealth of information about when and how metacognitive processes operate and the degree to which they are accurate. Current theoretical approaches, broadly defined, assume that metacognitive predictions are based on inferential processes. That is, rather than providing direct access to the contents of memory (as Hart, 1965, had suggested), measures such as FOK are based on a number of factors that are utilized to make an inferential judgment: Familiarity with topic, retrieval of episodic information associated with the topic/event, cue familiarity, accessibility of features (Connor, Balota, & Neely, 1992; Hanczakowski, Pasek, Zawadzka, & Mazzoni, 2013; Klin, Guzmán, & Levine, 1997; Koriat & Levy-Sadot, 2001; Koriat, 1993, 1995; Reder & Ritter, 1992; Reder, 1987; Schwartz & Metcalfe, 1992; Thomas, Bulevich, & Dubois, 2011, 2012). FOKs are commonly elicited when retrieval fails (but see, e.g., Hertzog, Dunlosky, & Sinclair, 2010; Koriat, 1995) and provide a measure of how individuals monitor the contents of their memory and evaluate the probability that they can retrieve the information therein. However, FOKs do not provide a clear window into the *experience* participants have during a retrieval failure. In other words, many metamemory judgments are predictive measures of future performance and do not always directly measure the actual "feeling" of knowing or not knowing that a participant might experience in the exact moment in which retrieval fails. The aim of the present work is to provide an exploration into the phenomenology of different types of retrieval failure.

The best (and seemingly only) example of the phenomenology of retrieval failures comes from research on tip-of-the-tongue states (TOT; Brown, 1991; Brown & McNeill, 1966). TOTs are the phenomenological experience of knowing but being unable to retrieve at that specific moment and critically, are characterized by a sense of imminent retrieval, as if the sought-after item were just out of reach (see Schwartz, 1999; Schwartz & Metcalfe, 2014, for reviews). Conversely, referring back to the doctrine of concordance, the cognitive process is the retrieval attempt and immediately subsequent failure. This issue was discussed by Schwartz (1999), who posited a distinction between the *feeling* of inaccessibility and the *state* of inaccessibility. The former refers to the phenomenological experience of the individual and the latter to the objective failure to retrieve the target item while retrieving partial information about it. In other words, the TOT state refers to the "experience of inaccessibility rather than inaccessibility itself" (Schwartz, 2006, pg. 150).

Research on TOTs has provided a wealth of information about what participants can and cannot access while in a TOT state, as well as information about the overall accuracy of participants' judgments concerning subsequent retrieval success. Much research has focused on whether participants are accurate in making TOT judgments, what information is accessible when one is in a TOT state, and on the mechanisms and theoretical explanations involved (Brown, 1991; Schwartz, 1999). The evidence concerning what type of information is accessible to participants while in a TOT state has provided critical information about the nature of retrieval processes in memory and psycholinguistics (see Schwartz & Metcalfe, 2014, for a review). However, as Schwartz (1999) has argued, one core element of TOTs concerns the specific and universal phenomenological experience they elicit. The very expression "on the tip of the tongue" appears to be relatively universal: In an informal survey of 51 languages, he found that the vast majority of expressions used to refer to the feeling of

[1] Throughout the manuscript, when we reference "episodic memory," we use it as originally defined (i.e., Tulving, 1972, 1983) and re-contextualized within explicit memory in Rubin and Umanath (2015).

imminent retrieval in fact included the word "tongue." This suggests that not only is the experience common to many cultures, but that the word choice used to capture that feeling is naturally occurring and readily understood. He suggested, in fact, that TOTs are a "lens that brings phenomenology into focus, as opposed to a 'window' on the nature of retrieval" (Schwartz, 1999, pg. 391). From this perspective, the function of TOTs is to alert a person in this state that the information is present, even if it is temporarily inaccessible. Recent work by Cleary and colleagues (e.g., Cleary & Claxton, 2015; Cleary, 2018) further indicates that TOT states can be used to draw inferences about features and characteristics of non-retrieved information. Thus, the phenomenological experience is useful, not a flaw in the system (see Schacter, 2001); what is generally considered a marker of retrieval failure serves an adaptive function because it provides information to the rememberer (Schwartz & Metcalfe, 2011) in a way that can guide future behavior, similar to other metacognitive assessments. For example, the experience of a TOT might serve to alert an individual that the information is available in memory but that additional search time is needed.

However, much like successful retrieval from memory can elicit distinct phenomenological qualities, *failing* to retrieve from explicit memory might result in different experiential or evaluative judgments. Retrieval failures likely come in a variety of "flavors," with TOTs being only one. Retrieval failures that do not elicit a TOT are "qualitatively different states of mind" (Schwartz, 1999, pg. 381). Thus, not all information is going to be "imminent" – if nothing else because of the adaptive nature of memory systems. How can we capture the phenomenology of some of these other types of retrieval failure experiences? For example, when attempting to retrieve information from event memory or more specifically, episodic memory, and failing to do so, a participant might report that they "do not remember," whereas a failed retrieval attempt from semantic memory might elicit a "do not know" response. Thus, the negative versions of the classic R/K dichotomy could similarly map on to cases in which retrieval fails. This distinction, then, would mirror Tulving (1985) original distinction between the states of consciousness associated with retrieval from episodic and semantic memory.

However, failing to remember and failing to know might also reflect fundamentally different phenomenological experiences than assessments of retrieval failures from different memory stores. We hypothesize that failing to retrieve a fact might elicit a response of "I don't know" if the information is deemed to be not part of the knowledge base (i.e., not available). Then, a response of "I don't remember" conveys a very different meaning: The information might be not accessible at the moment or is determined to have been known at one point in time but since forgotten. Thus, *not remembering* and *not knowing* might reflect the phenomenology associated with retrieval and storage, experiential differences in availability or accessibility failures more broadly, and might index different causes of retrieval failure.

Overview of the present studies

Given that *remembering* and *knowing* appear to refer to distinct phenomenological states associated with successful retrieval, our initial question was simply whether not remembering and not knowing would be used by participants to describe distinct phenomenological experiences. Our hypothesis, as noted above, is that not remembering is used to describe the experience of not being able to access information at the moment and not knowing is used to describe the experience of not having the information in memory.

As an initial step, we queried participants about what they mean when they use the expressions "I don't know" (DK) or "I don't remember" (DR) to determine whether different phenomenological states are indeed described. Because cognitive and phenomenological processes are separate, it becomes important to see how they are correlated (e.g., whether and how much TOTs predict subsequent performance). Thus, we also examined how younger and older participants used these expressions during a general knowledge task, by asking them to respond DK or DR when actually experiencing a retrieval failure. If distinct phenomenological states are expressed by DK and DR, these should manifest differently in terms of behavior, much like subsequent retrieval is more likely for items ascribed to a TOT state than those in a non-TOT state. By modeling on the R/K paradigm, which focuses directly on the first-person perspective of retrieval at the moment of its occurrence, our goal was to determine whether the phenomenology of not remembering and not knowing is associated with distinct experiences associated with inaccessibility and unavailability. Thus, this approach differs from measures of metacognition such as FOKs or JOLs, which are often predictions of future performance in quantitative terms. We next conducted two experimental studies, described in detail below, examining the use of these expressions during retrieval attempts.

Some research has directly examined DK responses in the context of FOK judgments. DK responses can be based on lack of knowledge or lack of familiarity (Glucksberg & McCloskey, 1981) if the sought-after information is determined to be unknowable based on one's current knowledge state. DK responses might reflect two distinct processes (Glucksberg & McCloskey, 1981; Klin et al., 1997). Some DK responses follow an initial search through memory for relevant or related information: When such a search yields no answer, a rapid DK assessment is made. Conversely, when related or relevant knowledge is located in memory, a secondary evaluation process occurs, during which the retrieved content is examined. If this process fails to yield a satisfactory answer, a slower DK response is given. However, in most cases using the FOK measure, the DK response is given prior to the FOK prediction; thus, all failures to respond are initially categorized similarly (i.e., when participants respond DK before making an FOK, they might use DK to signify a failure to remember). Here, we were interested in whether participants would use DK or DR on different occasions to signal a retrieval failure due to different underlying causes.

A handful of studies have examined negative FOK, or "feeling of not knowing" (FOnK). Luo, Kazuhisa, and Luo (2003) found distinct neural correlates between accurate and inaccurate episodic FOK and FOnK. Similarly, Liu, Su, Xu, and Chan (2007) reported these metacognitive judgments also differed in terms of cognitive processes, with FOKs affected by the ease with which targets could be retrieved and FOnKs affected by the familiarity of the cues. Thus, deciding something is not known appears to be distinct from standard FOKs both neurally and cognitively. However, to our knowledge, the FOK and FOnK measures have only been compared in episodic tasks; whether similar findings extend to semantic retrieval tasks is an open question. We discuss the relationship between our measure and the rich literature on FOK in the General Discussion.

Interestingly, there appears to be little work in the way of examining use of DR as a response. To our knowledge, only one study, on the use of linguistic cues during retrieval failures, explicitly addressed this. Smith and Clark (1993) found that participants' response latencies to say DR were approximately seven times longer than latencies to say DK or "I have no idea." Furthermore, FOK judgments and subsequent recognition performance were higher for DR responses than for the other two response options. Smith and Clark speculated that the use of these expressions occurred in part as a means of saving face. In other words, participants attempted to convey to the experimenter that they were actively searching through memory during an extensive delay in responding. They further suggested that DR served to indicate that the answer was part of their knowledge base. Thus, in natural language and in conversational modes, participants spontaneously use expressions such as DK and DR to express a continuing search or the outcome of a failed search. One caveat, however, is that Smith and Clark only obtained 25 DR responses from 25 participants answering 40 questions; thus, additional research is needed.

Experiment 1

As an initial exploration into the phenomenology of not knowing and not remembering, we asked lay and expert participants to provide definitions of what they mean when they use the expressions "I don't know" and "I don't remember." The primary goal was to determine whether and how people use these terms differently and to what extent not knowing and not remembering reflect failures in availability, accessibility, or some other factor. The secondary goal was to assess whether and how expertise in memory or other fields of psychology training affected participants' understanding of these terms.

The expert samples were included because a vast literature indicates significant differences in performance between expert and novice participants across a variety of tasks (e.g., Novick, 1988). Furthermore, because domain knowledge can influence metacognitive judgments such as JOLs (e.g. Shanks & Serra, 2014) and hindsight bias (Knoll & Arkes, 2017), we determined it was important and interesting to assess the presence or absence of expertise-mediated differences in how not remembering and not knowing are defined. For example, use of the Remember/Know paradigm generally entails extensive instructions, and a number of researchers have expressed concerns with the use of the terms *remember* and *know* because of pre-experimental associations (see Umanath & Coane, submitted for publication). Thus, for experts in the memory domain, the expressions DR and DK could be similarly laden with meaning but may not align with a lay audience's understanding. Experts from related disciplines were included to ensure that any differences between memory experts and lay participants were not due to factors such as general familiarity with psychological concepts and constructs or additional years of education. Additionally, it is important to ensure that researchers who might use DR/DK experimentally and participants in their studies have a similar understanding of what the terms mean.

Method

Participants

Seventy-six lay participants (39 women) completed the survey online. The data were collected in June 2016 via Amazon Mechanical Turk, and participation was limited to individuals with a United States IP address. Participants tested through this online platform have been demonstrated to provide data that is similar in quality to laboratory samples (Mason & Suri, 2012). Their ages ranged from 18 to 69 (M = 36.37, SD = 12.69), and they averaged 14.84 years of education (SD = 2.14, range = 12–20). Participants were compensated $.30.

One hundred and seventy-eight participants provided complete responses to the online survey for experts. An additional 30 participants' data were omitted because they failed to complete the survey or did not answer the questions meaningfully (i.e., in response to the questions, they restated the question). Six participants were removed from analyses because they indicated having some college (N = 2) or a Bachelor's degree (N = 4) and were not considered experts. Of the remaining 172, 36 (20 women) identified as Memory Experts, 43 (27 women) as Other Experts in Cognitive Psychology, with a specialization in a field other than memory (e.g., categorization, language, attention), and 93 (73 women, one other) as Non-Cognitive Experts (e.g., clinical, social, developmental). Memory Experts had a mean age of 40.25 (SD = 11.88; range 24–74), Other Cognitive Experts had a mean age of 37.88 (SD = 10.74; range 24–84), and Other Psychology Experts had a mean age of 36.38 (SD = 6.61; range 23–54). The majority of participants in all groups had a terminal doctoral degree in their field (88.9% of the Memory Experts, 81.4% of the Other Cognitive Experts, and 76.3% of the Other Psychology Experts). Other respondents had Master's degrees (N = 16), some graduate school (N = 10), or professional degrees (e.g., MD, JD; N = 8). Expert participants were not compensated. The study was approved by the Institutional Review Board at Colby College. Participants provided consent online prior to beginning the task.

Materials and procedure

After providing consent and answering questions about their gender, age, and education level, participants were presented with three questions, one at a time. The questions asked "When you say 'I don't know' something, it's because", "When you say 'I don't remember' something, it's because", and "If you say, 'I don't know' something, how is that different from saying 'I don't remember' something?" The instructions specified there were no right or wrong answers but to answer the questions to the best of their ability.

Response coding

The answers to the question asking about differences between not knowing and not remembering were not analyzed because an initial examination revealed that they tended to repeat the same information as the other two questions provided. Furthermore, the comparative approach might have biased participants to evaluate the two types of phenomenological experience differently. Thus, we report the results of the analyses on the open-ended definitions of not knowing and not remembering.

All valid responses (i.e., we did not code responses that simply restated the question) were scored along a number of dimensions. For all dimensions, a score of 1 indicated the dimension was present and 0 indicated it was absent. See Table 1 for the coding scheme and examples of participant responses. Each participant's response was given a

Table 1
Dimensions Used in Coding Participant Responses to "I don't Remember" and "I don't Know" and Sample Answers in Experiment 1.

Dimension	Definition	Sample Answers
Accessibility	Response specifically mentioned that the inability to retrieve information was temporary/at the present moment and included tip of the tongue states or mentions of not being able to retrieve the specific word/item	You know the answer, but just don't have the answer on the top of your head at the present moment; I know the answer, but can't recall it at the time I'm asked
Never	Response included the explicit use of the word "never" or the response indicated information was never learned or encountered	It is something I never learned, like geography; I don't have the knowledge and won't have the knowledge
Forgetting	Response makes an explicit reference to the loss of information over time	It is something you have forgotten
Availability	Response indicated that the information is or was stored in memory at some point or that the information was never available	I had knowledge but have forgotten; I have no knowledge of it
Learned	Response referenced that the information was learned or encoded at some point or that the individual had been exposed to the information previously. It also included the opposite extreme: cases in which the information was never learned	When I say I don't know something it's because I don't have that information available to me. I never learned what has to be known
Confidence	Response included reference to confidence or certainty of answer. References to high and low confidence were accepted.	I was almost positive I learned the information
Detail	Response referred to the amount of detail associated with the memory or to the specificity, high or low, of the information	I can't remember the details

Note: Responses were coded using multiple criteria. For example, if a participant responded "I am sure never learned it" this included the criteria of *Never, Learned, Confidence,* and *Availability*. Analyses regarding *Learned, Confidence,* and *Detail* are provided in the Supplementary Materials.

score of 1 or 0 for each dimension. Thus, the proportions reported refer to the proportion of participants that included a reference to a specific dimension or to the proportion of responses that included that dimension. Critically, a single response could be coded on multiple dimensions. Here, we focus on a subset of responses to the dimensions that most clearly provided insight into participants' phenomenological experiences and that most directly addressed the distinction between availability and accessibility failures. The remaining analyses are available in the Online Supplement.

Results

Participants' responses were coded for all dimensions by two independent coders, and correlations between the two coders ranged from 0.78 to 1. Discrepancies were then resolved through discussion. In this and the remaining experiments, effect sizes for significant comparisons were calculated using partial eta squared (η_p^2) for analyses of variance (ANOVAs) and using Cohen's *d* for t-tests. Where appropriate, given violations of sphericity or equal variance assumptions, reported degrees of freedom reflect a correction (Greenhouse-Geisser for ANOVAs).

Seven 2 (Question: I don't remember, I don't know) × 4 (Group: Lay, Memory Experts, Other Cognitive Experts, Other Psychology Experts) ANOVAs were conducted to examine the relative inclusions of the dimensions described in Table 1 in answering what "I don't remember" and "I don't know" meant across the participant groups. We organize the results by the dimension discussed and summarize across findings at the end. Note that to ensure against alpha error due to multiple comparisons, an initial MANOVA was conducted including comparisons between Question and Group for all 7 dimensions discussed below and yielded a significant effect of Question, Hotelling's Trace = 1.38, $F(7, 224) = 44.21$, $p < .001$, $\eta_p^2 = 0.58$, a main effect of Group, Hotelling's Trace = 0.38, $F(21, 668) = 3.99$, $p < .001$, $\eta_p^2 = 0.11$, and an interaction, Hotelling's Trace = 0.188, $F(21, 668) = 1.99$, $p < .01$, $\eta_p^2 = 0.06$. Because this greatly reduces the concern of alpha error due to multiple comparisons, we proceed with presenting and interpreting each ANOVA and the follow-up *t*-tests conducted.

Accessibility

Regarding references to accessibility in participant responses, there were significant main effects of Question, $F(1, 230) = 82.33$, $MSE = 0.16$, $p < .001$, $\eta_p^2 = 0.26$, and Group, $F(3, 230) = 5.46$, $MSE = 0.19$, $p = .001$, $\eta_p^2 = 0.07$, but no interaction between the two, $F(2, 230) = 2.28$, $MSE = 0.16$, $p = .08$. All experts and lay participants referenced ideas related to accessibility more when explaining how they use "I don't remember" ($M = 0.52$, $SEM = 0.03$) than how they use "I don't know" ($M = 0.16$, SEM = 0.02). Following up on the main effect of Group, the one way ANOVA was significant, $F(3, 246) = 5.30$, $MSE = 0.10$, $p = .001$. For both questions, Lay participants were less likely to reference accessibility ($M = 0.20$, $SEM = 0.03$) than the three expert groups [$M_{NotCog} = 0.31$, $SEM = 0.03$, $t(168) = 2.66$, $SED = 0.04$, $p = .009$, $d = 0.42$; $M_{Cog} = 0.41$, $SEM = 0.05$, $t(118) = 3.81$, $SED = 0.06$, $p < .001$, $d = 0.70$; $M_{Memory} = 0.38$, $SEM = 0.06$, $t(50.56) = 2.51$, $SED = 0.07$, $p = .015$, $d = 0.55$] among whom there were no significant differences, $ps > .10$. Thus, these data indicate that accessibility tends to be associated much more with not remembering than not knowing and that experts reference this dimension more often than lay participants. Critically, not all responses that referenced accessibility described what might be defined as a TOT; there were several responses that were broader, including statements from the lay participants like "You know that something and if you thought hard enough you could remember it but at that second you can't remember that something," "I am quite certain that I once possessed that information but I can't bring it to mind. I obviously have memories that lead me to believe that I must have once had that information, but when I try to access it I don't find it," and "At the present moment, for whatever reason, I cannot recall a certain piece of information that I already possess and which with I am familiar."

Never

Regardless of expertise level, participants referenced "never" more in relation to "I don't know" ($M = 0.21$, $SEM = 0.03$) than for "I don't remember" ($M = 0.01$, $SEM = 0.01$) where they hardly referenced it at all, $F(1, 230) = 46.03$, $MSE = 0.09$, $p < .001$, $\eta_p^2 = 0.17$. There was no main effect for Group, $p = .11$, but there was a significant interaction, $F(3, 230) = 2.99$, $MSE = 0.09$, $p = .03$, $\eta_p^2 = 0.04$. When defining "I don't remember," there were no differences across groups regarding inclusion of never having encountered information before, $p = .39$. However, for "I don't know," a marginal difference emerged across groups, $F(3, 241) = 2.48$, $MSE = 0.16$, $p = .06$, indicating that Lay people included "never" significantly more often ($M = 0.30$, $SEM = 0.05$) than Other Psych Experts [$M = 0.15$, $SEM = 0.04$, $t(133.17) = 2.30$, $SED = 0.07$, $p = .02$ $d = 0.37$] and Cognitive Experts [$M = 0.14$, $SEM = 0.05$, $t(106.73) = 2.13$, $SED = 0.08$, $p = .036$, $d = 0.40$], but no more so than Memory Experts ($M = 0.25$, $SEM = 0.07$, $t < 1$). Additionally, there was no difference across the experts, $ps > .19$. These data indicate that a complete lack of exposure or experience or as memory researchers might say, availability, represents not knowing rather than not remembering.

Forgetting

Only the main effect of Question was significant for Forgetting; $ps > .21$ for Group and the interaction. That is, forgetting was referenced more when participants defined not remembering ($M = 0.37$, $SEM = 0.03$) than not knowing ($M = 0.02$, $SEM = 0.01$), for which it was included quite rarely, $F(1, 230) = 104.88$, $MSE = 0.11$, $\eta_p^2 = 0.31$, $p < .001$. This pattern held across participants, indicating that regardless of expertise, all participants considered forgetting only in relation to not remembering.

Availability

Availability was more frequently referenced when participants defined not remembering ($M = 0.84$, $SEM = 0.03$) than not knowing ($M = 0.75$, $SEM = 0.03$), regardless of Group, $F(1, 230) = 4.29$, $MSE = 0.17$, $p = .04$, $\eta_p^2 = 0.02$. Following up on the main effect of Group, $F(3, 230) = 3.06$, $MSE = 0.18$, $p = .03$, $\eta_p^2 = 0.04$, the one-way ANOVA was significant, $F(3, 246) = 3.88$, $MSE = 0.11$, $p = .01$. Overall, the cognitive experts (Memory, $M = 0.86$, $SEM = 0.05$, and Other Cognitive Experts, $M = 0.84$, $SEM = 0.04$, which were equivalent, $t < 1$) included references to availability more often than the other two groups who did not have their area of expertise – lay participants [$M = 0.71$, $SEM = 0.04$; vs. Other Cognitive Experts: $t(105.82) = 2.24$, $SED = 0.06$, $p = .03$, $d = 0.41$; vs. Memory Experts, $t(83.35) = 2.43$, $SED = 0.06$, $p = .02$, $d = 0.48$] and Other Psychology Experts [$M = 0.69$, $SEM = 0.04$, vs. Other Cognitive Experts: $t(101.74) = 2.70$, $SED = 0.06$, $p = .01$, $d = 0.48$; vs. Memory Experts, $t(76.72) = 2.87$, $SED = 0.06$, $p = .005$, $d = 0.54$], which were equivalent, $t < 1$. The interaction between Question and Group was not significant, $F < 1$.

In sum, across participant groups, the responses to the survey indicated that not remembering was characterized by failures in accessibility of stored material and by forgetting, whereas not knowing was associated with the information never having been encountered and thus reflected a failure in availability. Forgetting was almost never mentioned in the context of not knowing. Although availability was also more likely to be mentioned in regards to not remembering than to not knowing, it is possible that this reflects the fact that not remembering – or failing to retrieve something that is deemed available – is more likely to elicit an acknowledgement of the information's availability. Conversely, when something is not known, participants might be less likely to mention issues of availability because the

information is simply not available. As such, participants might not define the concept of not knowing in terms of what is absent (i.e., availability). Importantly, the overall clear pattern of responses indicates that across levels of prior knowledge and expertise, not remembering and not knowing are tapping into very distinct types of retrieval failure. As a next step, we empirically tested whether participants use these expressions to describe distinct types of retrieval failures and to what extent they are accurate in doing so (i.e., whether not remembering is more likely to reflect temporary accessibility than not knowing).

Experiment 2

To examine how participants used DR and DK during a retrieval attempt, we administered a general knowledge test, which depends on accessing information stored in semantic memory. Semantic memory is defined as the storehouse of general knowledge or crystallized intelligence (Cattell, 1963) and includes knowledge about facts, people, and events (Balota & Coane, 2008), vocabulary, schemas, work-related skills, and various types of expertise (Bowles & Poon, 1985; McIntyre & Craik, 1987; Staudinger, Cornelius, & Baltes, 1989; Umanath & Marsh, 2014). According to Tulving (1972) conceptions, unlike episodic memory, semantic memory is supposed to be relatively stable, even increasing over the lifespan as individuals accumulate an increasing body of knowledge. Given the breadth of knowledge stored in semantic memory, information is likely on a continuum of accessibility at any given point in time. When asked a question, participants might generate the answer, either readily or after a search through memory, or they might fail to retrieve the information. This retrieval failure could reflect either a complete absence of the relevant knowledge or a temporary inability to retrieve the information, or a range of options in between.

To examine whether participants can reliably use the distinction between not remembering and not knowing to assess the contents of their knowledge base, we presented general knowledge (GK) questions and asked participants to retrieve the correct answer. When they were unable to do so, they were given the option to indicate DR or DK. No additional instructions were provided on when to use one phrase or the other. After a brief delay, participants were given a follow-up multiple choice test.

To assess generalizability of the phenomenology of retrieval failures as captured by the expressions DR and DK, we included a sample of older adult participants. One of the hallmarks of cognitive aging is a reported increase in retrieval failures (Cavanaugh, Grady, & Perlmutter, 1983). Be it failing to remember a name or remembering fewer words in a laboratory study relative to younger adults, older adults generally perform worse on many traditional tests of episodic memory (Balota, Dolan, & Duchek, 2000). Semantic memory, conversely, appears to remain stable or even increase (e.g., Park, 2000; Salthouse, 2004). Older adults' increased body of knowledge provides additional schematic or retrieval support in episodic tasks (Umanath & Marsh, 2014). However, retrieval failures or at least difficulties persist in the form of increased tip-of-the-tongue states (Burke, MacKay, Worthley, & Wade, 1991). In terms of metacognition, older adults sometimes perform poorly, especially in episodic tasks (but see Hertzog, Sinclair, & Dunlosky, 2010), where their FOK judgments are less well-calibrated than those of younger adults. Their semantic FOK accuracy appears to be comparable to that of younger adults (Hertzog & Dunlosky, 2011; Lachman, Lachman, & Thronesbery, 1979; Morson, Moulin, & Souchay, 2015).

Including older adults also allowed us to examine individuals who are 'knowledge experts.' Older adults typically have increased knowledge relative to younger adults and also likely have more experience with retrieval processes in general and retrieval failures in particular. Thus, they might be more sensitive to the distinction between not knowing and not remembering than younger adults (see Umanath, 2016). Increased knowledge might yield a higher rate of DR responses, reflecting the fact that locating one specific unit of information in a larger store requires more time and effort and is more likely to fail. DR responses might also increase due to a greater presence of related information. Conversely, if Smith and Clark (1993) argument that saving face is a driver of DR and DK responses, older adults might be less likely to use DR to avoid being perceived as having a memory failure (Barber, 2017; Eich, Murayama, Castel, & Knowlton, 2014).

In terms of Tulving (1989) doctrine of concordance, we examined whether memory performance on the final test (i.e., successful or unsuccessful retrieval) varied as a function of the phenomenological experiences captured by the overt responses given during the initial task. Such differences would suggest that the phenomenology described by DR and DK might indeed be indicative of differences in retrieval processes and underlying cognitive states.

Method

Participants

To determine sample sizes, we calculated the effect size for the difference between phenomenological reports of accessibility given by lay participants in Experiment 1 in reference to DR and DK. We chose the dimension of accessibility because it clearly mapped onto a theoretically-driven phenomenological experience. The power analysis protocol in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated a sample of 34 was necessary. Thus, we aimed to recruit approximately 35 participants in each age group in the following experiments. Thirty-eight younger adults (M_{age} = 18.84, SD = 0.86, range = 18–21; $M_{education}$ = 12.6, SD = 0.86, range = 12–15) and 34 older adults (M_{age} = 69.88, SD = 5.6, range = 62–89; $M_{education}$ = 16.74, SD = 2.98, range = 12–24) were tested. Three younger adults did not complete the Shipley vocabulary test. Older adults scored higher than younger adults did on vocabulary knowledge (M = 35.21, SD = 4.23 and M = 30.46, SD = 3.3, respectively), $t(67) = 5.20$, $p < .001$, and had more years of education, $t(70) = 8.19$, $p < .001$. Younger adults were recruited from the participant pool at Colby College, and older adults were community-dwelling residents of the surrounding area. Students earned partial course credit towards their psychology classes, and older adults were compensated at a rate of $10/hr.

Materials and procedure

The experiment was administered using E-Prime software (Schneider, Eschman, & Zuccolotto, 2012). We selected 84 GK questions from Cantor, Eslick, Marsh, Bjork, and Bjork (2014), who reported an average accurate retrieval rate of 36% in a young adult sample, indicating the items were fairly difficult for this population. We added 16 items from Tauber, Dunlosky, Rawson, Rhodes, and Sitzman (2013) knowledge norms as easy fillers (i.e., these items had an average correct response rate close to 100%). All questions had a single correct answer of one or two words (e.g., *What was the name of the body of water where Thoreau studied for two years?* The correct answer, not displayed to participants, is *Walden Pond*). Participants were encouraged to answer all the questions and were told that guessing was acceptable. The questions appeared one at a time in random order, and the task was self-paced. Participants typed their response directly into the computer. They were encouraged to answer each question to the best of their ability or told they could respond "I don't know" (DK) or "I don't remember" (DR). No explicit instructions for when or how to use these options were given and if participants asked for clarification, they were told to use their best judgment. Upon completion of this phase, participants solved arithmetic problems or completed a Sudoku puzzle for 5 min. After the delay, a multiple choice test was administered. The same questions were presented in a new random order, each with five responses: the correct response and four plausible alternatives (e.g., the options for the question about where Thoreau studied were: *Walden Pond*, *Golden Pond*, *Lake Champlain*, *Lake Huron*, *Lake Wobegon*). The position of the correct answer varied across the five options an equal

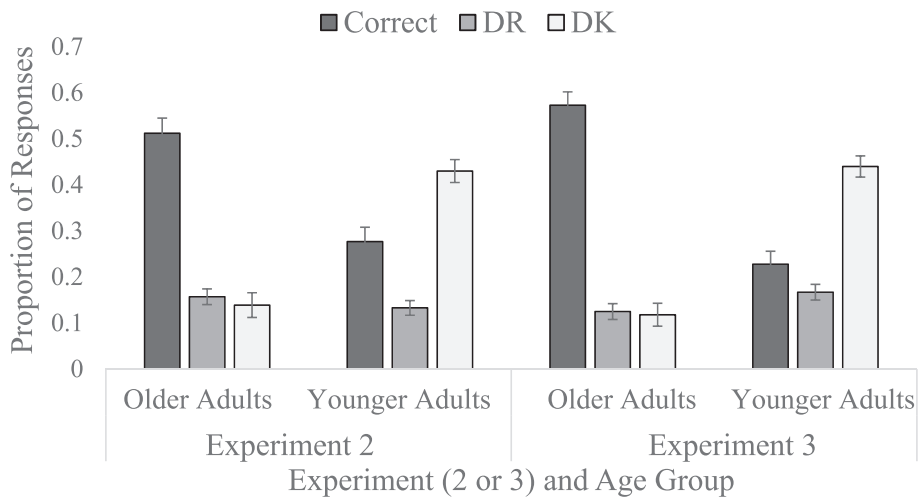

Fig. 1. Mean proportion of responses during the open-ended question task as a function of age in Experiments 2 (left panel) and Experiment 3 (right panel) (error bars reflect standard error of the mean).

number of times across items. Participants selected their response using the keyboard. After completing the forced choice task, participants were asked two open-ended questions to probe when and how they used the DR/DK response options during the initial phase. Specifically, the questions were "If you used "DK" [DR] in the first part of the study, what did you mean?" The question order was randomized. Finally, participants were administered the Shipley vocabulary task, thanked, and debriefed. Participation lasted less than an hour for most participants.

Results

Performance on initial GK test: open ended questions

Where appropriate, reported *p*-values for multiple comparisons have been adjusted following a Bonferroni correction. Only target items were included in the analyses. The open-ended questions were coded as incorrect (including errors and no answer given), correct (including minor spelling errors or morphological variations), DR, or DK. Performance was analyzed with a 3 (Response: Correct, DR, DK) × 2 (Age) mixed ANOVA. The effect of Age was not significant, $F(1, 70) = 1.68$, $p = .20$. The modal response was correct ($M = 0.39$, $SEM = 0.02$), $F(1.63, 96.01) = 33.84$, $MSE = 0.05$, $p < .001$, $\eta_p^2 = 0.33$. The second most frequent response was DK ($M = 0.28$, $SEM = 0.02$), followed by DR ($M = 0.14$, $SEM = 0.01$). Errors or no responses accounted for the remainder of cases. All pairwise comparisons were significant (all $ps < .045$).

Age and Response interacted, $F(1.83, 96.01) = 38.01$, $MSE = 0.05$, $p < .001$, $\eta_p^2 = 0.35$ (Fig. 1). Younger and older adults differed significantly in the proportion of correct responses, $t(70) = 5.22$, $p < .001$, $SED = 0.04$, $d = 1.22$, with older adults outperforming younger adults, and in the proportion of DK responses, $t(67.17) = -8.08$, $SED = 0.04$, $p < .001$, $d = -1.89$, where younger adults provided this response more than older adults. However, the two groups did not differ in the proportion of DR responses ($p = .32$). In sum, older adults were more knowledgeable than younger adults, consistent with other evidence showing preserved or increased semantic memory in aging (Umanath & Marsh, 2014).[2]

Standardized RTs during retrieval were analyzed using the same ANOVA. Raw response times were transformed to account for age-related differences in processing speed (Faust, Ferraro, Balota, & Spieler, 1999). If DR responses reflect a deficit in accessibility, they might be characterized by a longer search through memory than DK responses, which might be more driven by a lack of familiarity with the question or topic under consideration. RTs recorded initial key press following question onset (i.e., they included reading times, thereby further validating the use of standardized RTs).

Data from five older and three younger adults were omitted from the analyses because they never used one of the response options (DR or DK). The effect of Age was not significant, $p = .61$. There was a main effect of Response, $F(1.42, 87.78) = 42.371$, $MSE = 0.44$. $p < .001$, $\eta_p^2 = 0.41$. DR responses were the slowest ($M = 0.64$, $SEM = 0.10$), with RTs to correct responses ($M = 0.05$, $SEM = 0.05$) and DK responses ($M = -0.26$, $SEM = 0.04$) becoming faster. All pairwise comparisons were reliable (all $ps < .001$). Thus, DR responses did appear to reflect a search process that, when it failed or participants determined to stop searching, took longer than DK responses, which were the fastest. The Age by Response interaction was also significant, $F(1.42, 87.78) = 6.84$, $MSE = 0.44$, $p = .005$, $\eta_p^2 = 0.10$ (Fig. 2). The interaction was driven by the fact that older adults responded significantly faster than younger adults did to questions answered correctly, $t(64.92) = -4.09$, $SED = 0.09$, $p < .001$, $d = -0.98$, whereas the age effects for the other two responses were not significant (all $ps > .10$).

Performance on final multiple choice test

The proportion of correct responses on the final forced choice test were submitted to the same 3 (Response) × 2 (Age) ANOVA, where response referred to the response given to items during the open-ended question task. In other words, we performed an analysis of final test accuracy conditionalized upon performance in the initial GK question phase to address the question of whether participants could accurately assess failures in availability and accessibility. Because five older adults and three younger adults did not use DR or DK, the analyses include 29 older and 35 younger adults.

[2] In the analyses reported, we did not include null responses or items left blank. To address the concern that omission errors might have been the result of systematic biases such that some participants were more liberal or conservative in responding (a factor that would have affected the proportion of DR/DK responses given), we examined the number of omissions. Overall, they were very low (n = 215), accounting for 1% of responses across both experiments. The

(*footnote continued*)
majority of blank/no responses were made by five older adults, whose omission errors accounted for 67% of all omission errors (these participants made between 11 and 57 null responses). When the analyses were conducted omitting these five participants, none of the effects changed, suggesting that not responding did not seem to systematically reflect particular biases.

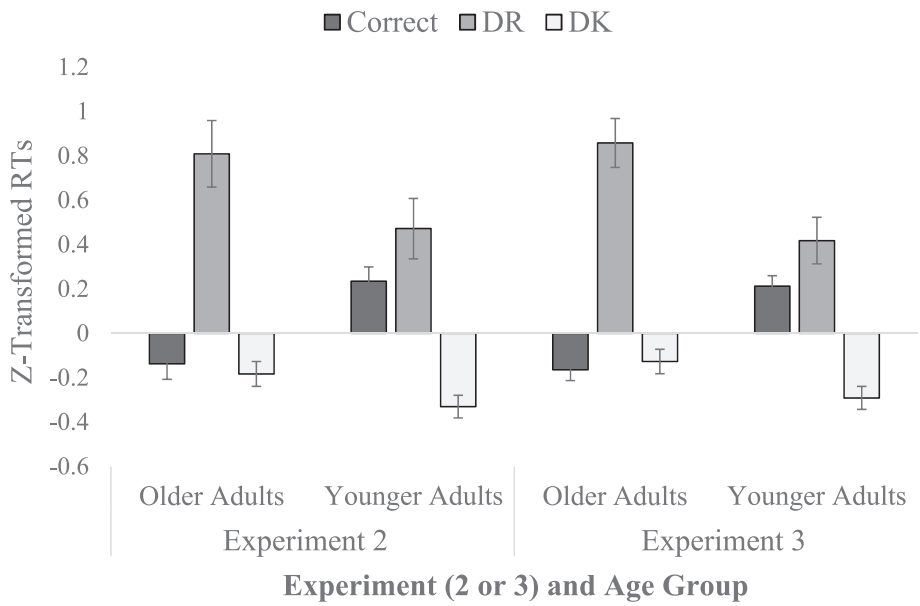

Fig. 2. Mean standardized response times on the open-ended question task as a function of age and response in Experiments 2 (left panel) and Experiment 3 (right panel) (error bars reflect standard error of the mean).

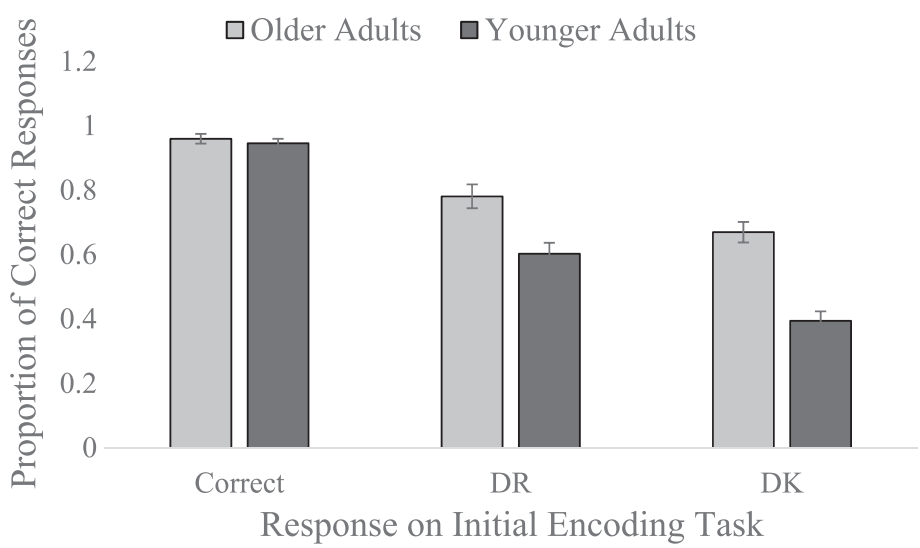

Fig. 3. Proportion of correct responses on the final multiple choice question task as a function of age and response during the encoding phase in Experiment 2 (error bars reflect standard error of the mean).

As during the encoding phase, older adults ($M = 0.80$, $SEM = 0.02$) correctly answered more questions than younger adults ($M = 0.65$, $SEM = 0.02$), $F(1, 62) = 31.08$, $MSE = 0.04$, $p < .001$, $\eta_p^2 = 0.33$. Importantly, accuracy on the multiple choice test varied as a function of response on the initial task, $F(2, 124) = 147.41$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = 0.70$. Initially correct items were recognized at a high rate ($M = 0.95$, $SEM = 0.01$), followed by items that were not remembered ($M = 0.69$, $SEM = 0.02$) and those that were not known ($M = 0.53$ $SEM = 0.02$). All pairwise comparisons were significant (all $ps < .001$).

The Age by Response interaction was significant, $F(2, 124) = 14.16$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = 0.19$ (Fig. 3). Performance on initially correct items did not differ as a function of Age ($p = .46$); however, older adults significantly outperformed younger adults on initially not remembered items, $t(52.13) = 3.72$, $SED = 0.05$, $p < .001$, $d = 0.91$, and on initially not known items, $t(66) = 6.39$, $SED = 0.04$, $p < .001$, $d = 1.53$. Importantly, both older and younger adults were more accurate on items that were initially classified as not remembered than not known, $t(28) = 2.94$, $p = .006$, $d = 0.58$, and $t(34) = 5.61$, $p < .001$, $d = 1.05$, respectively. Thus, a DR response was given to items that were temporarily inaccessible, whereas DK responses were more likely to be given to information that was not available.

Although both younger and older adults did correctly identify more items that were initially not remembered compared to those that were initially not known, both groups performed above chance (with five response options, chance was set at 0.20) even on not known items ($t[29] = 12.92$, $p < .001$, $d = 2.35$, and $t[37] = 9.03$, $p < .001$, $d = 1.5$, for older and younger adults respectively), and older adults in particular performed quite well on these items (almost 70% correct). This might reflect that older adults were under-estimating their knowledge or that they forgot they knew the information but were reminded of it at the final test. Alternatively, older adults might have developed more sophisticated or effective guessing strategies or may have been able to exclude foils on the basis of related prior knowledge. These possibilities are considered in the General Discussion. It is important to note, however, that the conditional analyses were based on a

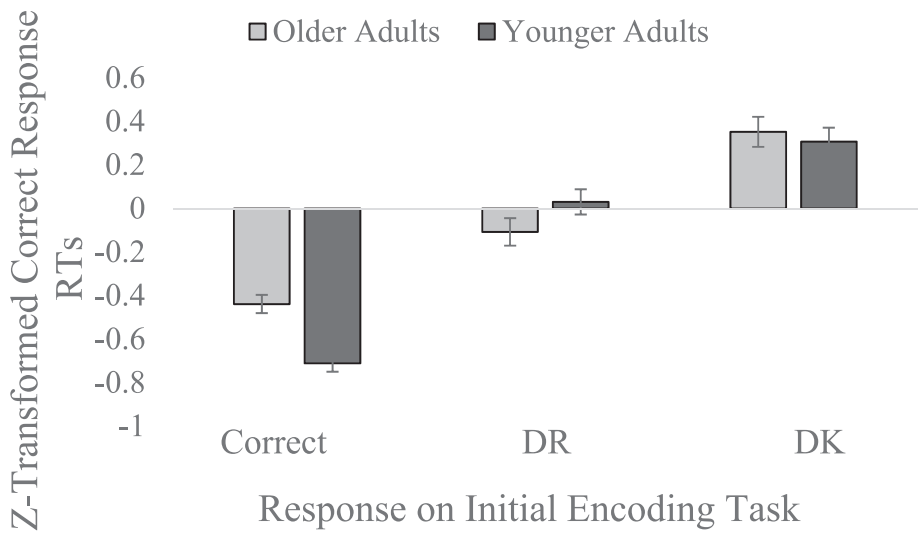

Fig. 4. Mean standardized response times on the multiple choice question task as a function of age and response in Experiment 2 (error bars reflect standard error of the mean).

relatively small number of items: Older adults only gave a DR response 16% of the time and a DK response 14% of the time (approximately 13 and 12 items out of the 84 target questions, respectively).

Standardized correct response latencies on the final multiple choice as a function of responses during the encoding phase were submitted to the same 3 × 2 ANOVA (see Fig. 4). The effect of Age was not significant, $p = .24$. Correct responses to initially correctly answered items were fastest ($M = -0.58$, $SEM = 0.03$), followed by items initially not remembered ($M = -0.04$, $SEM = 0.04$), and items initially not known were slowest ($M = 0.33$, $SEM = 0.05$), $F(1.67, 102.20) = 139.96$, $MSE = 0.11$, $p < .001$, $\eta_p^2 = 0.70$. All pairwise comparisons were significant (all $ps \leq .001$). Importantly, initially not remembered items were correctly recognized faster than initially not known items, confirming that not remembering was associated with a temporary failure in accessibility. The slow correct RTs to not known items might reflect the use of strategies or reasoning processes (i.e., participants could narrow the pool of correct responses by use of sophisticated guessing or excluding items they identified as incorrect).

The interaction between Age and Response category was significant, $F(1.68, 102.20) = 7.13$, $p = .001$, $MSE = 0.11$, $\eta_p^2 = 0.10$. Although both younger and older adults' correct response latencies followed the same ordering from fastest to slowest, the difference between younger adults' RTs to correct items and not remembered items was approximately twice as large as the difference for older adults. In terms of age effects, the only significant difference was for correct items; younger adults were faster than older adults at correctly selecting the correct response for those items they had initially answered correctly, $t(70) = 5.04$, $SED = 0.05$, $p < .001$, $d = 1.20$ (ps for DR and DK > 0.10).

In sum, the results of Experiment 2 indicate that both younger and older adults used DR for information that was subsequently correctly recognized at a higher rate than when they used DK. Furthermore, DR responses were slower in the initial encoding phase, presumably reflecting a search through memory, but subsequent correct answers to not remembered items were faster on the final multiple choice questions, where the correct answer was presumably identified more rapidly due to its greater accessibility. Although both age groups appeared to underestimate their actual knowledge, as reflected by the above chance performance on DK items on the final test, these data provide convincing initial evidence that a simple discrimination between DR and DK, which capitalizes on natural language use and reflects different phenomenological states, effectively allows individuals at different levels of knowledge expertise to make evaluate the contents of their memory.

Experiment 3

In Experiment 3, we adopted a slightly different paradigm to further explore how participants used responses of not remembering and not knowing. There were two major changes to the procedure used in Experiment 2. First, after answering the initial open-ended question, participants were given correct answer feedback. Second, instead of a final multiple choice test, we administered an open-ended cued recall test, which requires more generative processes. If DR responses reflect a temporary failure in accessibility, providing corrective feedback should serve as a refresher of existing knowledge and result in higher accuracy on the final test. Conversely, DK responses should reflect an initial learning episode as opposed to a reminder of previously known information and thus still result in lower performance on the final test (as in Experiment 2). In other words, the cued recall test allowed us to assess differences between re-learning or refreshing knowledge and new learning, by assuming that a DK response would reflect a lack of prior knowledge and DR a failure to retrieve the prior knowledge. This would extend the utility of relying on phenomenological experiences to discriminate between different cognitive states or retrieval processes beyond recognition performance to rates of learning from feedback. Whereas older adults performed as well as or better than younger adults on the recognition test in Experiment 2, it is possible that the more effortful processes required by cued recall tasks might allow age effects to emerge, given well-documented deficits in older adults' performance on tasks that require more controlled and self-generated processes (Craik & Byrd, 1982; Morris, Craik, & Gick, 1990). Furthermore, given the ample literature on age-related deficits in new learning (e.g., Old & Naveh-Benjamin, 2008), older adults might be at a particular disadvantage when presented with information they do not know.

Method

Participants

Thirty-six younger adults ($M_{age} = 18.69$, $SD = 0.82$, range = 18–21; $M_{education} = 12.53$, $SD = 0.77$, range = 12–14) and 33 older adults ($M_{age} = 68.91$, $SD = 7.29$, range = 60–96; $M_{education} = 15.97$, $SD = 2.57$, range = 11–20) from the same participant pool as the previous experiment were tested. None of the participants had been tested in Experiment 2. Data from two additional young adults were lost due to computer failures, and two young adults were missing scores on the Shipley vocabulary test. Again, younger adults ($M = 30.97$, $SD = 3.97$) scored lower than older adults

(M = 35.61, SD = 3.54) on vocabulary, $t(65) = 5.03$, $p < .001$ and had fewer years of education, $t(67) = 7.68$, $p < .001$.

Materials and procedure

The initial open-ended question task was the same as in Experiment 2, with the exception that after entering their response, participants were shown the correct answer for 2 sec before advancing to the next question. The final test consisted of the same open-ended questions, without feedback, in a random order, and participants entered their responses on the computer.

Results

Performance on initial GK test: open ended questions

Performance on the initial questions was analyzed in a 3 (Response) × 2 (Age) mixed ANOVA. The effect of Age was not reliable, $F < 1, p = .44$. Overall, correct answers accounted for 0.40 (SEM = 0.02) of responses, followed by DK (M = 0.28, SEM = 0.02), and DR (M = 0.15, SEM = 0.01), $F(1.69, 112.94) = 40.62$, $MSE = 0.033$, $p < .001$, $\eta_p^2 = 0.38$. All pairwise comparisons were significant (all $ps \leq .002$). Age and Response interacted, $F(1.69, 112.94) = 70.39$, $MSE = 0.033$, $\eta_p^2 = 0.51$ (Fig. 1). Consistent with Experiment 2, older adults were more likely to correctly answer questions than younger adults, $t(67) = 8.50$, $SED = 0.04$, $p < .001$, $d = 2.04$, and younger adults responded DK more than older adults, $t(60.99) = -9.65$, $SED = 0.03$, $p < .001$, $d = 2.31$. The difference in the proportion of DR responses was not significant, $t(67) = -1.72$, $p = .09$.

For the analyses on standardized response times, data from six younger adults were omitted from analyses because they never used a given response option. The effect of Age was not significant, $F(1, 62) = 1.44$, $p = .23$. There was a main effect of Response category, $F(1.50, 91.25) = 70.38$, $MSE = 0.23$, $p < .001$, $\eta_p^2 = 0.54$. DR responses were again the slowest (M = 0.64, SEM = 0.08), whereas correct responses (M = 0.02, SEM = 0.03) and DK responses (M = −0.21, SEM = 0.04) were faster. All pairwise comparisons were reliable (all $ps < .001$). The Age by Response interaction was also significant, $F(1.50, 91.25) = 15.84$, $MSE = 0.23$, $p < .001$, $\eta_p^2 = 0.21$ (Fig. 2). Older adults again retrieved the correct answer faster than younger adults, $t(52.85) = 5.85$, $SED = 0.07$, $p < .001$, $d = -1.39$, but were slower to say DR, $t(61) = 2.88$, $SED = 0.15$, $p = .005$, $d = 0.72$, suggesting they were engaging in a more extensive search through memory. The age difference in response latencies to say DK was not significant, $t(35.14) = 1.89$, $p = .07$. Overall, the pattern of performance in Experiment 3 was highly similar to that observed in Experiment 2.

Performance on final open-ended GK questions

Because our primary interest was in whether participants' use of DR and DK during the encoding phase was predictive of performance on the final test, we examined the proportion of correct responses on the final test as a function of initial response (DR vs. DK) and age as we did in Experiment 2. Older adults (M = 0.72, SEM = 0.04) were overall more accurate than younger adults (M = 0.52, SEM = 0.04), $F(1, 61) = 11.57$, $MSE = 0.11$, $p = .001$, $\eta_p^2 = 0.16$. Importantly, overall accuracy was higher following an initial DR response (M = 0.68, SEM = 0.03) than an initial DK response (M = 0.56, SEM = 0.03), $F(1, 61) = 30.21$, $MSE = 0.015$, $p < .001$, $\eta_p^2 = 0.33$. Thus, consistent with Experiment 2, participants' metacognitive assessments of the accessibility or availability of information were indeed predictive of subsequent performance. This main effect, however, was qualified by a significant interaction, $F(1, 61) = 15.59$, $MSE = 0.015$, $p < .001$, $\eta_p^2 = 0.20$. As is evident in Fig. 5, older adults only showed a small (M = 0.03, SEM = 0.03) and not significant accuracy advantage for DR over DK items, $t(29) = 1.01$, $p = .32$, whereas younger adults showed a robust (M = 0.21, SEM = 0.03) accuracy advantage for DR over DK items, $t(32) = 7.28$, $SEM = 0.03$, $p < .001$, $d = 1.12$.

Thus, in contrast to Experiment 2, in which both age groups performed better on items originally classified as DR than those classified as DK, when participants were given feedback, the effect only was present in younger adults. One possibility is that older adults underestimated their knowledge in the encoding task or that the knowledge assessed by the GK questions was more familiar to them or easier to integrate into their existing knowledge base. Another possibility is that older adults were more curious and motivated to learn the answers, and thus encoded them more deeply than younger adults (e.g., McGillivray, Murayama, & Castel, 2015). Clearly, these results suggest that older adults are not always at a disadvantage when it comes to learning or relearning information, even when testing conditions are similar to those in which older adults often show larger deficits in performance. We return to these points in the General Discussion. Additional analyses of performance on the final test, specifically on the frequency of DR and DK responses during the final cued recall task, are available in the Online Supplement.

Standardized response latencies to correct responses on the final test were analyzed using the same 3 (Initial Response) × 2 (Age) mixed ANOVA. The analyses included responses from 28 older and 32 younger adults. The effect of Age was not significant, $F < 1.0$. Response latencies to initially correct responses were fastest (M = −0.14, SEM = 0.03), followed by RTs to initially not remembered items (M = −0.06, SEM = 0.06). Initially not known items (M = 0.25, SEM = 0.05) were slowest, $F(1.80, 104.69) = 18.78$, $MSE = 0.135$, $p < .001$, $\eta_p^2 = 0.24$. All pairwise comparisons were reliable, $ps \leq .034$. Age and Initial Response interacted, $F(1.80, 104.69) = 8.43$, $MSE = 0.135$, $p = .001$, $\eta_p^2 = 0.13$. As can be seen in Fig. 6, the interaction is driven by the fact that older adults were faster at giving the correct answer to items originally given a DR response but much slower at giving the correct answer to items given a DK response, whereas younger adults did not differ in the RTs to items initially classified as DR and DK.

This interpretation of the interaction was supported by follow-up one-way ANOVAs on RTs to correct responses as a function of initial task performance within each age group. For older adults, the effect of Initial Response was significant, $F(2, 54) = 15.17$, $MSE = 0.161$, $p < .001$, $\eta_p^2 = 0.36$. Correct responses to initially correct and initially not remembered items did not differ ($p = .73$), but both were significantly faster than latencies to initially not known items ($ps \leq .004$). Thus, older adults retrieved the answer to items initially correctly answered and those that were initially inaccessible at the same speed, but searched longer before giving the correct answer to not known items. For younger adults, the effect of Initial Response was also reliable, $F(2, 62) = 8.61$, $MSE = 0.087$, $p = .001$, $\eta_p^2 = 0.22$. RTs to correctly answer items that were correct during the encoding phase were significantly faster than DR and DK items (both $ps \leq .004$), whereas the latter did not differ from one another ($p \geq .99$). This differential effect of response on latencies suggests that one interpretation of the high level of accuracy observed for DK items might be that older adults simply spent more time trying to retrieve the answer from memory after having learned it during the encoding phase.

Additional analyses

We examined whether the participants in Experiments 2 and 3 held generally similar conceptions of what "I don't know" and "I don't remember" meant as the lay participants in Experiment 1. For the lay participant group, 76 naïve individuals participated (see *Experiment 1: Participants* for more details). These individuals' responses were compared to those of all participants, younger and older, in Experiments 2 and 3. Note that the latter group defined not remembering and not knowing at the very end of the studies after having used these terms without instructions (see *Experiment 2: Materials and Procedure* for details). Thus, this group had actually utilized these terms during the task. Experiment 2 and 3 participants' responses were coded for all

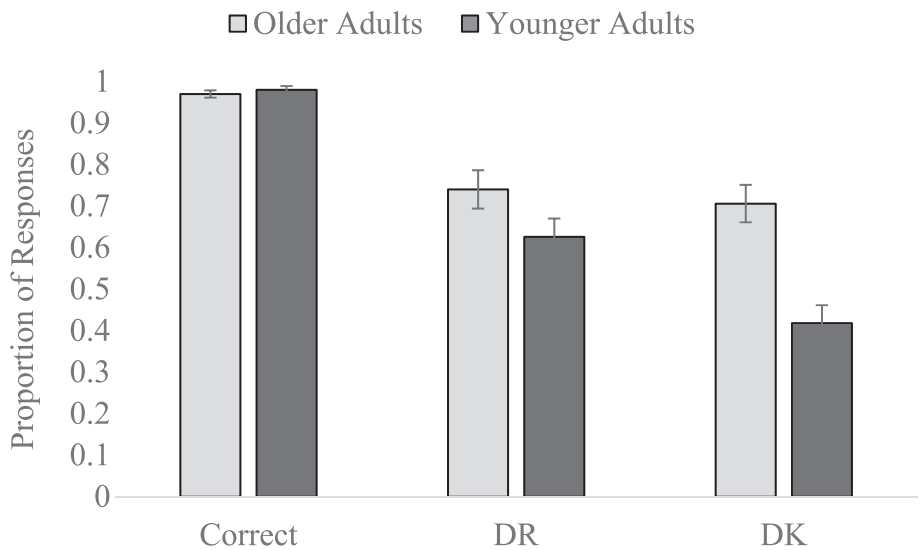

Fig. 5. Proportion of correct responses on the final open-ended task as a function of age and response during the encoding phase in Experiment 3 (error bars reflect standard error of the mean).

dimensions by two independent coders, and correlations between the two coders ranged from 0.89 to 1. Discrepancies were then resolved through discussion. Seven 2 (Question: I don't remember, I don't know) × 2 (Group: MTurk, Experiment Participants) ANOVAs were conducted to examine the relative inclusions of the dimensions in Table 1 in answering what "I remember" and "I know" meant between the participant groups. There were no influences (main effects or interactions) of which study participants had been in, Experiment 2 versus 3, so the analyses collapse over that factor. For the sake of brevity, only the analyses regarding the major dimensions from Experiment 1 are discussed in detail.

The analyses showed a remarkable consistency in terms of relative proportions of use. Specifically, DK responses were associated with never having learned or encountered the information, whereas DR responses were associated with forgetting and accessibility failures. Consistent with Experiment 1, all participants referenced "never" more in relation to DK ($M = 0.47$, $SEM = 0.04$) than to DR [$M = 0.007$, $SEM = 0.006$, $F(1, 173) = 159.77$, $MSE = 0.12$, $p < .001$, $\eta_p^2 = 0.48$], where they rarely referenced it at all. Forgetting was referenced more when for not remembering ($M = 0.36$, $SEM = 0.04$) than not knowing [$M = 0.03$, $SEM = 0.02$, $F(1, 173) = 66.36$, $MSE = 0.13$, $p < .001$, $\eta_p^2 = 0.28$], for which it was included quite rarely. Regarding references to Accessibility, both groups referenced ideas related to accessibility more when explaining how they used DR ($M = 0.63$, $SEM = 0.03$) than how they used DK [$M = 0.05$, $SEM = 0.02$, $F(1, 173) = 309.57$, $MSE = 0.09$, $p < .001$, $\eta_p^2 = 0.64$].

Although the MTurk participants were asked these questions out of context and without having experienced retrieval attempts or failures, experimental participants had just completed a difficult set of general knowledge questions. This experience magnified the distinction between not remembering and not knowing, especially for accessibility and forgetting, and never, three of the core dimensions associated with DR and DK, respectively. For Never, Experiment participants referenced this dimension overall more often ($M = 0.32$, $SEM = 0.02$) than did the MTurk participants [$M = 0.16$, $SEM = 0.03$, $F(1, 173) = 20.78$, $MSE = 0.11$, $p < .001$, $\eta_p^2 = 0.11$]. However, when defining DR, there were no differences across groups regarding inclusion of never having encountered information before; they very rarely referenced it at all ($Ms = 0.01$, $p = .20$). In contrast, for DK, Experiment participants included "never" more often ($M = 0.65$, $SEM = 0.05$) than MTurk

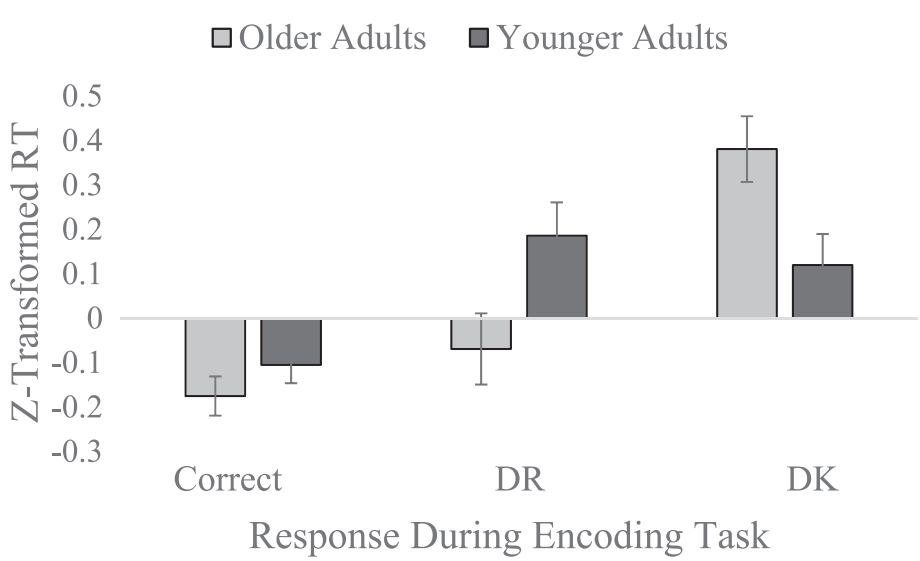

Fig. 6. Mean standardized response times for correct responses on the final open-ended task as a function of age and response in Experiment 3 (error bars reflect standard error of the mean).

participants [$M = 0.30$, $SEM = 0.06$, $t(179) = 4.84$, $SED = 0.07$, $p < .001$, $d = 0.75$]. For Forgetting, like almost all the other dimensions, Experiment participants referenced this dimension more often ($M = 0.26$, $SEM = 0.03$) than did MTurk participants [$M = 0.14$, $SEM = 0.03$, $F(1, 173) = 8.90$, $MSE = 0.13$, $p = .003$, $\eta_p^2 = 0.05$]. There was no interaction, $p = .22$. Experiment participants included much more information in their responses regarding accessibility ($M = 0.48$, $SEM = 0.02$) than did the MTurk participants [$M = 0.21$, $SEM = 0.03$, $F(1, 173) = 70.55$, $MSE = 0.09$, $p < .001$, $\eta_p^2 = 0.29$]. This effect was qualified by an interaction, $F(1, 173) = 86.99$, $MSE = 0.09$, $p < .001$, $\eta_p^2 = 0.34$, indicating that though there was no difference between the two groups for how much they referenced accessibility when they defined not knowing, which was very little ($M_{MTurk} = 0.07$, $SEM = 0.02$, $M_{Experiment} = 0.03$, $SEM = 0.02$, $t < 1$), the Experiment participants referenced it far more regarding not remembering ($M = 0.92$, $SEM = 0.04$) than did the MTurk participants [$M = 0.34$, $SEM = 0.04$, $t(198) = 10.56$, $SED = 0.05$, $p < .001$, $d = 1.48$].

These data confirm that these terms map onto natural language use and are intuitively understood by research participants, without the need for extensive instruction. Recent experience with retrieval, if anything, increases participants' intuitive understanding of the differences between retrieval failures due to lack of stored knowledge or lack of access.

General discussion

The goals of the research reported here were examine differences in phenomenological experience between failures in availability and failures in accessibility and how participants can use these terms to express and assess the content of their memory. Our rationale was that natural language use – specifically, the different ways in which people use the expressions "I don't know" and "I don't remember" – are associated with different phenomenological states which in turn reflect judgments about whether a to-be-remembered item is stored in and/or able to be retrieved from memory. Participants were expected to say "I don't know" when they failed to answer a question because of a storage failure whereas use of "I don't remember" was expected to reflect a retrieval failure.

An initial assessment of how lay and expert participants understand and use the expressions "I don't know" and "I don't remember" confirmed that not knowing was primarily associated with the dimension coded as Never: The information had simply never been encountered or encoded and, as such, was not available in the memory store. Conversely, not remembering was primarily associated with forgetting and inaccessibility. Thus, when participants said they did not remember they were indicating that they were aware that they currently or previously had stored the information, but could not retrieve it or a specific detail at the moment. Notably, references to accessibility did not necessarily draw on traditional phenomenological experiences of TOT states. Experimental participants from Experiments 2 and 3 who provided a definition of not remembering and not knowing at the end of the experimental task gave highly similar responses to lay participants from Experiment 1. The key difference was that recent experience with retrieval failures appeared to magnify the difference between these two states of metacognitive awareness. Thus, the dimensions selected clearly discriminated between the two response categories.

In two experiments, using different final tests, initial use of DR was associated with better performance on the final test than initial use of DK, with one exception, discussed below. Thus, DR responses were more likely to reflect initial accessibility or retrieval failures than DK responses, which were more likely to reflect the fact that the information could not be retrieved because it was not stored in memory. The effect of initial classification was quite large, with Cohen's d values ranging from 0.58 to 1.12 and η_p^2 values of 0.33 to 0.37. This suggests that our measure, which required no initial training or instruction, was effective at discriminating between different phenomenological experiences associated with memory performance. This effect was not only robust, but largely consistent across ages, with younger and older adults showing similar sensitivity to the distinction between DR and DK. Even though older adults outperformed younger adults in overall accuracy, the persistence of the effect across levels of knowledge expertise confirms its robustness and generalizability.

Across the three studies reported here, we have provided converging evidence that the cognitive states of *not knowing* and *not remembering* differ in noteworthy ways: First, overt explanations of what these expressions mean clearly discriminate the two states in terms of phenomenological experiences; second, they are associated with different levels of recognition performance; and third, they are also associated with different levels of learning from feedback. To frame this in the context of the doctrine of concordance (Tulving, 1989), the same behavioral output – i.e., a retrieval failure – can be associated with very distinct phenomenological and cognitive states, thus reinforcing the importance of examining processes, phenomenology, and behavior separately.

DR/DK vs. FOK

Although in many ways our measure shares some features with the standard FOK measure, there are some important differences. As the very name implies, in the FOK task, one has the "feeling" that some sought-after piece of information is known, but cannot be retrieved (see Thomas, Lee, & Hughes, 2015, for a review). However, we contend that, as typically administered, the FOK is relatively neutral with regards to capturing the actual phenomenological experience of participants. A key difference between our measure and the FOK is that, although lower FOK values presumably reflect not knowing and higher values not remembering, this is not explicitly stated, and participants might have different understandings of what the values on a scale reflect. Specifically, a FOK of 0 or responding "I am absolutely sure I do not know the answer" (Marquié & Huet, 2000) might reflect a lack of knowledge or unavailability, whereas a response of 100 or "I am absolutely sure I know the answer" seems to reflect a TOT (Bahrick, Baker, Hall, & Abrams, 2010) and temporary inaccessibility. Intermediate responses, whether on a six-point scale or on a 0–100 or 1–10 rating, ought to then reflect different levels of feelings of knowing. However, it is not clear from most FOK scales (e.g., *sure I would not recognize the answer* to *sure I would recognize the answer*; Singer & Thiede, 2008), what ascribing intermediate values really means. Furthermore, responding *sure I would not recognize the answer* could indicate either a failure of accessibility, if the information was forgotten, *or* a failure in availability, if the information had never been learned. Similar issues can arise in TOT research, where some participants gave a rating of 4 on a 6-point scale to items that were comparable to non-TOT trials, whereas others used the same rating for TOT-type trials, suggesting that multipoint scales might be inconsistently used by participants and are not always transparent (Bahrick et al., 2010).

As noted by Thomas et al. (2015), "FOK judgments influence subsequent control processes. Individuals select specific cognitive operations based on their subjective assessments of knowing or not knowing" (pg. 85). According to this and similar positions, what FOKs do not clearly provide is the distinction between not *knowing* and not *remembering* or between the phenomenology of accessibility and availability failures. Instead, FOKs are predictive assessments of future performance. Similarly, the literature on TOTs does not directly discriminate between these two types of failures but focuses on a subset of retrieval failures clearly attributed to deficits in accessibility.

The DR/DK measure presented here represents an initial attempt to develop a tool to simply and intuitively discriminate between storage and retrieval failures as two independent sources and causes of memory failures by relying on phenomenological assessments. Thus, this measure is complementary to FOKs by providing information about a

different type of judgment. Whereas the sensitivity of FOKs to cue or question familiarity might be problematic, in that familiarity inflates estimates of metamemory but has no direct effect on actual memory performance (Hanczakowski et al., 2013), a prediction from the DR/DK measure is that familiarity should lead to a DR response, because not remembering can reflect either a failure to retrieve information due to temporary inaccessibility or to trace decay. A next step in this research is to provide a direct comparison between FOK and DR/DK.

Predictive accuracy of DR/DK responses

As noted above in the quote by Thomas et al. (2015), metamemory judgments influence control processes, which in turn determine behavioral outcomes such as whether to continue studying or not (e.g., Hanczakowski, Zawadzka, & Cockcroft-McKay, 2014; Metcalfe & Finn, 2008) or whether to continue to search for an item in memory (e.g., Koriat, 1993; Singer & Thiede, 2008). Take the example of a student preparing for an upcoming exam: Failing to accurately assess that she is under-prepared will result in a poor grade or score. Thus, the student's academic performance depends not only on how she has studied (e.g., Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013), but also on her ability to make an accurate metacognitive assessment and determine to what extent the to-be-remembered material is not only stored but also accessible. Such an assessment following a retrieval failure can lead to additional study time or other strategies to ensure learning occurs.

Thus, metamemory involves not only an assessment of one's memory contents and function, but also the use of that information to guide and control behavior. Accurate metamemory assessments can support effective use of one's time and cognitive resources. The accuracy of such assessments is critical: When information is not known, additional learning is necessary but additional searching is unlikely to be fruitful; conversely, when information is not remembered, additional searches or use of cues is likely to yield a positive outcome, but additional learning will benefit performance. Although metamemory judgments are used for guiding subsequent behavior, the research on how FOKs result in changes in control processes is somewhat scarce (Hanczakowski et al., 2013).

In Experiment 3, the inclusion of post-retrieval feedback provided a potential framework for examining how these decisions might guide future control processes, by allowing participants to learn or re-learn the information. Although in the present study feedback was presented for the same duration regardless of response accuracy to all participants, future studies might include manipulations such as giving the option of seeing the correct answer, allowing participants to regulate the time they spent with feedback, or asking them, after the feedback whether they knew the answer before (i.e., whether it was an accurate DR or DK response).

Although DR and DK responses were associated with different levels of accuracy on the final tests, participants generally performed significantly above chance on items they gave an initial DK response to, with older adults in particular performing quite well on these items. This might reflect a number of factors: A face-saving mechanism (Smith & Clark, 1993), where a lack of knowledge is judged to be less threatening to one's sense of self and less anxiety provoking than a retrieval failure; more sophisticated guessing during the final multiple choice phase; the fact that information that was initially not retrieved became accessible after the initial search; or the fact that some DK responses should have been DR responses and reflected an underestimation of the knowledge base. We briefly address each of these in turn.

As Smith and Clark (1993) argued, use of DR and DK might be driven in part by a desire to avoid embarrassment about not being able to answer a question. A DK response in particular might serve to reduce negative affect: Saying "I don't know," as our survey participants indicated, indicates that the information was never learned and thus not able to be retrieved. Thus, admitting to ignorance might be less threatening than admitting to a lapse in memory, especially for a population with pre-existing negative stereotypes about age-related cognitive decline (Hess, Auman, Colcombe, & Rahhal, 2003).

High performance on DK items might also reflect use of sophisticated guessing or strategies used during the final test. For example, on a multiple choice test, a true lack of knowledge would be expected to result in chance performance, unless other strategies were used to increase one's chances of correctly guessing. Examination of response latencies is consistent with the idea that participants were engaging in some strategic processes, such as elimination of options known to be incorrect (e.g., if one knows that Lake Wobegon is a fictional site in a popular radio program, one can exclude that as an option when trying to determine where Thoreau spent his time in the wilderness). In recall tasks, participants can retrieve information about an item, such as number of syllables or initial letter while in a DK state, at rates better than chance (Koriat & Lieblich, 1974), presumably by relying on knowledge of related terms or concepts or the regularities of the language. Similar processes likely contribute to high performance levels observed here.

The third alternative is that participants gained access to correct answers over the duration of the study either because of fluctuations in accessibility or additional search processes. Although the initial search through memory might have failed to elicit a desired answer, some low-level persistent activation of related information might have rendered the correct answer more accessible. Older adults' TOT experiences support this claim in that they have more spontaneous retrievals or "pop ups" of the correct answers compared to younger adults (Burke et al., 1991; Cohen & Faulkner, 1986). they tend to require more time to resolve TOT states (Burke et al., 1991), and they are also more likely to resolve most of them (Heine, Ober, & Shenaut, 1999) than are younger adults. There is also evidence that access to knowledge is not stable but fluctuates over time (e.g., marginal knowledge, Bahrick & Phelps, 1988; Berger, Hall, & Bahrick, 1999).

Finally, above chance accuracy on DK items suggests that participants might have underestimated their knowledge base or mistaken a failure in accessibility for a failure in availability and responded DK when, in reality, the "correct" response was DR. When no familiar information is activated or when one is certain one does not have any relevant knowledge in a field, a fast DK response might be made. However, some DK responses might be made after a more deliberate and slow search, especially if some of the questions were familiar or activated related knowledge in memory. Thus, we examined whether fast or slow DK RTs were equally predictive of subsequent accuracy. If some failures in availability are made rapidly upon determining that one simply does not have the relevant knowledge, final test accuracy on these items should be lower than that for slower DK RTs.

Given the relatively low number of DK responses, especially from older adults, we divided encoding DK RTs into three bins (fast, medium, and slow). We chose three bins instead of a median split or quartiles because the former might not have allowed for sufficient spread between fast and slow DK RTs and the latter yielded too few observations per quartile. Fast, medium, and slow DK RTs were calculated separately for younger and older adults. Final accuracy on the multiple choice test in Experiment 2 was then analyzed as a function of encoding DK RT and age. Thirteen older and 34 younger adults' data were included (few older adults used DK as a response). None of the effects were significant, all $Fs < 2.3$, $ps > .13$. Examination of mean accuracy indicated that, for older adults, accuracy on the final test increased from 0.44 to 0.59 to 0.77 as DK RTs went from fast to slow, consistent with predictions (a one-way ANOVA yielded a p-value of 0.15). Younger adults, however, showed the opposite pattern, with accuracy dropping from 0.62 at the fastest DK RTs to 0.44 for both medium and slow DK RTs (the one-way ANOVA was not significant, $p = .30$).

In Experiment 3, a similar analysis was performed on final cued recall performance. Fourteen older and 35 younger adults' data were included in the analysis. The effect of age was reliable, $F(1,$

47) = 11.15, $MSE = 0.12$, $p = .002$, $\eta_p^2 = 0.19$, with older adults ($M = 0.64$, $SEM = 0.05$) outperforming younger adults ($M = 0.42$, $SEM = 0.03$). Accuracy on the final test varied as a function of encoding DK RT, $F(1.7, 79.86) = 10.48$, $MSE = 0.05$, $p < .001$, $\eta_p^2 = 0.18$. Final recall performance was highest at the slowest DK RT ($M = 0.64$, $SEM = 0.04$), decreasing to 0.53 ($SEM = 0.04$, $p = .02$) at the medium DK RT, and to 0.42 ($SEM = 0.05$, $p = .001$) at the fastest DK RT. Fast and medium DK RTs did not differ, $p = .081$. The interaction was not significant, $p > .90$.

In sum, the analyses based upon DK RT are partially consistent with the hypothesis that DK responses might be based on two distinct processes: A rapid assessment of familiarity followed by a slower search through memory. The fast RTs were associated with lower accuracy in Experiment 3 and, albeit not reliably, for older adults in Experiment 2. Taken in concert with the overall results, these additional analyses suggest that participants can assess not knowing rapidly and relatively accurately, especially when the sought-after information is really not available. However, the limited number of observations available for these analyses, although intriguing, suggests that future studies should develop materials that would yield a higher rate of DR and DK responses to provide more stable RT data. Although we acknowledge that many of our conclusions were based on rather small numbers of trials in the DR and DK cells, in particular for older adults, the effects obtained, as noted, were rather large, thus giving us confidence that the effect would replicate across different materials.

As noted, there was one exception to the finding that initial DR responses were associated with higher rates of subsequent performance than initial DK responses: Older adults in Experiment 3 performed equally well on the final cued recall test for items initially classified as DR and those classified as DK. Examination of response latencies reveals a potential explanation. Younger adults, whose accuracy on the final test differed as a function of initial classification, were equally fast to retrieve the correct answer on DR and DK items. Both RTs were slower than on trials that were initially correct, suggesting they might have been relying on retrieval from episodic stores (i.e., the encoding phase with feedback) for DR and DK trials, whereas initially correct trials reflected the advantage of prior retrieval from semantic stores (i.e., a testing effect; Roediger & Karpicke, 2006). Older adults, however, responded equally fast on initially correct and DR trials during the final test phase and slower on DK trials. Thus, for these participants, successful retrieval attempts and DR trials did not differ in subsequent accessibility, possibly reflecting the fact that DR items were indeed temporarily inaccessible and, once refreshed via feedback, were equally facilitated after a brief delay. The slower RTs on the correct DK trials suggest that participants were engaging in a search through memory, presumably to the encoding phase. The fact that older participants engaged in this search for longer than younger adults can therefore explain their high accuracy rates: Although the information was not known initially (or judged to be not known), the feedback provided participants with a trace to retrieve to respond.

Another possibility is that older adults, during the encoding phase, were better able to retain the less familiar or unknown material in memory. It is likely they had more related or relevant knowledge previously stored (Umanath & Marsh, 2014), providing a platform for integrating new information. A DK response might have served to orient participants to the fact that the information was not part of their knowledge base, thereby encouraging them to attend to it to learn it. Older adults might also have been more intrinsically interested in the GK questions or curious about the answers when they failed to retrieve them. Interest can positively predict subsequent memory performance (McGillivray et al., 2015). Similarly, a DR response would have increased attention because it would be akin to resolving a TOT state for participants. In a recent study, Bloom, Friedman, Xu, Vuorre, and Metcalfe (2018) examined the differences between DK responses and TOT states. Not surprisingly, final test accuracy was higher for the latter. More relevant to the present discussion was the finding that differences in final test performance were mediated in part by processing of feedback, as evidenced by event-related potentials. Because older adults tend to have more TOTs, a larger knowledge base, and might be more motivated to perform well than younger adults (e.g., Hess, 2005; Hess, Germain, Swaim, & Osowski, 2009), their motivations to attend to and subsequently remember the feedback might have been greater. In sum, when retrieval failed, it is possible that older adults attended more deeply to the answer, thereby supporting subsequent retrieval.

One might argue that, because of above-chance performance on DK items, the measure used here falls short in accurately assessing what participants do and do not know or remember. As discussed, a number of factors might explain the fact that participants generally underestimated what they knew. Furthermore, metacognitive judgments based on phenomenology, like we used here, are not expected to be perfect at predicting retrieval success because they are the subjective judgments of participants based on their experience at the moment. Experience-based metacognitive approaches (Moulin & Souchay, 2014) suggest that feelings and experiences serve a specific function in memory to guide search and processing strategies. These feelings are most likely to be noticed when they are in some way discordant with the goals of the individual – as in the case of a retrieval failure. Importantly, the phenomenology is associated with the metacognitive process, not the cognitive one.

In fact, most measures of metacognition in the literature fail to perfectly predict performance. In the context of the R/K paradigm, which resembles the measure used here, *remember* responses are sometimes attributed to false alarms, indicating that participants attribute a recollective experience to a non-studied event. In fact, some researchers have suggested that such responses be omitted from analyses (Rotello, Macmillan, Reeder, & Wong, 2005); however, the fact remains that participants do make such attributions. As an additional example, despite initially providing a non-TOT designation, participants were able to recognize almost 60% of correct answers to general knowledge questions (Bahrick et al., 2010). Similarly, JOLs "tend to overestimate learning and demonstrate modest relative accuracy" (Shanks & Serra, 2014, pg. 445). FOKs, as measured by Gamma correlations, typically do not reveal a perfect association and, in some cases, participants are unable to make accurate FOK predictions (i.e., their FOKs and actual recognition are uncorrelated; e.g., Thomas, Bulevich, & Dubois, 2011, Experiment 1; see also Busey, Tunnicliff, Loftus, & Loftus, 2000, for a discussion of confidence/accuracy dissociations). Lachman et al. (1979) also reported that older adults tended to underestimate their knowledge, correctly recognizing over 40% of items classified as not known in an initial recall test.

In our first experiment, participants at all levels of expertise clearly distinguished between DR and DK in terms of phenomenological experiences and in the experimental studies they further discriminated based on these terms. Clearly, no measure of cognitive or metacognitive performance is going to be perfect. However, we note that since many measures of metacognition reveal over-confidence (such as JOLs), a measure that seems to demonstrate under-confidence might be useful in allowing participants to adopt more conservative biases in their own assessment. Consistent with our findings, Luo et al. (2003) found that among items given a FOnK judgment, over 50% were recognized accurately on the final test, indicating that participants tended to underestimate their performance. As argued by Son and Kornell (2010), knowing what one does not know can enhance learning and promote use of more effective learning strategies. Furthermore, underestimating knowledge, in an applied context such as studying for an exam, is likely to have less aversive consequences than overestimating it.

To address the issues raised above, the DR response could be partitioned into two distinct response categories (i.e., I used to know it but forgot vs. It's on the tip of my tongue/I can't remember it right now but I know I know it). Singer and Thiede (2008) examined the use of an additional response – "I once knew it" (OKI) and found that such

responses were associated with longer response times (RTs) prior to responding DK and with higher recognition accuracy than responses that did not reflect prior knowledge (although it is important to note that different participant groups provided OKI and DK ratings). An OKI response could be akin to saying DR if not remembering reflects acknowledging that information was at one point available and is no longer accessible (or no longer available due to decay).

As evidenced by the results of Experiment 1, participants might use DK to communicate that they have never been exposed to the information being queried. Not remembering, conversely, refers both to having once known but since forgotten the information, or to temporary inaccessibility such as a TOT. Such a partitioning of DR might increase the sensitivity of the measure and capture some of the nuances that were identified in the survey data (e.g., the distinction between forgetting or temporary inaccessibility). What the research reported here represents is an initial exploration of how the phenomenology of retrieval failures can be expressed beyond TOTs by the expressions not knowing and not remembering. This dichotomy, as noted in the Introduction, mirrors the R/K distinction used in hundreds of studies since Tulving (1985) proposal that conscious experiences could discriminate between sources of retrieval.

Limitations and future directions

In addition to the limited number of trials in which participants used the DR/DK options, we note that our conclusions regarding the feasibility of using a DR/DK measure to discriminate between availability and accessibility failures are, at this point, limited to retrieval failures from the knowledge base or semantic memory. Whether similar effects would emerge in traditional episodic memory tasks remains to be determined.

A second concern is whether participants would have spontaneously used DR and DK or whether the findings reported here reflect demand characteristics. We acknowledge this might be the case, but also note that virtually all metacognitive measures currently in use involve a similar conundrum. For example, in a TOT study, participants are often asked what they can recall about the un-retrievable target. Would they spontaneously provide such responses? Similarly, it seems unlikely that participants spontaneously assign numerical probabilities to the likelihood they will recognize an item from a list as is done in FOK studies. We do not wish to diminish the value of such research, of course; we simply note that some demand characteristics are inevitable. There is some evidence that the demand characteristics here are fairly minimal and unlikely to distort the results excessively. First, the results of Experiment 1 clearly indicate that participants do conceptualize these terms differently and can provide elaborate explanations without any context or instruction. Second, the fact that participants in our experiments showed the same patterns, albeit magnified, suggests that the use of these terms is consistent both when a query provides no indication that the use is within the context of memory and when participants have just completed a memory test. Third, in the instructions, participants were given the minimal possible cuing, in that they were simply told they could respond, say *I don't know*, or say *I don't remember* with no additional elaboration or instruction. Fourth, our results converge nicely with Smith and Clark (1993) study, in which they recorded spoken protocols and examined spontaneous use of DR and DK, suggesting that such responses are part of participants' vernacular.

Conclusions

As noted in the introduction, the terms *remember* and *know* were used by Tulving (1985) to reflect retrieval from episodic and semantic stores, respectively. Based on our findings, *not* remembering and *not* knowing appear to reflect differences in accessibility and availability. The measure developed and tested here provides some initial insight into participants' understanding of and use of the expressions "I don't remember" and "I don't know." Across age groups and levels of knowledge, participants appeared to use DR more so for information that was stored but unable to be retrieved, whereas DK was used more for information that was deemed unlikely to have ever been stored. Accuracy and response latencies provided converging evidence for the face validity of this phenomenologically-driven judgment of retrieval failure. That is, participants were more overall likely to correctly recognize information following a DR response than a DK response. By capitalizing on natural language use, this measure allows participants to readily discriminate between failures in retrieval and failures in storage. By building on and extending the strong tradition of the R/K and FOK literatures, this research emphasizes that memory researchers and lay participants alike make an important distinction between retrieval failures reflecting what is not known or not remembered – i.e., what is not available and what is not accessible.

Declaration of Competing Interest

None.

Acknowledgements

The authors would like to thank Cher Almeida, Shuofeng Xu, Yi Feng, Cole Walsh, Jake Lester, and Sarah Boland for their contributions to data collection and data coding. We also thank Maciej Hanczakowski and two anonymous reviewers for feedback on an earlier version of the manuscript.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jml.2019.05.002.

References

Bahrick, H. P., Baker, M. K., Hall, L. K., & Abrams, L. (2010). How should we define and differentiate metacognitions? In A. S. Benjamin (Ed.). *Successful remembering and successful forgetting: Essays in honor of Robert A. Bjork* (pp. 329–346). New York, NY: Psychology Press. https://doi.org/10.4324/9780203842539.

Bahrick, H. P., & Phelps, E. (1988). The maintenance of marginal knowledge. In U. Neisser, & E. Winograd (Eds.). *Remembering reconsidered: Ecological and traditional approaches to the study of memory* (pp. 178–192). New York: Cambridge University.

Balota, D. A., & Coane, J. H. (2008). Semantic memory. In J. H. Byrne, H. Eichenbaum, R. Menzel, H. L. Roediger, & D. Sweatt (Eds.). *Handbook of learning and memory: A comprehensive reference* (pp. 511–534). Amsterdam: Elsevier.

Balota, D. A., Dolan, P. O., & Duchek, J. M. (2000). Memory changes in healthy young and older adults. In E. Tulving, & F. I. M. Craik (Eds.). *The Oxford handbook of memory* (pp. 395–410). Oxford: Oxford University Press.

Barber, S. J. (2017). An examination of age-based stereotype threat about cognitive decline: Implications for stereotype-threat research and theory development. *Perspectives on Psychological Science, 12*(1), 62–90. https://doi.org/10.1177/1745691616656345.

Berger, S. A., Hall, L. K., & Bahrick, H. P. (1999). Stabilizing access to marginal and submarginal knowledge. *Journal of Experimental Psychology: Applied, 5*, 438–447.

Bloom, P. A., Friedman, D., Xu, J., Vuorre, M., & Metcalfe, J. (2018). Tip-of-the-tongue states predict enhanced feedback processing and subsequent memory. *Consciousness and Cognition, 63*, 206–217. https://doi.org/10.1016/j.concog.2018.05.010.

Bowles, N. L., & Poon, L. W. (1985). Aging and retrieval of words in semantic memory. *Journals of Gerontology, 40*(1), 71–77. https://doi.org/10.1093/geronj/40.1.71.

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109*(2), 204–223. https://doi.org/10.1037/0033-2909.109.2.204.

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior, 5*(4), 325–337. https://doi.org/10.1016/S0022-5371(66)80040-3.

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30*(5), 542–579. https://doi.org/10.1016/0749-596X(91)90026-G.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review, 7*(1), 26–48. https://doi.org/10.3758/BF03210724.

Cantor, A. D., Eslick, A. N., Marsh, E. J., Bjork, R. A., & Bjork, E. L. (2014). Multiple-choice tests stabilize access to marginal knowledge. *Memory and Cognition, 43*(2), 193–205. https://doi.org/10.3758/s13421-014-0462-6.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1–22. https://doi.org/10.1037/h0046743.

Cavanaugh, J. C., Grady, J. G., & Perlmutter, M. (1983). Forgetting and use of memory aids in 20 to 70 year olds everyday life. *International Journal of Aging & Human Development, 17*(2), 113–122. https://doi.org/10.2190/H7L2-K3XK-H32K-VW89.

Cleary, A. M. (2018). The biasing nature of the tip-of-the-tongue experience: When decisions bask in the glow of the tip-of-the-tongue state. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0000520.

Cleary, A. M., & Claxton, A. B. (2015). The tip-of-the-tongue heuristic: How tip-of-the-tongue states confer perceptibility on inaccessible words. *Journal of Experimental Psychology: Learning Memory and Cognition, 41*(5), 1533–1539. https://doi.org/10.1037/xlm0000097.

Cohen, G., & Faulkner, D. (1986). Does 'elderspeak' work? the effect of intonation and stress on comprehension and recall of spoken discourse in old age. *Language and Communication, 6*(1–2), 91–98. https://doi.org/10.1016/0271-5309(86)90008-X.

Connor, L. T., Balota, D. A., & Neely, J. H. (1992). On the relation between feeling of knowing and lexical decision: Persistent subthreshold activation or topic familiarity? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 544–554. https://doi.org/10.1037/0278-7393.18.3.544.

Craik, F. I. M., & Byrd, M. (1982). Aging and cognitive deficits: The role of attentional resources. In F. I. M. Craik, & E. Trehub (Eds.). *Aging and cognitive processes* (pp. 191–211). New York: Plenum Press.

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

Eich, T. S., Murayama, K., Castel, A. D., & Knowlton, B. J. (2014). The dynamic effects of age- related stereotype threat on explicit and implicit memory performance in older adults. *Social Cognition, 32*(6), 559–570. https://doi.org/10.1521/soco.2014.32.6.559.

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*, 175–191.

Faust, M. E., Ferraro, F. R., Balota, D. A., & Spieler, D. H. (1999). Individual differences in information-processing rate and amount: Implications for group differences in response latency. *Psychological Bulletin, 125*(6), 777–799. https://doi.org/10.1037/0033-2909.125.6.777.

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

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory, 7*(5), 311–325. https://doi.org/10.1037/0278-7393.7.5.311.

Hanczakowski, M., Pasek, T., Zawadzka, K., & Mazzoni, G. (2013). Cue familiarity and 'don't know' responding in episodic memory tasks. *Journal of Memory and Language, 69*(3), 368–383. https://doi.org/10.1016/j.jml.2013.04.005.

Hanczakowski, M., Zawadzka, K., & Cockcroft-McKay, C. (2014). Feeling of knowing and restudy choices. *Psychonomic Bulletin and Review, 21*(6), 1617–1622. https://doi.org/10.3758/s13423-014-0619-0.

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56*(4), 208–216. https://doi.org/10.1037/h0022263.

Heine, M. K., Ober, B. A., & Shenaut, G. K. (1999). Naturally occurring and experimentally induced tip-of-the-tongue experiences in three adult age groups. *Psychology and Aging, 14*(3), 445–457. https://doi.org/10.1037/0882-7974.14.3.445.

Hertzog, C., & Dunlosky, J. (2011). Metacognition in later adulthood: Spared monitoring can benefit older adults' self-regulation. *Current Directions in Psychological Science, 20*(3), 167–173. https://doi.org/10.1177/0963721411409026.

Hertzog, C., Dunlosky, J., & Sinclair, S. M. (2010). Episodic feeling-of-knowing resolution derives from the quality of original encoding. *Memory & Cognition, 38*(6), 77–84. https://doi.org/10.3758/MC.38.6.771.

Hertzog, C., Sinclair, S. M., & Dunlosky, J. (2010). Age differences in the monitoring of learning: Cross-sectional evidence of spared resolution across the adult life span. *Developmental Psychology, 46*(4), 939–948. https://doi.org/10.1037/a0019812.

Hess, T. M. (2005). Memory and aging in context. *Psychological Bulletin, 131*(3), 383–406. https://doi.org/10.1037/0033-2909.131.3.383.

Hess, T. M., Auman, C., Colcombe, S. J., & Rahhal, T. A. (2003). The impact of stereotype threat on age differences in memory performance. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences, 58*(1), 3–11. https://doi.org/10.1093/geronb/58.1.P3.

Hess, T. M., Germain, C. M., Swaim, E. L., & Osowski, N. L. (2009). Aging and selective engagement: The moderating impact of motivation on older adults' resource utilization. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences, 64*(4), 447–456. https://doi.org/10.1093/geronb/gbp020.

Jersakova, R., Allan, R., Booth, J., Souchay, C., & O'Connor, A. R. (2017). Understanding metacognitive confidence: Insights from judgment-of-learning justifications. *Journal of Memory and Language, 97*, 187–207. https://doi.org/10.1016/j.jml.2017.08.002.

Klin, C. M., Guzmán, A. E., & Levine, W. H. (1997). Knowing that you don't know: Metamemory and discourse processing. *Journal of Experimental Psychology: Learning Memory and Cognition, 23*(6), 1378–1393. https://doi.org/10.1037/0278-7393.23.6.1378.

Knoll, M. A. Z., & Arkes, H. R. (2017). The effects of expertise on the hindsight bias. *Journal of Behavioral Decision Making, 30*(2), 389–399. https://doi.org/10.1002/bdm.1950.

Koriat, A. (1993). How do we know that we know? The accessibility model of the feeling of knowing. *Psychological Review, 100*(4), 609–639. https://doi.org/10.1037/0033-295X.100.4.609.

Koriat, A. (1995). Dissociating knowing and the feeling of knowing: Further evidence for the accessibility model. *Journal of Experimental Psychology: General, 124*(3), 311–333. https://doi.org/10.1037/0096-3445.124.3.311.

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 34–53. https://doi.org/10.1037/0278-7393.27.1.34.

Koriat, A., & Lieblich, I. (1974). What does a person in a "TOT" state know that a person in a "don't know" state doesn't know. *Memory & Cognition, 2*(4), 647–655. https://doi.org/10.3758/BF03198134.

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(1), 283–294. https://doi.org/10.1037/a0037850.

Lachman, J. L., Lachman, R., & Thronesbery, C. (1979). Metamemory through the adult life span. *Developmental Psychology, 15*(5), 543–551. https://doi.org/10.1037/0012-1649.15.5.543.

Light, L. L., & Carter-Sobell, L. (1970). Effects of changed semantic context on recognition memory. *Journal of Verbal Learning and Verbal Behavior, 9*(1), 1–11. https://doi.org/10.1016/S0022-5371(70)80002-0.

Liu, Y., Su, Y., Xu, G., & Chan, R. C. K. (2007). Two dissociable aspects of feeling-of-knowing: Knowing that you know and knowing that you do not know. *Quarterly Journal of Experimental Psychology, 60*(5), 672–680. https://doi.org/10.1080/17470210601184039.

Luo, J., Kazuhisa, N., & Luo, Y. (2003). Neural correlates of "feeling-of-not-knowing": Evidence from functional MRI. *Chinese Science Bulletin, 48*(2), 144–147. https://doi.org/10.1360/03tb9029.

Marquié, J. C., & Huet, N. (2000). Age differences in feeling-of-knowing and confidence judgments as a function of knowledge domain. *Psychology and Aging, 15*(3), 451–461. https://doi.org/10.1037/0882-7974.15.3.451.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6.

McGillivray, S., Murayama, K., & Castel, A. D. (2015). Thirst for knowledge: The effects of curiosity and interest on memory in younger and older adults. *Psychology and Aging, 30*(4), 835–841. https://doi.org/10.1037/a0039801.

McIntyre, J. S., & Craik, F. I. M. (1987). Age differences in memory for item and source information. *Canadian Journal of Psychology, 41*(2), 175–192. https://doi.org/10.1037/h0084154.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review, 15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174.

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

Morris, R. G., Craik, F. I., & Gick, M. L. (1990). Age differences in working memory tasks: The role of secondary memory and the central executive system. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 42*, 67–86. https://doi.org/10.1080/14640749008401208.

Morson, S. M., Moulin, C. J. A., & Souchay, C. (2015). Selective deficits in episodic feeling of knowing in ageing: A novel use of the general knowledge task. *Acta Psychologica, 157*, 85–92. https://doi.org/10.1016/j.actpsy.2015.02.014.

Moulin, C., & Souchay, C. (2014). Epistemic feelings and memory. In T. J. Perfect, & D. S. Lindsay (Eds.). *The Sage handbook of applied memory* (pp. 520–540). Los Angeles, CA: Sage. https://doi.org/10.4135/9781446294703.n29.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In G. H. Bower (Vol. Ed.), *The psychology of learning and motivation: Vol. 26*, (pp. 125–141). San Diego: Academic Press.

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In 1. Metcalfe, & A. P. Shimamura (Eds.). *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 510–520. https://doi.org/10.1037/0278-7393.14.3.510.

Old, S. R., & Naveh-Benjamin, M. (2008). Differential effects of age on item and associative measures of memory: A meta-analysis. *Psychology and Aging, 23*(1), 104–118. https://doi.org/10.1037/0882-7974.23.1.104.

Park, D. C. (2000). The basic mechanism accounting for age-related decline in cognitive function. In D. C. Park, & N. Schwarz (Eds.). *Cognitive aging: A primer* (pp. 3–21). Philadelphia, PA: Psychology Press.

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.

Reder, L. M. (1987). Strategy selection in question answering. *Cognitive Psychology, 19*(1), 90–138. https://doi.org/10.1016/0010-0285(87)90005-3.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 435–451. https://doi.org/10.1037/0278-7393.18.3.435.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181–210. https://doi.org/10.1111/j.1745-6916.2006.00012.x.

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. https://doi.org/10.3758/BF03196778.

Rubin, D. C., & Umanath, S. (2015). Event memory: A theory of memory for laboratory, autobiographical, and fictional events. *Psychological Review, 122*(1), 1–23. https://doi.org/10.1037/a0037907.

Salthouse, T. A. (2004). What and when of cognitive aging. *Current Directions in Psychological Science, 13*(4), 140–144. https://doi.org/10.1111/j.0963-7214.2004.

00293.x.
Schacter, D. L. (2001). *The seven sins of memory: How the mind forgets and remembers.* Boston: Houghton Mifflin.
Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime user's guide.* Pittsburgh, PA: Psychology Software Tools Inc.
Schwartz, B. L. (1999). Sparkling at the end of the tongue: The etiology of tip-of-the-tongue phenomenology. *Psychonomic Bulletin & Review, 6*(3), 379–393. https://doi.org/10.3758/BF03210827.
Schwartz, B. L. (2006). Tip-of-the-tongue states as metacognition. *Metacognition and Learning, 1*(2), 149–158. https://doi.org/10.1007/s11409-006-9583-z.
Schwartz, B. L., & Metcalfe, J. (1992). Cue familiarity but not target retrievability enhances feeling-of-knowing judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(5), 1074–1083. https://doi.org/10.1037/0278-7393.18.5.1074.
Schwartz, B. L., & Metcalfe, J. (2011). Tip-of-the-tongue (TOT) states: Retrieval, behavior, and experience. *Memory & Cognition, 39*(5), 737–749. https://doi.org/10.3758/s13421-010-0066-8.
Schwartz, B. L., & Metcalfe, J. (2014). Tip-of-the-tongue (TOT) states: Mechanisms and metacognitive control. In B. L. Schwartz, & A. S. Brown (Eds.). *Tip-of-the-Tongue states and related phenomena* (pp. 15–31). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139547383.002.
Shanks, L. L., & Serra, M. J. (2014). Domain familiarity as a cue for judgments of learning. *Psychonomic Bulletin & Review, 21*(2), https://doi.org/10.3758/s13423-013-0513-1.
Singer, M., & Thiede, H. L. (2008). Feeling of knowing and duration of unsuccessful memory search. *Memory & Cognition, 36*, 588–597. https://doi.org/10.3758/MC.36.3.588.
Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language, 32*(1), 25–38. https://doi.org/10.1006/jmla.1993.1002.
Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta- analysis. *Psychonomic Bulletin and Review, 8*(2), 203–220. https://doi.org/10.3758/BF03196157.
Son, L. K., & Kornell, N. (2010). The virtues of ignorance. *Behavioral Processes, 83*, 207–212. https://doi.org/10.1016/j.beproc.2009.12.005.
Squire, L. R. (1992). Declarative and nondeclarative memory: Multiple brain systems supporting learning and memory. *Journal of Cognitive Neuroscience, 4*(3), 232–243. https://doi.org/10.1162/jocn.1992.4.3.232.
Staudinger, U. M., Cornelius, S. W., & Baltes, P. B. (1989). The aging of intelligence: Potential and limits. *The Annals of the American Academy of Political and Social Science, 503*(1), 43–59. https://doi.org/10.1177/0002716289503001004.
Tauber, S. K., Dunlosky, J., Rawson, K. A., Rhodes, M. G., & Sitzman, D. M. (2013). General knowledge norms: Updated and expanded from the Nelson and Narens (1980) norms. *Behavior Research Methods, 45*(4), 1115–1143. https://doi.org/10.3758/s13428-012-0307-9.
Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2011). Context affects feeling-of-knowing accuracy in younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(1), 96–108. https://doi.org/10.1037/a0021612.
Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2012). An analysis of the determinants of the feeling of knowing. *Consciousness and Cognition, 21*(4), 1681–1694. https://doi.org/10.1016/j.concog.2012.09.005.
Thomas, A. K., Lee, M., & Hughes, G. (2015). Feeling-of-Knowing judgments (FOKs). In J. Dunlosky, & S. Tauber (Eds.). *The Oxford handbook of metamemory* (pp. 81–95). New York, NY: Oxford University Press.
Thomson, D. M., & Tulving, E. (1970). Associative encoding and retrieval: Weak and strong cues. *Journal of Experimental Psychology, 86*(2), 255–262. https://doi.org/10.1037/h0029997.
Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, & W. Donaldson (Eds.). *Organization of memory* (pp. 381–402). Oxford, England: Academic Press.
Tulving, E. (1983). *Elements of episodic memory.* Oxford, England: Clarendon Press.
Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne, 26*(1), 1–12. https://doi.org/10.1037/h0080017.
Tulving, E. (1989). Memory: Performance, knowledge, and experience. *European Journal of Cognitive Psychology, 1*(1), 3–26. https://doi.org/10.1080/09541448908403069.
Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5*(4), 381–391. https://doi.org/10.1016/S0022-5371(66)80048-8.
Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352–373. https://doi.org/10.1037/h0020071.
Umanath, S. (2016). Age differences in suggestibility to contradictions of demonstrated knowledge: The influence of prior knowledge. *Aging, Neuropsychology, and Cognition, 23*(6), 744–767. https://doi.org/10.1080/13825585.2016.1167161.
Umanath, S., & Coane, J. H. (submitted for publication). Face validity of remembering and knowing: Empirical consensus and disagreement between participants and researchers.
Umanath, S., & Marsh, E. J. (2014). Understanding how prior knowledge influences memory in older adults. *Perspectives on Psychological Science, 9*(4), 408–426. https://doi.org/10.1177/1745691614535933.
Wais, P. E. (2008). fMRI signals associated with memory strength in the medial temporal lobes: A meta-analysis. *Neuropsychologia, 46*, 3185–3196. https://doi.org/10.1016/j.neuropsychologia.2008.08.025.
Wixted, J. T., & Squire, L. R. (2011). The medial temporal lobe and the attributes of memory. *Trends in Cognitive Sciences, 15*, 210–217. https://doi.org/10.1016/j.tics.2011.03.005.