Journal of Memory and Language 95 (2017) 124–137

Contents lists available at ScienceDirect

Journal of Memory and Language

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

Predicting others' memory performance: The accuracy and bases of social metacognition

Jonathan G. Tullis [a,*], Scott H. Fraundorf [b]

[a] Department of Educational Psychology, University of Arizona, United States
[b] Learning Research and Development Center and Department of Psychology, University of Pittsburgh, United States

ARTICLE INFO

Article history:
Received 9 May 2016
revision received 23 January 2017
Available online 22 March 2017

Keywords:
Metacognition
Cue generation
Perspective-taking

ABSTRACT

Successful teaching, effective advertising, and happy interpersonal relationships depend upon accurately anticipating what others will remember. Across three experiments, we tested how precisely subjects judged the mnemonic effectiveness of cues for supporting other subjects' episodic memories. Some subjects generated cue-target word pairs and made judgments of learning (JOLs) for these word pairs while other subjects studied the pairs and made JOLs. Across all three experiments, subjects' JOLs for others were more accurate than chance, but less accurate than subjects' JOLs for themselves. Further, JOLs for others were similarly accurate across cues that subjects generated for others and cues that subjects read but did not themselves generate. Idiosyncratic cue generation processes impacted subjects' JOLs for others; however, this bias was not the primary reason for the inaccuracy of JOLs for others. Rather, our results suggest that the accuracy of judgments about others' memories suffers because people do not have access to the personal idiosyncrasies of others' encoding and processing.

Introduction

Predicting what other people will remember has significant social, educational, and economic consequences. Successful teachers, effective negotiators, electable politicians, and productive advertisers rely upon the ability to put themselves in others' shoes in order to generate cues that others will strongly remember (Nickerson, 1999, 2001). Similarly, creating and maintaining friendships, communicating effectively, and making plans rely upon our abilities to anticipate the likelihood of other people retaining task-relevant information and to generate cues that will support their recall (Selman, 1981).

Despite these processes being "indispensible to human social functioning" (p. 85; Nelson, Kruglanski, & Jost, 1998), little research has examined metacognition related to others' memories. In the current experiments, we examined how *accurately* subjects judged the effectiveness of mnemonic cues for others, and we analyzed the *bases* upon which subjects made those judgments of effectiveness. These topics lie at the intersection of three important areas in cognitive psychology: cue generation, metacognitive monitoring, and perspective-taking. We will outline questions of interest in each area in turn before describing the current experiments in greater detail.

Generating mnemonic cues for self and others

When taking notes, naming computer files, and outlining readings, learners use sophisticated strategies to generate external mnemonic cues in order to reduce the demands placed on the memory system and to support future retrieval. Learner-generated cues can lead to very high levels of memory performance, even for long lists of items (Bäckman & Mäntylä, 1988; Hunt & Smith, 1996; Mäntylä, 1986; Mäntylä & Nilsson, 1983, 1988). One reason for the effectiveness of learner-generated cues is that they can be more distinctive (i.e., point to fewer possible targets) and more idiosyncratic as compared to, for instance, experimenter-provided cues (Tullis & Benjamin, 2015b). For instance, when given the target word "sibling", a strongly-associated generic cue for the target might be "sister" while an idiosyncratic, self-generated mnemonic cue might be the name of one's own sister (e.g., "Gillian").

Learners adjust the types of cues they generate when they know those cues will be used by other people. For instance, when generating mnemonic cues for others as opposed to themselves, subjects generate cues that have greater normative associative strength to the target word, but point to more possible targets (Tullis &

* Corresponding author at: Department of Educational Psychology, University of Arizona, 1430 E 2nd Street, Tucson, AZ 85721, United States.
E-mail address: jonathantullis@gmail.com (J.G. Tullis).

http://dx.doi.org/10.1016/j.jml.2017.03.003
0749-596X/

Benjamin, 2015a). Further, when generating cues for themselves, subjects create unique and idiosyncratic cues (as described above), but when generating cues for others, they generate more shared (or "common") cues (Kraus, Vivekananthan, & Weinheimer, 1968). These adjustments effectively support others' recall: Cues that are generated with the intent of being used by others lead to more accurate recall for those individuals than cues generated without such intent (Tullis & Benjamin, 2015a).

Predicting memory performance of others

However, perspective-taking in cue generation does not end with generating the cues. Another important task is *judging* the effectiveness of those cues in supporting others' memory performance, such as determining which of several candidate ad slogans is the most memorable. Although little is known about this topic, some plausible hypotheses are suggested by research on how learners make predictions about their *own* recall. When subjects can rely upon personal mnemonic experiences to make predictions, their judgments of learning (JOLs) are very accurate at predicting which items they will remember and which they will forget (Dunlosky & Nelson, 1994). However, people usually do not have direct access to the personal mnemonic experiences of others, so their predictions regarding others' memories may be much less accurate than predictions of one's own memories. This exemplifies the cognitive challenge of *perspective-taking*—making judgments that reflect someone else's knowledge and experience rather than one's own.

Perspective-taking

Predicting what others will remember necessitates stepping outside of one's own experiences and taking the perspective of others. Although research generally suggests that taking another's perspective is difficult and subject to systematic errors (Hanna, Tanenhaus, & Trueswell, 2003; Keysar, Barr, Balin, & Brauner, 2000), perspective-taking is a domain-specific skill that may be easier or harder in some tasks than others (Ryskin, Benjamin, Tullis, & Brown-Schmidt, 2015). Thus, there is a need to consider how—and how effectively—people take perspective specifically within the domain of creating and judging the effectiveness of mnemonic cues. In the present study, we contrast four hypotheses—inspired by the broader literature on why perspective-taking is frequently difficult—about how accurately people can judge the effectiveness of mnemonic cues for other learners.

Under what we term the *misleading-information* hypothesis, perspective-taking is difficult because people struggle to ignore their own knowledge when attempting to take the perspective of others (i.e., the "curse of knowledge"; Birch, 2005). Across many paradigms and stimuli, people's own knowledge and experiences influence their messages for others even when they are irrelevant to others' perspectives (Birch, 2005; Brown-Schmidt & Hanna, 2011; Camerer, Lowenstein, & Weber, 1989; Kelley & Jacoby, 1996; Stone, Baron-Cohen, & Knight, 1998). According to this hypothesis, judgments about others' knowledge are contaminated by irrelevant (or systematically misleading) aspects of one's own experience, and consequently people should show some ability to predict others' memory, but with reduced accuracy. For instance, Tullis and Benjamin (2015a) demonstrated that cues for the self benefit from different features than cues from others, so if subjects erroneously predict the memories of others using the same mnemonic indicators as they might use to predict their own memories, their predictions may be biased. Crucially, in this view, it is the *presence* of misleading information (i.e., one's own experience) that contributes to reduced accuracy in predicting others' memory performance.

A second source of difficulty in perspective-taking may be a *lack* of adequate knowledge about a particular person's perspective. Under the *inadequate-information* hypothesis, predicting of others' memory may show reduced accuracy because people simply have less information about other learners than about themselves. In particular, subjects lack access to others' idiosyncratic mnemonic experiences, which may serve an essential, diagnostic role in predicting episodic memory performance (Lovelace, 1984; Underwood, 1966). That is, this hypothesis also predicts that judging others' memories should be more difficult, but this difficulty reflects an *absence* of information rather than the presence of misleading information. Some evidence in favor of the inadequate-information hypothesis comes from Vesonder and Voss (1985), who divided subjects into (a) *learners* who studied a list of sentences and predicted whether they would remember the sentences on a later memory test and (b) *observers* who predicted whether the learners would remember each studied sentence. Observers' judgments of which sentences the learners would and would not remember were near chance, but increased sharply when observers were provided information about the learners' idiosyncratic performance (i.e., the learners' prior success or failure at retrieving specific sentences). This pattern suggests that the source of perspective-taking failure in this task was a lack of information, rather than an inability to use information about the learners' subjective encoding experiences.

The literature also suggests two other, more extreme hypotheses about perspective-taking. One, which we term the *equivalent-accuracy* hypothesis, is that learners should be equally accurate at predicting another person's future memory as predicting their own future memory. This hypothesis is suggested by the more general proposal (Nelson et al., 1998) that taking the perspective of one's future self may have much in common with taking the perspective of another person: Both require judging a mental state different from one's immediate mental state (e.g., Fraundorf & Benjamin, 2014). Predicting others' memories and predicting one's future memory performance may rely on the same cues and processes (Jost, Kruglanski, & Nelson, 1998), and consequently result in similar outcomes. Finally, under what we term the *no-accuracy* hypothesis, people may completely fail to predict what other learners can and cannot remember. This hypothesis is suggested by the finding of Vesonder and Voss (1985) that, without specific information about other learners' mnemonic experiences, observers' predictions were no more accurate than chance.

The relative contributions of the misleading influence of one's own perspective and inadequate information about another's perspective to perspective-taking failures are unclear. Moreover, because perspective-taking is a domain-specific ability (Ryskin et al., 2015), testing perspective-taking abilities across many situations is crucial in order to identify specific perspective-taking strengths and weaknesses. Here, we do so in the domain of creating and judging mnemonic cues.

Overview of experiments

Across three experiments, subjects generated mnemonic cues and judged their effectiveness for others. We directly evaluated subjects' metacognitive monitoring by testing how *accurately* they predicted their own and others' memories. Second, we examined the *bases* on which people make judgments about the efficacy of mnemonic cues for others: Do their predictions of memory appropriately reflect the same cue characteristics that actually correlate with actual memory?

Cue characteristics

To characterize the efficacy of cues and the bases upon which subjects make metacognitive judgments, we analyzed three char-

acteristics of subject-generated cues (cue-to-target associative strength, number of associates, and cue commonality) that are related to a cue's effectiveness (Tullis & Benjamin, 2015a, 2015b), as well as one characteristic of subjects' experience generating the cue (cue generation time).

We operationalized these cue characteristics as in the prior literature (Ryskin et al., 2015). *Cue-to-target associative strength* describes how strongly a cue points to a target and was determined using the normative cue-to-target associative strength found in the University of South Florida Free Association Norms (Nelson, McEvoy, & Schreiber, 1998). *Number of associates* was operationalized by counting the number of targets associated to each cue in the Free Association Norms (Nelson et al., 1998) and provides a measure of cue distinctiveness. A cue with few associates other than the target (i.e., a more distinctive cue) is advantageous because it can limit the group of possible targets, and therefore the search space, during recall. Finally, *cue commonality* was the proportion of subjects that generated a specific cue for a given target. Cue commonality plays a similar role as cue-to-target associative strength, but may be additionally important because it is measured within the experimental task presented here and thus suggests how shared the knowledge is for the specific mnemonic cue task used here (as opposed to a general free association task).

Each of these cue characteristics can vary independently. For example, in the first experiment, subjects generated a cue word for the target word "car" (among others). One subject generated "honk" and several generated "drive." "Drive" has a stronger cue-to-target associative strength to "car" than "honk" does, and since several subjects generated "drive," this cue also has greater cue commonality. But, the cue "drive" is associated with a greater number of other targets in the database (e.g., "fast", "stop", "steer") than "honk".

In addition to the three stimulus characteristics described above, we analyzed one characteristic of the cue generation process: the time needed to generate the cue. Cue generation time may be an idiosyncratic mnemonic experience specific to the subject who generates the cue. The learner who studies the cue-target pair has no knowledge of how long it took the generator to create it, and so cue generation time should have no direct influence on learners' JOLs or recall. Thus, consideration of cue-generation time in a generator's JOLs would represent an intrusion of one's own idiosyncratic experiences into judgments of another's perspective.

Experiment 1

In Experiment 1, we examined how accurately subjects predicted others' memory performance and whether subjects rely upon the same information to make judgments for the self versus for others. Experiment 1 included two between-subjects conditions. In one, *generators* created cue words for others and rated the memorability of these cues for others. In the other, *learners* viewed the intact cue-target word pairs, rated their memorability for themselves, and were tested on them. This allowed us to directly compare the accuracy and bases of predictions for the self and for others.

Method

Participants

One hundred and fifty-four introductory psychology students at Indiana University participated for partial course credit.

Materials

Eighty words were collected from the University of South Florida Free Association Norms (Nelson et al., 1998). We chose items for which intuition suggested that college-aged subjects would have some personal experiences, such as *hobby* and *roommate*. Appendix A presents the complete list of targets.

Procedure

The experiment was presented in MATLAB using the Psychophysics Toolbox (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007) and CogToolbox (Fraundorf et al., 2014) on personal computers in individual testing rooms. The assignment of subjects to conditions alternated between two conditions ($n = 77$ each). Specifically, every other subject in each testing room was assigned to the generation condition (*generators*) and was instructed to generate cues that would best support the memory for a new, different learner. Generators were told that another subject in the experiment would study the cue-target word pair and then have to recall the target when given the cue. Generators viewed the to-be-remembered targets one at a time in a random order in black 25 point Arial font on the right side of the computer screen. To the left of each target item, an empty cue box was presented. For each word, generators typed a single word into the cue box and pressed the Enter key. If generators entered the target word as its own cue, the computer program erased the cue and asked them to enter a different cue. The length of time needed to generate the cue was recorded. Immediately after generating a cue, the subject rated the effectiveness of their cue for a new learner on a 1–4 scale. Generators were instructed that, "1 means the cue will NOT help another learner remember the target, 2 means the cue probably will NOT help the learner remember the target, 3 means the cue will probably help the learner remember the target, and 4 means that the cue will definitely help the learner remember the target." They were further instructed, "Generate the best cues for a different learner as you can. But also try to use the whole rating scale."

Learners were yoked to the immediate prior subjects on the same computer, who were always in the generation condition. Learners studied the list of 80 targets with the cues that the yoked prior subject generated. Learners were told that they would study a list of word pairs, rate the memorability of the pairs, and then take a cued recall test for the targets. Learners studied each word pair for 500 ms, at which point the metacognitive rating scale appeared at the bottom of the screen. Each word pair stayed on the screen until learners entered their JOL. Learners received the following instructions: "You will rate the cue-target pairs on a scale of 1 to 4. 1 means that the cue will NOT support your memory for the target, 2 means the cue probably will NOT help you remember the target, 3 means the cue will probably help you remember the target, and 4 means that the cue will definitely help you remember the target. Try to use the whole scale." After rating each word pair, learners immediately took the cued recall test. The cues were presented in a new random order, and learners had to type in the corresponding target for each cue.

Results

Analytic strategy

We calculated the accuracy of subjects' predictions using the gamma correlations between JOLs and recall, as has been done in prior research (Benjamin & Diaz, 2008; Tullis & Benjamin, 2012).

To assess the relationship between cue characteristics and recall and JOLs, we used linear mixed-effects models (Baayen, Davidson, & Bates, 2008; Murayama, Sakaki, Yan, & Smith, 2014). The unit of analysis in a mixed-effect model is the outcome of a single trial; i.e., whether or not a particular target was recalled by a particular participant, or the rating of a particular cue's effectiveness. These trial-level outcomes can be modeled as a function of multiple *fixed effects*—those of theoretical interest—as well as multiple *random effects*—effects for which the observed levels are

sampled out of a larger population (e.g., subjects sampled out of a population of possible subjects).

Linear mixed-effects models solve four statistical problems related to the investigation of self-generated cues. First, unlike experiments in which cues and targets are selected and counterbalanced by the experimenter, participants generated their own cues. Participant generation of cues was important to our experimental design because it introduced large and ecologically valid variability in the generated cues, ensured that generators had personal experiences with the cues, and, most importantly, provided us with a measure of the generators' idiosyncratic experience (cue generation time). However, this methodology also required an analysis that takes account of this item-level variability (Freeman, Heathcote, Chalmers, & Hockley, 2010). Mixed-effect models solve this problem by simultaneously accounting for random variation both across participants and across items (Baayen, 2008; Murayama et al., 2014). Second, because the cues were participant-generated, they varied continuously across the full range of important variables, such as cue-to-target associative strength, rather than existing only in categorically distinct experimental conditions. Mixed-effects models can take full advantage of the information contained in this continuous variation. By contrast, dichotomizing such variables for an ANOVA model (e.g., through a median split) greatly reduces statistical power (Cohen, 1983). Third, we were interested in how several different characteristics of the participant-generated cues, such as cue-to-target associative strength and the number of associates, each influenced recall and participants' JOLs. Mixed effects models are an ideal solution because, like all multiple regression models, they can include multiple predictors at the same time, testing the effect of one while controlling for the other. Finally, mixed effects models can appropriately deal with binary outcomes—such as whether or not a particular item was recalled—by modeling the log odds (or *logit*) of recall (Jaeger, 2008).

For each mixed-effect model, we included multiple fixed effects of theoretical interest. We also included *random intercepts* that capture baseline differences in memorability and in JOLs across participants, target words, and cue words. Using likelihood-ratio tests, we assessed the contribution of each random slope to each model, and we report the model with the maximal random-effects structure justified by the data. Except where otherwise noted, all predictor variables were mean-centered to produce main effects equivalent to those obtained from an ANOVA. For models with continuous outcomes (i.e., JOLs), we assessed the significance of fixed-effects by comparing the t statistic to the normal distribution because in mixed-effect models with hundreds of observations, the t-distribution effectively converges to the normal (Baayen, 2008); for models with categorical outcomes (i.e., recall), p-values can be directly obtained from the Wald z statistics; All models were fit in R using the *lmer()* function of the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015).

Cued recall performance

We assessed cued recall performance using strict scoring, such that only responses that were identical matches to the target were counted as correct, because it provided the most objective measure of recall. Nevertheless, the results were unchanged if we scored changes in plurality or misspelled words as correct (which affected 3% of the data). Overall, learners correctly recalled 57% (SD = 9%) of the targets.

Accuracy of learners' predictions

The distribution of JOLs across conditions in each experiment is shown in Table 1. We evaluated the accuracy of mnemonic predictions by computing the gamma correlations between predictions and recall. The gamma correlation between predictions of others' recall and their actual recall (M = 0.25, SD = 0.21) was significantly above zero, $t(76) = 10.17$, $p < 0.001$, indicating that predictions about others' memories were more accurate than chance. Nevertheless, gamma correlations for others' recall were significantly smaller than gamma correlations for one's own recall (M = 0.33, SD = 0.23); $t(76) = 2.93$, $p = 0.005$, Cohen's $d = 0.34$, indicating that predictions about others' recall were less accurate than predictions about one's own recall.

Characteristics of effective cues

Why were predictions of others' memories less diagnostic of others' actual recall than predictions of one's own memory? One explanation is that judgments about others' memories were based on different, less diagnostic cue characteristics. To test this hypothesis, we examined (a) first, which cue characteristics were *actually* associated with a higher probability of recall (i.e., cue diagnosticity: see Van Loon, de Bruin, van Gog, van Merrienboer, & Dunlosky, 2014) and (b) next, which cue characteristics were associated with *predictions* of memory for oneself and others (i.e., cue utilization: see Van Loon et al., 2014).

To address the first question, we examined recall performance in the Learn condition as a function of characteristics of the cues previously provided by the generators. (Recall that only the learners' recall performance was assessed.) Four such characteristics were tested: Three characterized the cue itself (cue-to-target associative strength, number of associates, and cue commonality), and one characterized the generation process (cue generation time). Note that these characteristics are measured on different, unrelated scales. For example, cue generation time is measured in milliseconds whereas cue commonality is measured in a proportion of participants. To make their effects more comparable, we standardized (z-scored) each cue characteristic so that the regression coefficients corresponded to the effects of a 1-standard-deviation change in any of the cue characteristics.

Table 2 displays the mean of each of these cue characteristics (in both the original units and on the standardized scale) separately for items that the learner eventually remembered and for items that the learner eventually did not remember. However, any differences here must be interpreted with some caution. Because the variables characterize participant-generated cues that were not orthogonally manipulated, it is possible that an apparent effect of one cue characteristic (e.g., cue commonality) could arise from that variable being partially confounded with another cue characteristic (e.g., number of associates). Instead, the cue characteristics are better analyzed in a multiple-regression context that allows us to examine the effect of one cue characteristic while holding all other variables constant.

Table 3 lists the results of such a mixed-effects regression predicting the log odds of recall success from the cue characteristics. Cue-to-target associative strength was the strongest predictor of recall; the odds of recall increased by 1.97 times (95% confidence interval: [1.58, 2.46]) for every 1-standard-deviation increase in associative strength. Number of associates and cue commonality also had significant effects, and these two effects were of roughly equal size. A 1-standard-deviation increase in the number of associates (thus producing a *less* distinctive cue) was associated with a 1.38 times decrease in the odds of correct recall (95% CI: [1.28, 1.50]). A 1-standard-deviation increase in the percentage of generators providing that cue (a more common cue) was associated with a 1.31 times increase in the odds of recall (95% CI: [1.15, 1.49]). By contrast, cue generation time was entirely unrelated to the learners' recall. Its effect did not approach significance, $p = 0.64$, and the point estimate of the regression coefficient indicated an effect that was an order of magnitude smaller than the others.

Table 1
The proportion of cue-target word-pairs assigned to each JOL value and the mean JOL value in each condition across Experiments 1–3.

	JOL 1	JOL 2	JOL 3	JOL 4	Mean JOL
Experiment 1					
Generator for Learner	0.11 (0.09)	0.27 (0.10)	0.39 (0.14)	0.24 (0.13)	2.75 (0.29)
Learner for Self	0.15 (0.10)	0.28 (0.08)	0.32 (0.12)	0.24 (0.13)	2.65 (0.31)

	JOL 1	JOL 2	JOL 3	JOL 4	Mean JOL
Experiment 2					
Generator for Learner	0.05 (0.04)	0.22 (0.12)	0.42 (0.13)	0.31 (0.18)	2.99 (0.31)
Observer for Learner	0.08 (0.08)	0.19 (0.09)	0.35 (0.14)	0.38 (0.18)	3.03 (0.35)
Learner for Self	0.10 (0.10)	0.26 (0.13)	0.34 (0.15)	0.30 (0.20)	2.83 (0.37)

	JOL 1	JOL 2	JOL 3	JOL 4	JOL 5	JOL 6	JOL 7	Mean JOL
Experiment 3								
Generator for Self	0.08 (0.05)	0.10 (0.05)	0.12 (0.05)	0.13 (0.06)	0.16 (0.08)	0.16 (0.08)	0.25 (0.16)	4.69 (0.54)
Generator for Learner	0.15 (0.09)	0.12 (0.05)	0.11 (0.06)	0.13 (0.06)	0.14 (0.06)	0.17 (0.07)	0.17 (0.12)	4.19 (0.60)
Learner for Self	0.11 (0.07)	0.11 (0.08)	0.10 (0.07)	0.11 (0.06)	0.14 (0.05)	0.16 (0.08)	0.28 (0.22)	4.63 (0.92)
Learner for Generator	0.08 (0.07)	0.10 (0.07)	0.09 (0.06)	0.11 (0.07)	0.15 (0.07)	0.19 (0.10)	0.28 (0.23)	4.85 (0.89)

Table 2
Means and standard deviations in Experiment 1 of cue characteristics for cue-target pairs that were and that were not remembered, measured in original units (top half) or standardized scores (bottom half).

	Items Remembered		Items Not Remembered	
	M	*SD*	*M*	*SD*
Cue characteristic				
Cue-to-target associative strength	0.09	0.16	0.03	0.09
Number of associates	9.93	7.24	11.17	7.32
Cue commonality	0.19	0.21	0.13	0.18
Cue generation time (s)	8.10	7.44	8.34	7.67
Cue characteristic (z-scored)				
Cue-to-target associative strength	0.16	1.16	−0.23	0.63
Number of associates	−0.07	0.99	0.10	1.00
Cue commonality	0.11	1.05	−0.15	0.90
Cue generation time (s)	−0.01	0.99	0.02	1.02

Table 3
Fixed effect estimates for mixed effects logit model of cued recall accuracy in Experiment 1 as a function of cue characteristics (N = 6064, log-likelihood: −3565).

Fixed effect	$\hat{\beta}$	*SE*	Wald *z*	*p*
Intercept (baseline recall)	0.662	0.118	5.63	<0.001
Cue-to-target associative strength	0.682	0.112	6.07	<0.001
Number of associates	−0.325	0.005	−8.15	<0.001
Cue commonality	0.274	0.065	4.24	<0.001
Cue generation time	0.017	0.036	0.47	0.64

Note. SE = standard error.

Bases for judgments

To summarize the above model, some cue characteristics (associative strength, number of associates, and cue commonality) predicted recall whereas another (cue generation time) did not. Thus, one possible explanation of inaccurate predictions by cue generators is that they underweighted the cue characteristics that actually predicted recall and/or that they were influenced by an irrelevant characteristic (generation time).

To test this hypothesis, we examined whether these same cue characteristics were predictive of the JOLs made by the cue generators and by the learners. We conducted a second mixed-effects regression with the JOL as the dependent measure and the cue characteristics as predictor variables. Further, we included subject condition (Generate or Learn) to assess whether subjects in one condition generally rated the cues as more effective than subjects in the other condition (a main effect of Condition). Finally, and most crucially, we allowed Condition to interact with each of the cue characteristics in predicting JOLs to assess whether the two conditions differ in how subjects weighted the cue characteristics in making their JOLs (a Condition × Cue Characteristic interaction). We dummy-coded Condition with the Generate condition as the baseline level; under this system, the simple main effect terms represents the effects of the cue characteristics within the Generate condition whereas the interaction terms indicate whether the effects of the cue characteristics reliably differed between learners and cue generators.

Table 4 lists the results of this regression. First, consider the effects within the Generate condition. Participants who generated cues assigned higher JOLs to cues with higher cue-to-target associative strength, to cues with fewer associates, and to more common cues, consistent with the observed effects of all of these characteristics on actual recall. In addition, however, cue generators gave lower JOLs to cues that took them longer to generate even though cue generation time had no relation to the learners' recall.

Next, we turn to the interaction terms. The Condition × Cue-to-target associative strength interaction was non-significant; there was no evidence that learners considered cue-to-target associative strength differently than did generators. By contrast, for the other three cue characteristics, there were significant positively-signed interactions with Condition. In a regression model, positively-signed interactions amplify positively-signed main effects; that is, learners were even more sensitive to cue commonality than

Table 4
Fixed effect estimates for mixed effects model of JOLs in Experiment 1 (N = 12,128, log-likelihood: −15,208).

Fixed effect	$\hat{\beta}$	*SE*	*t*	*p*
Intercept (baseline rating)	2.755	0.042	64.97	<0.001
Cue-to-target associative strength (Generate condition)	0.144	0.024	5.88	<0.001
Number of associates (Generate condition)	−0.102	0.014	−7.27	<0.001
Cue commonality (Generate condition)	0.247	0.028	8.97	<0.001
Cue generation time (Generate condition)	−0.221	0.018	−12.19	<0.001
Learn condition	−0.083	0.054	−1.53	0.13
Learn condition × Cue-to-target associative strength	<0.001	0.020	0.01	0.99
Learn condition × Number of associates	0.087	0.019	4.71	<0.001
Learn condition × Cue commonality	0.077	0.026	2.94	<0.01
Learn condition × Cue generation time	0.167	0.020	8.39	<0.001

Note. SE = standard error.

were generators. However, positively-signed interactions *reduce* negatively-signed main effects; in other words, learners relied less than the generators on number of associates and were less sensitive to cue generation time in making JOLs. Indeed, the similarly sized parameter estimates for the simple main effects (−0.102 for number of associates and −0.221 for cue generation time) as for the corresponding interaction terms (0.087 and 0.167, respectively) indicate that the influence of these variables was essentially eliminated in the Learn condition. That is, learners appeared to disregard the number of associates and cue generation time when assigning JOLs for their own memories—but, by contrast, they were *more* sensitive to cue commonality. Finally, a marginal main effect of Condition indicated that learners tended to give somewhat lower overall JOLs than generators, but this effect did not reach conventional levels of significance.

Discussion

Subjects' predictions about others' memories were more accurate than chance, but were less accurate than subjects' predictions about their own memories. This result is evidence against strong accounts of perspective-taking in which metacognitive judgments about others' memories either are wholly inaccurate (*no-accuracy* hypothesis) or are as accurate as predictions of one's own memory (*equivalent-accuracy* hypothesis). People show *some* ability to predict other learners' memories, even without access to others' personal mnemonic experiences, but their predictions of others are less accurate than predictions about their own memory.

One possible reason for reduced accuracy when predicting others' memories, as proposed by the misleading-information hypothesis, is that JOLs made for other learners relied upon different, less informative cue characteristics than JOLs made for oneself. Experiment 1 provided some support for this view. Cue generators' JOLs were influenced by cue generation time even though generation time was unrelated to others' recall; therefore, using it to assign JOLs adds noise to judgments and may reduce the accuracy of JOLs for others. The process of generating a mnemonic cue may provide distracting, non-diagnostic information about the cue that subjects cannot ignore.

Further, when generators predicted others' recall, they relied more upon the number of associates and less upon cue commonality than did subjects predicting their own recall. Prior research has suggested that the number of associates is more strongly related to a generator's recall than to a learner's recall whereas cue commonality is more strongly related to a learner's recall than to a generator's recall (Tullis & Benjamin, 2015a). In strongly weighting the number of associates and weakly weighting cue commonality, generators may be predicting the performance of others by focusing on cue characteristics related to their own recall rather than those related to others' recall. However, it is not clear whether these differences were necessarily harmful. The regression model in Table 3 indicated that, for the present task, number of associates (to which generators were more sensitive than learners) and cue commonality (to which generators were less sensitive than learners) had roughly equal effects on recall. Thus, it is not clear that differentially relying on these cues would actually result in a net decrease in relative metamnemonic accuracy.

These two possible causes of the inaccuracies in predicting others' memories—the inability to one's own experiences and the lack of information about another learner—are difficult to distinguish on the basis of data from Experiment 1 alone. In this experiment, generators differed from learners both in having to predict someone else's recall *and* in rating cues that they generated. One way to tease apart these influences would be to examine the accuracy of JOLs made by individuals who are predicting someone else's memory but who did not generate the memory cues themselves. Subjects who do not generate the cues cannot be misled by the idiosyncratic cue generation process. Thus, if the primary difficulty in predicting others' memories is the inability to ignore one's own idiosyncratic cue generation experience (as proposed by the misleading information hypothesis), subjects who do not generate the cues themselves should be better able to predict the effectiveness of cues for others than subjects who generated those cues. We tested this prediction in Experiment 2.

Experiment 2

In this second experiment, we examined whether generating a cue—and the potentially misleading idiosyncratic experiences that come with it—underlies the difficulty of judging cue-target memorability for others. We replaced the generation condition of Experiment 1 with a *generate-or-observe* condition in which subjects also judged the effectiveness of already-generated cues for others. Specifically, each generate-or-observe subject was yoked to a particular generate-condition subject from Experiment 1. For half of the items, these subjects *generated* their own cue and then rated it (as in prior experiments). For the other half of the items, these subjects *observed* and rated the existing cue-target pair created by their yoked partner from Experiment 1.

If the process of generating cues provides misleading idiosyncratic information about the memorability of cues (i.e., cue generation time) and consequently impairs predictions of others' recall, then JOLs made when generating cues will be less accurate than JOLs made when merely observing the cues. By contrast, if the difficulty of mnemonic perspective-taking simply owes to the lack of information about the other learner, as proposed by the inadequate-information hypothesis, then observer JOLs and generator JOLs should be equally non-diagnostic relative to learners' predictions of their own memory because both observer JOLs and generator JOLs require taking the perspective of another learner.

Method

Participants

One hundred and fifty-six introductory psychology students at Indiana University participated for partial course credit.

Materials

The same eighty items used in the first experiment were used in this experiment. Further, half of the cues generated by subjects in Experiment 1 were used within this experiment.

Procedure

Subjects were alternatively assigned to the generate-or-observe condition and the learn condition. In the generate-or-observe condition, subjects generated a new cue for some targets and observed existing cues for other targets. Specifically, each generate-or-observe participant from Experiment 2 was yoked to a different subject from Experiment 1 (see Tullis & Benjamin, 2011). For a random half of the target items, subjects saw only the target item and they generated cue words to help other subjects recall the target items (as the generators did for all the targets in Experiment 1). For the other random half of the targets, subjects read the cue-target pairs created by the yoked subject from Experiment 1.[1] Immediately after reading or generating each cue word, subjects judged the memorability of the cues for a new learner, as in Experiment 1. The generated and observed words were ordered randomly throughout the list.

Subjects in the learn condition were asked to remember and rate the memorability of intact cue-target word pairs, just as learners did in Experiment 1. Word pairs were presented on the screen for 500 ms before the rating scale was also displayed. Learners rated the memorability of the cue-target word pair for themselves on a scale of 1–4. Half of the word pairs had cues generated by the immediate prior subject (i.e., one assigned to the generate-or-observe condition from this experiment), and half of the word pairs had cues generated by the matched subject from Experiment 2 (the memorability of which had been judged by both the immediate prior subject in this experiment and the generator in Experiment 1). In this manner, the consecutive subjects in the generate-or-observe condition and the learn condition rated memorability of the exact same 80 cue-target pairs in the same presentation order. After learners had studied and rated all 80 word pairs, they immediately began a cued recall test. The cue word was presented on the left side of the screen and subjects typed in the corresponding target item in a box on the right side of the screen, as in Experiment 1.

Results

As in Experiment 1, cued recall performance was assessed using strict scoring, although the conclusions were unchanged if we counted misspellings or changes in plurality as correct (affecting 4% of the data). The subjects in the learn condition recalled 56% (*SD* = 14%) of the word pairs on average.

[1] One potential hypothesis related to this procedure is that seeing someone else's cues may affect the quality and kind of cues that a generator would produce and how they assign JOLs. We compared the effectiveness and characteristics of cues generated by generate-or-observe subjects from this experiment with the generate subjects in Experiment 1. We found no significant differences between experiments either in terms of the cues' ability to support recall (i.e., learners' eventual recall did not significantly differ between generate-or-observe cues and generate-only cues) or the cue characteristics. Further, we found no differences in how subjects assigned JOLs to cues across experiments. It, therefore, seems unlikely that observing someone else's cues changed the types of cues generate-or-observe subjects produced or how they judged the effectiveness of cues.

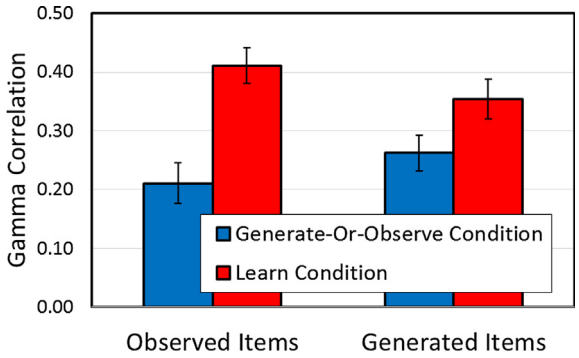

Fig. 1. Gamma correlations between predictions and learners' recall in Experiment 2, as a function of subjects' conditions and whether subjects in the generate-or-observe condition observed or generated the cues. Subjects in the learn condition never generate cues, but each cue they studied was either generated or observed by their yoked subject in the generate-or-observe condition. Error bars show one standard error of the mean above and below the sample mean.

Accuracy of learners' predictors

When judging the efficacy of cues for other subjects, do subjects make more accurate predictions for cues they merely observed than for cues they generated? Half of the cue-target pairs were observed by the subjects in the generate-or-observe condition, and half were generated. We computed the gamma correlations for these two subsets of items across all subjects in Experiment 2, and the gammas are displayed in Fig. 1.[2] Gamma correlations between predictions about others' recall and others' actual recall were significantly greater than zero for both observed ($t(77) = 6.16$, $p < 0.001$) and generated ($t(77) = 8.56$, $p < 0.001$) items, indicating that predictions about others' memories were more accurate than chance. However, the generate-or-observe subjects were less effective at predicting the learners' memories than the learners were at predicting their own memories (effectively replicating Experiment 1); this was true both for cues that the generate-or-observe subjects generated themselves ($t(77) = 2.23$, $p = 0.03$, Cohen's $d = 0.25$), and, critically, for cues that these subjects only observed ($t(77) = 4.61$, $p < 0.001$, Cohen's $d = 0.52$). That is, regardless of whether a subject generated or observed a cue, their predictions of learners' memories were less accurate than learners' self-predictions. Indeed, within the generator-or-observe condition, no significant difference was found between the predictive accuracy for observed and generated cues ($t(77) = 1.17$, $p = 0.25$, Cohen's $d = 0.13$); in fact, metamnemonic accuracy was numerically higher for generated cues than observed cues. There was no evidence that generating a particular cue made it harder to judge its effectiveness for others; predicting others' memories was simply difficult in general.

Characteristics of effective cues

As in Experiment 1, we first determined which cue characteristics were actually associated with a higher probability of recall, and we then examined whether those cue characteristics were leveraged differently across the conditions. We sought to (a) replicate the finding in Experiment 1 that predictions of one's own recall were made based on somewhat different, more diagnostic cue characteristics than predictions of others' recall and (b) determine

[2] The primary hypotheses of this experiment concerned the simple comparisons between the learn and generate-or-observe conditions, and these analyses are presented in the main text. For the sake of completeness, however, we conducted a 2 (subject condition: generate-or-observe or learn) × 2 (item condition: generated or observed) repeated measures ANOVA on the gamma correlations, and this analysis revealed a significant interaction between conditions ($F(1,77) = 4.64$, $p = 0.03$, $\eta_p^2 = 0.06$) and a main effect of subject condition ($F(1,77) = 18.19$, $p < 0.001$, $\eta_p^2 = 0.19$). No significant main effect of item condition obtained ($F(1,77) = 0.01$, $p = 0.93$, $\eta_p^2 < 0.001$).

Table 5
Means and standard deviations in Experiment 2 of cue characteristics for cue-target pairs that were not remembered and for pairs that were remembered, measured in original units (top half) or standardized scores (bottom half).

	Items Remembered		Items Not Remembered	
	M	*SD*	*M*	*SD*
Cue characteristic				
Cue-to-target associative strength	0.09	0.16	0.04	0.09
Number of associates	10.14	7.23	11.12	7.26
Cue commonality	0.19	0.21	0.14	0.18
Cue generation time (s)	8.67	8.55	9.53	9.73
Cue characteristic (z-scored)				
Cue-to-target associative strength	0.17	1.17	−0.21	0.69
Number of associates	−0.06	1.00	0.07	1.00
Cue commonality	0.10	1.05	−0.12	0.91
Cue generation time (s)	−0.04	0.94	0.05	1.07

whether predictions of others' recall were made on different bases if one had generated the relevant cues oneself.

Consider that half of the test items in Experiment 2 (the ones for which the cues had been generated in Experiment 1) were rated by three participants: the subject taking the eventual cued recall test (the *learner*), the subject who generated the cue (the *generator*), and the *observer* subject in Experiment 2, who neither generated that particular cue nor was tested on it. We selected these items for analysis because these items received JOLs under all three conditions. First, we examined whether the characteristics of the cues received by the learners predicted which cue-target pairings would be recalled. (Although we had no reason to expect this pattern to be any different in Experiment 2, we wanted to confirm that this was the case, and, additionally, to conduct a replication of the Experiment 1 results.)

Table 5 lists the means and standard deviations of each cue characteristic for the pairs that were remembered and for the pairs that were not remembered. Again, however, the most informative analysis is a multiple mixed-effects regression model that tests the effect of each cue characteristic while holding the others constant. The results of that model, listed in Table 6, closely replicated those of Experiment 1 both in the patterns of significance and in the relative size of the effects. Cue-to-target associative strength was once again the strongest predictor of recall; the odds of recall increased by 2.03 times (95% confidence interval: [1.63, 2.52]) for every 1-standard-deviation increase in associative strength. The number of associates also impacted recall: For every 1-standard-deviation increase in the number of words associated with the cue (i.e., a *less* distinctive cue), the odds of recall decreased 1.25 times (95% CI: [1.17, 1.36]). Cue commonality further predicted recall; for every 1-standard-deviation increase in the percentage of participants giving the cue, the odds of correct recall increased 1.10 times (95% CI: [1.02, 1.20]). Finally, the time taken to generate the cue was not significantly related to the learners' recall, $p = 0.82$, and the estimated effect was an order of magnitude smaller than the others. Thus, the pattern across all four cue characteristics was the same as in Experiment 1.

Judgments of cue effectiveness

As in Experiment 1, some cue characteristics (associative strength, number of associates, and cue commonality) predicted recall whereas one other (cue generation time) did not. The pattern of metacognitive accuracy across conditions seen in Fig. 1 might thus be explained in part by differences across conditions in which of these cue characteristics influenced predictions of recall. We modeled cue effectiveness rating as a function of cue-to-target associative strength, number of associates, cue commonality, and cue generation time, as well as the interaction of these cue characteristics with the condition under which the rating was made.

Table 7 lists the results of this regression. Overall, cues were given higher ratings when they had higher cue-to-target associative strength, had fewer associates, were more common across the subject population, and were generated more quickly. Our primary interest, however, was differences across conditions in how subjects used these variables to make JOLs. First, consider the contrast between learners making JOLs for themselves versus the two conditions in which participants made JOLs for someone else. The Learn × Cue Generation Time interaction was in the opposite direction as the main effect, indicating that—as in Experiment 1—cue generation time was less influential on JOLs made by learners. Indeed, the magnitude of the condition × cue generation time was virtually the same as that of the main effect,[3] indicating that the effect of cue generation time on JOLs was essentially eliminated for the learners (as expected, since the learners had no knowledge of the cue generation time). Making predictions about someone else's recall (rather than one's own) did not affect how subjects used the other cue characteristics when making JOLs. All of the other Learn × cue characteristic interactions were non-significant, though there was a marginal decrease in the effect of number of associates. Thus, it appears that there were not large differences in how cue characteristics were utilized between judgments of one's own memory versus judgments of someone else's memory.

Rather, if any such differences exist, they must be driven specifically by subjects rating cues they had themselves generated. We next test for this possibility by considering the difference between the two conditions in which participants predicted someone else's recall. Overall, generated cues elicited lower JOLs than observed cues. In addition, reliable Generate × Number of Associates and Generate × Cue Generation Time interactions in the same direction as the main effects indicate that these variables were more influential when participants rated their own generated cues rather than someone else's. By contrast, there was a significant Generate × Cue Commonality interaction in the opposite direction as the main effect; cue generators were less influenced by the commonality of the cue for the subject population as a whole. That is, observers judging existing cues more heavily weighted cue commonality whereas cue generators were less sensitive to this cue characteristic.

[3] It might be asked why the magnitude of the interaction term was not *exactly* equal to the main effect of cue generation time given that, in the Learn conditions, participants had no knowledge of how long the cue took the generator to create. The most likely explanation is that this simply reflects sampling error; even under the null hypothesis that cue generation time is irrelevant in the Learn condition, it is improbable that the observed effect of cue generation in that condition would be exactly equal to zero.

Table 6
Fixed effect estimates for mixed effects logit model of cued recall accuracy in Experiment 2 as a function of cue characteristics (N = 6160, log-likelihood: −3747).

Fixed effect	$\hat{\beta}$	*SE*	Wald *z*	*p*
Intercept (baseline recall)	0.384	0.104	3.67	<0.001
Cue-to-target associative strength	0.708	0.111	6.35	<0.001
Number of associates	−0.231	0.038	−6.09	<0.001
Cue commonality	0.101	0.043	2.34	<0.05
Cue generation time (s)	−0.007	0.032	−0.23	0.82

Note. SE = standard error.

Table 7
Fixed effect estimates for mixed effects model of JOLs in Experiment 2 (N = 9240, log-likelihood: −11,667).

Fixed effect	$\hat{\beta}$	*SE*	*t*	*p*
Intercept (baseline rating)	2.872	0.029	98.44	<0.001
Cue-to-target associative strength	0.094	0.019	4.76	<0.001
Number of associates	−0.047	0.014	−3.42	<0.01
Cue commonality	0.291	0.029	9.99	<0.001
Cue generation time	−0.088	0.019	−4.73	<0.001
Learn condition (vs rating for others)	−0.076	0.054	−1.41	0.16
Learn condition × cue-to-target associative strength	−0.005	0.022	−0.21	0.83
Learn condition × number of associates	0.035	0.019	1.88	0.06
Learn condition × cue commonality	−0.016	0.022	−0.73	0.43
Learn condition × cue generation time	0.065	0.023	2.82	<0.01
Generate condition (vs Observe condition)	−0.291	0.053	−5.52	<0.001
Generate condition × cue-to-target associative strength	−0.005	0.022	−0.73	0.47
Generate condition × number of associates	−0.065	0.021	−3.13	<0.01
Generate condition × cue commonality	−0.067	0.023	−2.89	<0.01
Generate condition × cue generation time	−0.174	0.036	−4.81	<0.001

Note. SE = standard error.

Discussion

Experiment 2 replicated the key results of Experiment 1. Subjects accurately predicted the recall of others, but not as accurately as they predicted their own recall. Again, people showed *some* ability to predict other's memories, but this ability was limited. Further, Experiment 2 provided a more stringent test of whether the idiosyncratic information about the cue generation process drove these deficits in prediction accuracy by comparing judgments of cues that participants had themselves generated versus cues that were merely observed. Critically, the process of generating a cue did not affect the overall accuracy of subjects' predictions relative to simply observing the cues; in fact, subjects judged the effectiveness of cues they generated numerically more accurately than observed cues. This suggests that the generators' personal cue generation experiences did not cloud or impair their predictions of others' memories. Rather, predicting others' cued recall appears to be intrinsically more difficult than predicting one's own, regardless of whether one has generated the relevant cues or merely observed them, consistent with the inadequate-information hypothesis.

The conditions under which JOLs for others were made did, however, shape how subjects weighed different stimulus characteristics. When subjects generated cues for others, their JOLs relied more upon the number of associates and cue generation time, but less upon cue commonality, compared to when subjects observed already-generated cues. However, these differences between generators and observers in *how* they made JOLs for other people neither improved nor impaired the metamnemonic resolution of those JOLs. The aforementioned differences likely canceled each other out. Number of associates and cue commonality were *both* related to recall, so the fact that generators relied more on number of associates and observers more on cue commonality would not lead one condition to be more accurate than the other. Rather, these can be viewed as two separate strategies for arriving at the same level of metamnemonic accuracy.

Thus, Experiment 2 provides evidence against a misleading-information hypothesis in which the difficulty of predicting others' memories stems from the misleading effects of cue characteristics more relevant to one's own cognitive processing; the significant differences across conditions in which cues were used to make memory judgments did not match the differences across conditions in relative metamnemonic accuracy. Rather, the results seem to favor an inadequate-information view of perspective-taking in which the difficulty of predicting someone else's memory stems from a lack of information about others' memory.

The inadequate-information hypothesis makes another, stronger prediction: If predicting others' memory is generally difficult (rather than reflecting misleading effects of the cue generation process), *learners* should be disadvantaged in predicting the memory of cue *generators*, even though they never generated any cues. Experiment 3 tests this prediction.

Experiment 3

In the first two experiments, generators only ever predicted someone else's memory, and learners only ever predicted their own memory. In Experiment 3, we expanded the design to include the full set of conditions suggested by our experimental variables: cue generators predicting their own memory, cue generators predicting someone else's memory, learners predicting their own memory, and learners predicting someone else's memory. Specifically, generators produced mnemonic cues, rated the cues' effectiveness for both themselves and others, and took a memory test on all cues. Learners, similarly, rated the mnemonic cues for both themselves and generators before being tested on all word-pairs. This design allowed us to examine whether the relationship between JOLs and actual recall differed as a function of two factors that were now completely orthogonal: (a) whether the prediction was being made by the cue generator or the learner and (b) whether the prediction concerned one's own memory or someone

else's memory. The inadequate-information hypothesis predicts that judgments of someone else's memory should always be worse than judgments of one's own memory (i.e., a main effect of factor B above). By contrast, the misleading-information hypothesis predicts that cue generators may be unable to avoid being influenced by their idiosyncratic experience generating the cues even where that experience is irrelevant to ultimate recall, thus making them especially impaired at predicting the learners' recall (and perhaps even their own).

This design also removed any confound with test expectancy: In the first two experiments, learners always expected (and received) a memory test, whereas generators never received or expected a test. Thus, because cue generators did not expect a memory test, their predictions might have overvalued the generation process at the expense of the appropriate considerations of the final memory test. Experiment 3 removed this confound; all groups of participants expected and received a memory test.

Similarly, in the first two experiments, generators were always produced cues for others, and subsequently judged the effectiveness of those cues for others. Outputting a cue for someone else likely involves an inherent judgment process whereby generators first assess the quality of the cue for others before outputting them. Generators likely withhold cues that would be unsuitable for others; withholding cues ultimately restricts the types of cues used, may limit or bias the metacognitive judgments produced, and artificially depresses the metamnemonic accuracy of judgements for others. In Experiment 3, generators output cues for themselves, and this should reduce or eliminate the bias in metacognitive judgments produced for others.

Method

Participants

Sixty members of the University of Arizona community participated in this experiment over the course of 2 days for $20.

Materials

The same 80 words used in the first two experiments were used here.

Procedure

In this experiment, generators were told to produce cues that would help them remember the target words for a later memory test, whereas prior experiments asked generators to produce cues helpful for other subjects. After generating cues for all of the target words, the generators judged the effectiveness of the cues for themselves and for other learners. Generators were alternatively assigned to the self-first or other-first judgment conditions as a means of counter-balancing.[4] In the self-first condition, generators viewed each cue-target pair one-at-a-time and rated the effectiveness of the cue for their own memory before repeating the entire process for someone else's memory. In this experiment, all subjects rated each cue twice: once for themselves and once for others. A concern with the JOL scale used in the previous two experiments was that it would not allow enough flexibility for subjects to provide different JOLs for self and for others. Therefore, subjects rated cues on a large 1 (least memorable) to 7 (most memorable) scale. The larger JOL scale utilized in this experiment gave a larger range for subjects to select and allowed more movement between ratings for self and for other. In the other-first judgment condition, generators rated the effectiveness of all of the cues for a different learner before they rated the cues for themselves. Generators left the lab after rating each cue twice. They returned two days later to take the final memory test where they were presented with each cue they generated and had to recall the corresponding target item. The retention interval for generators was 2 days to produce cued recall performance that was not at ceiling and similar to that of learners.[5]

Each *learner* was yoked to a single earlier generator and studied the cue-target pairs that the generator produced. The learners, like the generators, were alternatively assigned to first rate the generator's memory or their own memory. Learners in the self-first condition viewed each cue-target pair and rated the effectiveness of the cue for their own memory on a scale of 1–7. Next, they viewed the list of cue-target pairs in a new random order and rated the effectiveness of the cues for the generator. When rating the effectiveness of the cues for the generator, the learners were instructed: "Each cue you study was created by a previous subject. The prior subject generated the cue and took a memory test. During the test, they got their cue back and had to type in the corresponding target. You will judge how effectively each cue supported the memory of the prior subject." After rating each word pair twice (once for themselves and once for the generator), learners completed an unrelated math task for 30 min. Finally, learners took the cued recall test. The cues were presented in a new random order, and the learners had to type in the corresponding target for each cue.

Results

Accuracy of recall

Cued recall performance was assessed using strict scoring, as in the prior two experiments. Generators recalled numerically more targets ($M = 65\%$, $SD = 20\%$) than learners ($M = 58\%$, $SD = 19\%$), but this difference did not reach significance according to a paired t-test ($t(29) = 1.40$; $p = 0.17$; Cohen's $d = 0.26$).

Experiment 2 provided evidence that differences in metamnemonic resolution across conditions did not stem from how cue characteristics were weighted when making JOLs. Thus, for brevity, we focus our analysis on Experiment 3 primarily on the relation of JOLs to recall in each experimental condition.[6]

Accuracy of predictions

Fig. 2 depicts the gamma correlations between JOLs and recall in each of the four cells in this design. A 2 (JOL by generator or learner) × 2 (JOL for generator or learner) repeated measures ANOVA showed a significant interaction between conditions ($F(1,29) = 5.88$, $p = 0.02$, $\eta_p^2 = 0.17$) and a main effect of for whom the JOLs were made ($F(1,29) = 20.42$, $p < 0.001$, $\eta_p^2 = 0.41$). By contrast, who produced the JOLs did not have a main effect on the accuracy of those JOLs ($F(1,29) = 0.96$, $p = 0.34$, $\eta_p^2 = 0.03$). Planned comparisons revealed that generators were more accurate at predicting their own recall than the learners' recall ($t(29) = 2.35$, $p = 0.03$, Cohen's $d = 0.44$) and that learners were more accurate at predicting their own recall than the generators' recall ($t(29) = 4.01$, $p < 0.001$, Cohen's $d = 0.75$). Further, as in the prior experiments, predictive accuracy in each condition was compared to zero. Generators predicted their own recall ($t(29) = 9.02$, $p < 0.001$) and learners' recall ($t(29) = 4.93$, $p < 0.001$) more accurately than chance. Learners predicted their own recall more accurately than chance ($t(29) = 7.86$, $p < 0.001$), but did not predict

[4] No differences on metamnemonic accuracy between orders of JOLs were expected or found.

[5] We first collected data with generators and learners both having a 30 min retention interval, but ceiling effects in generators' memory performance prevented meaningful analysis of these data.

[6] Analysis of the cue characteristics using the Experiment 3 data replicates the findings from Experiment 2: Cue-to-target associative strength, cue distinctiveness, and cue commonality all positively predicted recall whereas cue generation time did not; however, differences across conditions in which cues were used did not correspond to differences across conditions in metamnemonic accuracy. These analyses are presented in the corresponding online supplement.

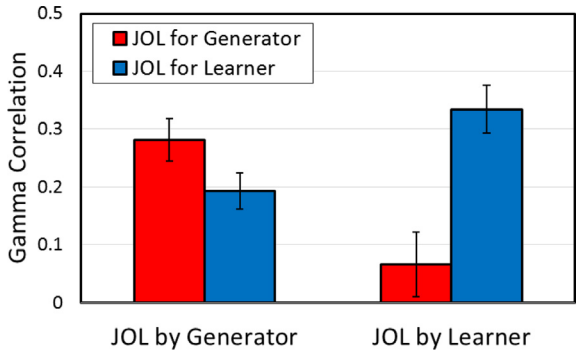

Fig. 2. Gamma correlations between predictions and recall in Experiment 3, as a function of who made the JOLs and for whom they were made. The two outer bars show predictions made for oneself, while the two inner bars show predictions made for others. Error bars show one standard error of the mean above and below the sample mean.

generators' recall more accurately than chance ($t(29) = 1.28$, $p = 0.21$).

Discussion

Experiment 3 introduced a fully orthogonal design in which both cue generators and learners predicted both their own memory and their yoked counterpart's. The results showed that predictions of someone else's memory were consistently less accurate than predictions of one's own memory. That is, generators were less accurate at predicting learners' recall than their own, and learners were less accurate predicting generators' recall than their own. Interestingly, the learners' predictions of the generators' recall were no more accurate than chance.

The misleading-information hypothesis, under which predicting others' memories is difficult because of the misleading effects of one's own experience, would have predicted that one's own experience generating the cues would have contaminated JOLs and made cue generators especially impaired at predicting the learners' recall. The results actually show the opposite pattern: learners seem to be especially impaired at predicting the generators' recall.

These results provide evidence in support of the inadequate-information hypothesis. In this experiment, generators are likely to have meaningful idiosyncratic experiences with the cues (and the cues' relationship to the targets) that will affect their recall; conversely, the learners are less likely to have idiosyncratic experiences with the cues because they did not generate them. Therefore, predicting the generators' recall should be especially difficult for learners who have no knowledge of the generators' idiosyncratic experiences with the cues. The results bear out this hypothesis and reveal that predicting someone else's memory is largely difficult because people have less information about others' mnemonic experiences than they do their own.

General discussion

Across three experiments, we evaluated the accuracy of mnemonic predictions for oneself versus others, both when those predictions concerned memory cues that a subject had themselves generated (as teachers, advertisers, and significant others often do) and when those predictions concerned cues already created by another individual. We also examined what stimulus characteristics characterized effective cues for recall and whether subjects' JOLs for themselves versus others were differentially sensitive to those stimulus characteristics.

The accuracy and bases of mnemonic predictions

The current experiments consistently showed that subjects can predict others' memories more accurately than chance, but less accurately than they can predict their own memories. While other research had suggested that subjects' accuracy in predicting others' memory performance can be very low (Vesonder & Voss, 1985), the present results provide evidence against a strong no-accuracy hypothesis in which subjects are entirely unable to predict others' memories.

Instead, we found that people have some ability to predict others' recall, but with reduced accuracy. One potential explanation of this pattern is the misleading-information hypothesis: Judgments for others are contaminated by aspects of one's own experience that are irrelevant to others' memory. For instance, subjects generating cues might be largely biased by their idiosyncratic experiences producing the cues. Indeed, generators' predictions of others' memories reflected the amount of time needed to generate the cue, even though cue generation time was unrelated to others' mnemonic performance. Similarly, the number of associates influenced generators' predictions of others' recall more than it influenced learners' predictions of their own recall, which likely reflects the fact that the number of associates is strongly related to generators' recall, but only very weakly related to learners' recall (Tullis & Benjamin, 2015a). Generators' judgments of others' recall were less influenced by cue commonality than learners' judgments of their own recall, and cue commonality is more strongly related to others' recall then to generators' recall (Tullis & Benjamin, 2015a). Together, these patterns indicate that generators consistently valued stimulus characteristics that were related to their own future recall rather than ones that were related to others' recall.

While we found some differences in how cue generators and observers weighted different stimulus characteristics, these differences in weighting did not appear to be the primary source of differences in relative metamnemonic accuracy. In Experiment 2, when subjects observed (rather than generated) cues, they weighted the stimulus characteristics more similarly to learners, but were no more accurate than cue generators at predicting others' memories. And, in Experiment 3, learners who had never generated any cues at all were as impaired at predicting cue generators' memory as cue generators were in predicting learners' memory. These results suggest that it was not the contaminating experience of generating the cue that reduced the relative accuracy of the JOLs. Rather, the evidence favors an insufficient-information hypothesis: Reduced metamnemonic accuracy may arise because generators and observers do not have access to learners' personal, idiosyncratic experiences. Learners who are predicting their own future recall can rely upon privileged access to personal, idiosyncratic experiences with the stimuli—such as cue familiarity (Metcalfe, Schwartz, & Joaquim, 1993) and encoding fluency (Hertzog, Dunlosky, Robinson, & Kidder, 2003)—that often accurately align with future recall. The key to increasing mnemonic accuracy, then, may lie in the idiosyncratic experiences that only the learner can access (Lovelace, 1984; Underwood, 1966; Vesonder & Voss, 1985). Providing generators more information about a specific learner whose memory they are predicting may provide information about others' idiosyncratic experiences and thereby increase predictive accuracy about others. In fact, providing speakers with detailed information about their audience can improve the speakers' perspective taking (Isaacs & Clark, 1987; Kahneman & Tversky, 1973); similarly, judgers may be able to make more accurate metacognitive predictions about specific others than unknown others.

Some theorists have argued that making JOLs for others involves the same general processes as making JOLs for oneself (Jost et al., 1998). That is, in both cases, subjects consider and weigh informa-

tion in order to infer future memory performance. This might have led to an equivalent-accuracy hypothesis in which JOLs for someone else and JOLs for one's future self involve similar perspective-taking processes and are similarly (in)accurate; here, however, we found that JOLs for others were less accurate than JOLs for one's own future self. Nevertheless, the results support the general idea that subjects consider a variety of clues, including stimulus characteristics and personal mnemonic experiences with the stimuli, both when making JOLs for others and for oneself. Further, much like the processes involved with predicting one's own memories (Dunlosky & Nelson, 1994), the conditions under which JOLs are elicited impact the stimulus characteristics that are used to infer others' mnemonic performance. In these regards, making predictions about others' mnemonic performance and making predictions about one's own mnemonic performance are very similar. The processes may only differ in kinds of available information to make inferences and the weight given to different cues. How different clues impact judges' predictions about others' memory remains an open question. Future research may explore, for example, how extrinsic study characteristics (e.g., study time and number of repetitions) or characteristics of the learners (e.g., their age or idiosyncratic interests) impact judges' predictions of others' memories.

Implications for perspective-taking

The experiments presented here also make two significant contributions to the broader literature on perspective-taking. First, we showed that learners can accurately judge how well they have taken the perspective of a different learner when creating mnemonic cues for them. While much of the literature on perspective-taking has focused on whether and how learners *cognitively* take perspective of others, we showed that learners can *metacognitively* judge when they have effectively taken the perspective of others. This finding suggests that the degree to which people can accurately metacognitively judge their perspective-taking in other domains, and their reasons for the accuracy or inaccuracy of these judgments, could be a fruitful topic for future research.

Second, our results suggest that failures of perspective-taking do not always result primarily from the interfering effects of one's own perspective. Past work on perspective-taking, such as that on perspective-taking in language comprehension, has often emphasized interference from one's own perspective as a source of perspective-taking failure (e.g., Brown-Schmidt & Hanna, 2011; Hanna et al., 2003; Keysar et al., 2000; Stone et al., 1998). In the present study, however, we found that the contaminating effects of one's own experiences was not the best explanation of the relative inaccuracy of JOLs for another learner was best explained; rather, it was the lack of access to another learner's idiosyncratic mnemonic experiences. Thus, another important constraint on perspective-taking may simply be a lack of information about the other person whose perspective one is trying to take. Although this is a somewhat different constraint than what has been emphasized in the literature on language processing, differences across studies are plausible given that perspective-taking is a complex, domain-specific skills, as evidenced by the fact that individual differences in perspective-taking correlate very weakly across domains (Ryskin et al., 2015). The major constraints on perspective-taking in memory may be different from the major constraints on perspective-taking in language processing.

Conclusion

JOLs for others accurately predicted others' recall, but not as accurately as a learner's own mnemonic predictions. The reduced accuracy of JOLs for others appeared to be largely driven by missing information about the idiosyncratic experiences of the learners. In addition, JOLs for others were influenced differently by stimulus characteristics than JOLs for oneself.

The ability to accurately anticipate others' memories may play a crucial role across many social interactions. For instance, teachers need to judge how well their lectures and notes will support their students' performance. Learning suffers when instructors fail to accurately predict a student's knowledge (Nückles, Wittwer, & Renkl, 2005). Biased estimates of others' knowledge can impair instruction and cause students to pose more questions and ultimately learn less (Herppich, Wittwer, Nückles, & Renkl, 2014). Improving the accuracy of metacognitive judgments of others may be one means of improving instruction. Providing learners access to others' mnemonic experiences may prove to be an important aspect of improving social metacognitive judgments, enhancing instruction, and ultimately supporting student learning.

Appendix A

Target items used across experiments.

admit
analogy
appeal
atlas
band
bowl
breakfast
change
charm
close
clothes
college
culture
dance
decision
democracy
digest
dressing
enigma
exercise
fad
fraud
freedom
game
grief
gross
gym
hero
hobby
holiday
homework
horoscope
instrument
iron
juice
just
legend
lemon
library
logic

(*continued on next page*)

magazine
major
march
medicine
mold
movie
name
novel
obey
occupation
office
olympics
operation
palm
perfume
personality
plumber
professor
restaurant
roommate
selfish
senator
shore
sibling
size
social
street
success
team
television
throne
tradition
valentine
value
virtue
volunteer
weakness
weather
would
yard

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jml.2017.03.003.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005.

Bäckman, L., & Mäntylä, T. (1988). Effectiveness of self-generated cues in younger and older adults: The role of retention interval. *International Journal of Aging and Human Development, 26*, 241–248.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using *lme4*. *Journal of Statistical Software, 67*, 1–48. http://dx.doi.org/10.18637/jss.v067.i01.

Benjamin, A. S., & Diaz, M. (2008). Measurement of relative metamnemonic accuracy. In J. Dunlosky & R. A. Bjork (Eds.), *Handbook of memory and metamemory* (pp. 73–94). New York, NY: Psychology Press.

Birch, S. A. J. (2005). When knowledge is a curse: Children's and adults' reasoning about mental states. *Current Directions in Psychological Science, 14*, 25–29.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: Incremental perspective-taking in language processing. *Dialogue and Discourse, 2*, 11–33. http://dx.doi.org/10.5087/dad.2011.102.

Camerer, C., Lowenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. *Journal of Political Economy, 97*, 1232–1254.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement, 7*, 249–253.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*, 545–565.

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language, 71*, 17–38. http://dx.doi.org/10.1016/j.jml.2013.10.002.

Fraundorf, S. H., Diaz, M. I., Finley, J. R., Lewis, M. L., Tooley, K. M., Isaacs, A. M., ... Brehm, L. (2014). *CogToolbox for MATLAB [computer software]*. Available from <http://www.scottfraundorf.com/cogtoolbox.html>.

Freeman, E., Heathcote, A., Chalmers, K., & Hockley, W. (2010). Item effects in recognition memory for words. *Journal of Memory and Language, 62*, 1–18. http://dx.doi.org/10.1016/j.jml.2009.09.004.

Hanna, J. E., Tanenhaus, M. K., & Trueswell, J. C. (2003). The effects of common ground and perspective on domains referential interpretation. *Journal of Memory and Language, 49*, 43–61. http://dx.doi.org/10.1016/S0749-596X(03)00022-6.

Herppich, S., Wittwer, J., Nückles, M., & Renkl, A. (2014). Addressing knowledge deficits in tutoring and the role of teaching experiences: Benefits for learning and summative assessment. *Journal of Educational Psychology, 106*, 934–945.

Hertzog, C., Dunlosky, J., Robinson, A. E., & Kidder, D. P. (2003). Encoding fluency is a cue used for judgments about learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 22–34.

Hunt, R. R., & Smith, R. E. (1996). Accessing the particular from the general: The power of distinctiveness in the context of organization. *Memory & Cognition, 24*(2), 217–225.

Isaacs, E. A., & Clark, H. H. (1987). References in conversation between experts and novices. *Journal of Experimental Psychology: General, 116*, 26–37.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*, 434–446.

Jost, J. T., Kruglanski, A. W., & Nelson, T. O. (1998). Social metacognition: An expansionist review. *Personality and Social Psychology Review, 2*, 137–154.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*, 237–251.

Kelley, C. M., & Jacoby, L. L. (1996). Adult egocentrism: Subjective experience versus analytic basis for judgment. *Journal of Memory and Language, 35*, 157–175. http://dx.doi.org/10.1006/jmla.1996.0009.

Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science, 11*, 32–38.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36* (ECVP Abstract Supplement).

Kraus, R. M., Vivekananthan, P. S., & Weinheimer, S. (1968). Inner speech and external speech: Characteristics and communication effectiveness of socially and nonsocially encoded messages. *Journal of Personality and Social Psychology, 9*, 295–300.

Lovelace, E. A. (1984). Metamemory: monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 75–766.

Mäntylä, T. (1986). Optimizing cue effectiveness: Recall of 500 and 600 incidentally learned words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 66–71.

Mäntylä, T., & Nilsson, L. G. (1983). Are my cues better than your cues? Uniqueness and reconstruction as prerequisites for optimal recall of verbal materials. *Scandinavian Journal of Psychology, 24*, 303–313.

Mäntylä, T., & Nilsson, L. G. (1988). Cue distinctiveness and forgetting: Effectiveness of self-generated retrieval cues in delayed recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*, 502–509.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The Cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 851–861.

Murayama, K., Sakaki, M., Yan, V. X., & Smith, G. M. (2014). Type I error inflation in the traditional by-participant analysis to metamemory accuracy: A generalized mixed-effects model perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40*, 1287–1306. http://dx.doi.org/10.1037/a0036914.

Nelson, T. O., Kruglanski, A. W., & Jost, J. T. (1998). Knowing thyself and others: Progress in metacognitive social psychology. In V. Y. Yzerbyt, G. Lories, & B. Dardenne (Eds.), *Metacognition: Cognitive and social dimensions* (pp. 69–89). London: Sage.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms*. <http://www.usf.edu/FreeAssociation/>.

Nickerson, R. S. (1999). How we know- and sometimes misjudge – What others know: Imputing one's own knowledge to others. *Psychological Bulletin, 125*, 737–759.

Nickerson, R. S. (2001). The projective way of knowing: A useful heuristic that sometimes misleads. *Current Directions in Psychological Science, 10*, 168–172.

Nückles, M., Wittwer, J., & Renkl, A. (2005). Information about a layperson's knowledge supports experts in giving effective and efficient online advice to laypersons. *Journal of Experimental Psychology: Applied, 11*, 219–236.

Ryskin, R. A., Benjamin, A. S., Tullis, J. G., & Brown-Schmidt, S. (2015). Perspective-taking in comprehension, production, and memory: An individual differences approach. *Journal of Experimental Psychology: General, 144*, 898–915.

Selman, R. L. (1981). The child as a friendship philosopher: A case study in the growth of interpersonal understanding. In S. R. Asher & J. M. Gottman (Eds.), *The development of children's friendships*. Cambridge: Cambridge University Press.

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience, 10*, 640–656.

Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language, 64*, 109–118.

Tullis, J. G., & Benjamin, A. S. (2012). Consequences of restudy choices in younger and older adults. *Psychonomic Bulletin & Review, 19*, 743–749.

Tullis, J. G., & Benjamin, A. S. (2015a). Cue generation: How learners flexibly support future retrieval. *Memory & Cognition, 43*, 922–938.

Tullis, J. G., & Benjamin, A. S. (2015b). Cueing others' memories. *Memory & Cognition, 43*, 634–646.

Underwood, B. J. (1966). Individual and group predictions of item difficulty for free learning. *Journal of Experimental Psychology, 71*, 673–679.

Van Loon, M. H., de Bruin, A. B. H., van Gog, T., van Merrienboer, J. J. G., & Dunlosky, J. (2014). Can students evaluate their understanding of cause-and-effect relations? The effects of diagram completion on monitoring accuracy. *Acta Psychologica, 151*, 143–154.

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language, 24*, 363–376.