Spring 5-15-2015

Rapid Detection and Use of Non-verbal Confidence Cues During Adaptive Memory Biasing

Jihyun Cha
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychology

Rapid Detection and Use of Non-verbal Confidence Cues during Adaptive Memory Biasing

by

Jihyun Cha

A thesis presented to the
Graduate School of Arts and Sciences of
Washington University in
partial fulfillment of the
requirements for the
degree of Master of Arts

May 2015

Saint Louis, Missouri

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to thank Ian Dobbins for his support and advice throughout the project. I also appreciate valuable comments from other members of my committee, Henry Roediger and David Balota.

ABSTRACT OF THE THESIS

Rapid Detection and Use of Non-verbal Confidence Cues

during Adaptive Memory Biasing

by

Jihyun Cha

Master of Arts in Psychology

Washington University in St. Louis, 2015

Professor Ian G. Dobbins, Chair

Prior literature has demonstrated that participants use probabilistic, verbal memory cues ('Likely Old' or 'Likely New') to adaptively bias their recognition judgments. Here we tested whether this is more effective when the cues are the actual videotaped responses of others taking the same recognition test, based on the possibility that observers might use non-verbal confidence signs to modulate their degree of cue reliance on each trial. Experiment 1 demonstrated observers could reliably rate the confidence of others (Models) from single recognition responses ('old' or 'new') and that when doing so, the latency of the model's response was the primary influence, with a secondary influence of non-latency information presumably linked to prosody or facial expression. In Experiment 2, subjects were asked to use these video-taped recognition responses as memory cues while they undertook the same recognition test. The model's responses reliably biased the observer's recognition judgments and were reliably moderated by the model's response latency; non-latency signs of confidence were not reliably influential. In Experiment 3, observers were asked to explicitly rate the confidence of the model's responses before using them during their own recognition judgments. Their initial ratings of the model's confidence were sensitive to latency and non-latency confidence signs; however, the subsequent recognition judgments of the observers were again only sensitive to the latency of the model’s recognition judgments. Overall, subjects can rapidly read non-verbal confidence information contained in brief single recognition responses. However, when using these to inform their own recognition judgments, only response latency appears to reliably moderate the biasing of recognition judgments.

Introduction

Optimal decision-making requires using information from all available sources. In the case of recognition judgments, this means using not only one's internal memory evidence, but also capitalizing on environmental cues that potentially signal the likely familiarity or novelty of encountered stimuli. Previous work using an Explicit Memory Cueing paradigm has shown that observers use anticipatory verbal hints ('Likely Old' or 'Likely New') during recognition to elevate overall performance (Dobbins, Jaeger, Studer, & Simons, 2012; Jaeger, Konkel, & Dobbins, 2013; Jaeger, Lauris, Selmeczy, & Dobbins, 2012; O'Connor, Han, & Dobbins, 2010; Selmeczy & Dobbins, 2013). Since the hints or cues are probabilistic (~75% valid) this requires joint consideration of recommendations and internal evidence, a conclusion bolstered by the fact that gains during the paradigm vary as a function of individual differences in meta-mnemonic awareness (Selmeczy & Dobbins, 2013).

Although observers robustly use the cues in the verbal variant of the Explicit Memory Cueing paradigm, here we extend the procedure to more social cues by having observers view the video recorded responses of other individuals (from here on called 'Models[1]') taking the same recognition test. The reason for examining these socio-perceptual cues was twofold. First, we anticipated that observers might be motivated to more carefully consider the cues given their greater social value (i.e., other students taking the same recognition test). Second, and more critically, we were interested in whether subjects would be swayed at the trial level by the varying confidence of the Models despite the fact that each clip merely documented the Model responding 'old' or 'new' with no explicit statements of confidence. Thus, if they were swayed by confidence, it would necessarily mean they were very sensitive to non-verbal indicators of Model confidence, which would be an important skill because, as we explain below, confidence signals recognition reliability.

Rapidly appreciating the confidence of another's report would improve the use of cues because trial-wise recognition accuracy and subjective confidence are generally positively related (Desoto & Roediger, 2014; Roediger & Desoto, 2014; Roediger, Wixted, & Desoto, 2012). More concretely, consider a Model who is 75% accurate overall and conveys no confidence information from trial to trial. If the Model responds 'old' on any given trial, an ideal observer knows there are 3:1 odds that the upcoming stimulus is in fact old even before seeing it. Under a basic signal detection model observers can benefit from such cues by shifting a decision criterion in either a lax or strict direction based on the direction of the cue (viz., 'old' or 'new'). Under this approach, the basic signal detection model indicates that an observer with a d' of 1 (69% correct) could elevate performance to 78% correct. While appreciable, the observer is restricted to the same strategy on each trial (Jaeger et al., 2012).

In contrast, consider a Model that still is correct 75% of the time overall, but varies across trials from complete guessing (50%) on half of the trials, to complete certainty and perfect performance (100%) on the remaining half of the trials. If our hypothetical observer could "read" the confidence state of the Model then he or she would presumably completely disregard the Model during guessing (yielding 69% correct, his or her baseline ability) and totally defer to the Model when the Model was perfectly certain (yielding 100% correct). This would then earn a net performance of 84.5%, which is an improvement over the case in which all Model responses were weighted equally across trials. Critically, this improvement occurs even though the Model still has the same overall reliability as before, namely, 75%. While this is an extreme example, the logic generally applies to more continuous gradations of confidence and accuracy; namely, an observer that is sensitive to confidence cues can modulate their degree of reliance on an external source during each individual trial. Given this ability, the observer would presumably compare his or her own confidence to that of the external source and on trials where the judgments initially conflict, defer to the source when it appeared more confident (and hence was presumably more likely to be correct).

Recent work on perceptual decision-making in fact demonstrates that joint gains above individual performance levels depend on the ability of observers to share confidence information explicitly (Bahrami et al., 2010). In an oddball detection task using Gabor patches, dyads of participants shared their interim answer (individual decision) and discussed with each other before reaching a final joint decision. The results showed that two observers of similar visual sensitivity yielded better performance as a dyad than their respective individual performance, and a computational model premised on sharing of weighted confidence (the probability of being correct) predicted the actual data better than any other models assuming sharing of mere decision outcome or sharing of direct sensory signal of individual. However, in Bahrami et al. 2010, the sharing of confidence information was explicit because the observers were free to interact extensively before committing the joint decision. Here we instead examine whether observers are sensitive to confidence information even when its conveyance is neither explicit nor interactive, but instead carried in the non-verbal facial expressions or prosody information accompanying the single recognition utterance of another individual.

More direct demonstration of people’s tendency to use explicitly delivered confidence of others is shown by Schneider and Watkins (1996). In their experiment 2, the confederate and the participants took turn to say the recognition responses, accompanied with their 3-point confidence ratings. As in many memory conformity studies, this study also focused more on people’s bias toward other’s responses rather than the potential benefits from doing so, however, they did show that there was a linear trend of people following the confederate’s confidence rating, which is, whenever the confederate makes high confidence judgment, the participant not only follow the judgment, but also they rated their own confidence to be high (and vice versa for low confidence). Again, here the confidence of another person was explicitly given to the participants in a form of self-rating.

Social psychology studies have demonstrated that people can make accurate social judgments from a brief (from a few seconds to less than 5 minutes) excerpt of expressive behavior sampled from the behavioral stream (referred to as ‘thin-slices’), even in the absence of personal interaction with those whom they are rating (Ambady, Bernieri, & Richeson, 2000). It is demonstrated that people can infer a wide spectrum of constructs such as internal states, personality, social relations or even the long-term performance outcomes from these very limited samples of behavior. Studies on thin-slicing have often focused on demonstrating how this very limited exposure can lead to accurate inference of prolonged social, personality characteristics which one might assume would require more extended observation to reach reliable judgment.

In the current report, however, we are not interested in long-term, or trait-like Model characteristics, such as a general tendency towards over-confidence, but in the trial-wise variation in confidence as a Model responds to individual memory probes. One type of cue that can signal the trial-wise confidence of the speaker is the relative order of the response when there is more than one speaker who provides external source of information. Wright and Carlucci (2011) suggested that participants believed the speaker who responded first to be more accurate and more confident than a subsequent speaker, even when they chose the response order themselves. As the researchers suggested, this might be related to the fact that in most situations the person who introduces a topic into a discussion might be more accurate. However, when there is only one person who provides the hints and thus, there is no such contextual cue that signals relative confidence of the person compared to another, can people still infer how confident his/her response is?

Some evidence suggesting that extracting this information may be possible comes from spoken language research demonstrating that listeners use various prosodic features such as latencies, intonation, and pitch to estimate others' metacognitive states (Brennan & Williams, 1995; Pon-Barry, 2008). For example, Brennan and Williams showed rising intonation and longer latencies are related to lower feeling-of-knowing (FOK) of speaker, and demonstrated these prosodic features led to lower ratings of feeling-of another's knowing (FOAK) by listeners. Critically, the audio or video stimuli used in these speech sample studies were mostly recorded in a question-answer format, where the speakers were freely responding to general knowledge questions. Thus there were ample opportunities to extract prosodic information in the speech sample. In contrast to this, as we describe below, we examine whether extremely restricted videotaped reports could nonetheless convey confidence information, which speaks to the general sensitivity of observers to non-verbal confidence cues.

The brief video clips in current study provide both visual as well as prosodic properties that potentially cue the memorial confidence of the Models, however, the clips only consist of single utterance of “Old.” or “New” for each trial lasting several seconds. Additionally, since the Models reported their recognition judgments in isolation and were unaware that we were interested in the ability of other participants to extract or sense the confidence of these simple reports, they were unlikely to strive in being particularly expressive of their varying confidence levels and thus the current stimuli can be seen as representing a fairly demanding situation for confidence detection and influence. Any demonstration that observers were sensitive to the trial-wise variation in confidence under these circumstances would necessarily mean that memory confidence would be easily detectable and likely more influential with more extended exposures or direct interactions.

Summary

Optimal decision-making assumes full use of available information during judgments. Our previous research demonstrates that probabilistic verbal cues with a fixed, experiment-wide reliability elevate recognition performance (Jaeger et al., 2012; Selmeczy & Dobbins, 2013). However, in more naturalistic settings, observers may be able to profitably vary their cue reliance on individual trials if they are sensitive to the signals of confidence of individual reports. Given that confidence anticipates success in recognition memory, such a skill would be useful.

Experiment 1 first examined whether or not subjects could judge the confidence of Models when directly asked to do so without the concomitant requirement of taking a recognition test themselves. That is, after filming four Models taking a verbal recognition test, we brought in a pool of subjects and asked them to try to judge the confidence of each trial/clip of the Models' responding. If subjects were unable to do so, then there would clearly be no need to examine how observers might use the non-verbal confidence information contained within these same clips when trying to integrate them into their own memory judgments. Because Experiment 1 demonstrated observers could in fact reliably gauge the confidence of the Models during individual taped responses, in Experiment 2 we then went on to examine if they naturally used this confidence information when asked instead to bolster their own recognition performance using the taped Model responses during the same recognition test (matched materials and order). Critically, no mention of Model confidence was made and observers were simply told to use the Models' 'old' or 'new' responses as additional information when they themselves were evaluating the same recognition materials. Experiment 3 then examined whether confidence information in the clips had a greater effect on subjects' own recognition judgments when, during each trial, they first explicitly rated the confidence of the Model in the clip and then proceeded to use this information to bolster their own recognition judgments.

Experiment 1

Experiment 1 examined whether subjects could rapidly 'read' the subjective confidence of Models taking a basic single item recognition test and being filmed responding either 'old' or 'new' on each individual trial. Such ability would likely be a prerequisite for using the non-verbal confidence cues of others to more efficiently bias one's own judgments during recognition based on others' responses during the same test.

Methods

Participants Five undergraduate students from Washington University in St. Louis were paid $20 for their participation and served as Models. Data from one were excluded for failure to follow the instructions (viz., failure to make a verbal response for the half of the trials). The experiment also included 40 undergraduate students (average age = 19.2 years; 30 female) who served as raters and attempted to gauge the confidence of the Models during the individual trials of a recognition test. The four Models provided permission for the video recordings of their faces and voices to be used later as stimuli in other experiments and all the raters provided informed consent in accordance with the university's institutional review board. Data from one trial of Model 3 was excluded from all the association analyses due to failure to make a verbal response. Also, the initial trial from all the Models was also excluded due to the large reaction time demonstrated on the first trial of the test procedure.

Materials For each Model, a total of 120 words were randomly selected from a pool of 1215 words with an average of 7.09 letters, 2.35 syllables and average Hyperspace Analogue to Language (HAL) frequency (Balota et al., 2007; Lund & Burgess, 1996) of 4624.39. From this set, 60 were assigned as old items and 60 as new items. The study and test materials were presented via Matlab's Psychophysics Toolbox (Brainard, 1997; Pelli, 1997). During the Models' tests, responses were video-recorded using a Logitech HD Webcam C310.

Procedure for Models During the study, each word was preceded by a fixation cross for 500 milliseconds, signaling the start of each trial. Study items were presented for 3 seconds while the participants were asked to intentionally memorize each item for a later unspecified memory test. At test, the recognition test procedure was explained and the Models underwent 6 to 18 practice trials in order to familiarize themselves with the keyboard response requirements and prevent them from looking at the keyboard during testing. The test was self-paced and each test trial was started with a brief beep that served as a marker for isolating trial onset in the video clip for each trial. While each item was presented on the screen, the Models indicated their recognition judgment verbally by saying either "old" or "new" aloud. After making a verbal response, they then pressed the space bar to proceed to a rating screen and wherein they indicated high, medium or low confidence using 'j', 'k', or 'l' keys, for old judgments and 'f', 'd', or 's' key for new judgments. All keys were numbered and colored to indicate each confidence and old-new judgment combination and there was a prompt on the screen to help the participants remember the key assignment.

Procedure for raters The four sets of 120 video clips were used as stimuli for a rating task in which new observes attempted to rate the confidence of each individual response. The participants were randomly assigned to a set of clips (1 Model) as raters and each set was rated by 10 separate raters. The raters watched the serially presented clips, using headphones to hear the responses. They were accurately informed the responses were of another group of students taking a recognition test and that their task was to rate how confident the Model's recognition response was, regardless of the whether the Model's verbal response was either "old" or "new". After watching each clip, a screen asking "How confident do you think this person is in his/her own response" appeared together with a prompt of scale instructing pressing key 1 for low confidence, 2 for medium confidence and 3 for high confidence. The rating was self-paced and the order of video presentation was kept the same as the order of trials during the Model's original recognition test. Critically, the raters were not informed that the Models keyed in their response confidence during each trial and this keyed response was not visible in the clip itself.

The length of these video clips was ranged from 2.7 s to 14.65 (extreme values were from the first trials of each Model), and the average was 3.92 s (Mean *SD* across Models = 1.02 s). The voice response latency ranged from .48 to 5.66 s (Mean = 1.89 s, *SD* = .75 s).

Coding of Model clips Each clip lasted from the beginning of the trial until a confidence key was pressed. Response latency was defined as the latency before each "old" or "new" response as determined by software described below.

Verbal onset latencies were identified by converting each clip to an audio file and using audio software (eBpoweramp Music Converter, *illustrate*) to automatically estimate the initial period of silence, defined as periods of less than -20 dB continuing for longer than 20 milliseconds. For one Model with noticeably louder ambient noise, a -10dB filter was instead applied. Any initial period below these levels was considered silence. This procedure was used to identify the general location of the response onset. Following this, each waveform was plotted visually to verify that the waveform onset did in fact correspond to an “old” or “new” utterance. In cases where fillers (e.g., “um”, or “uh”) occurred or the response was changed, the visual waveform was used to identify the final ‘old’ or ‘new’ utterance (WavePad, NCH Software). Thus for each trial, verbal response latency simply equaled the duration of measured silence in the beginning of the clip.

Results

Model Performance The Models were generally accurate during recognition with *d’* discrimination values of 1.88, 1.32, 0.91, and 1.70 respectively for Models 1 through 4. Table 1 illustrates further characteristics of their performance focusing on their meta-mnemonic awareness (viz., the trial-wise relationship between confidence and success), the relationship between the latency measure and success, and finally the relationship between the latency measure and subjective confidence. To examine the correspondence between the Models' self-rating of confidence and their memory accuracy, we computed the Gamma correlation (Goodman & Kruskal, 1954) between confidence and success at the trial level. The relationship between latency and accuracy was measured using simple correlation coefficient (see Table 1.).

The data are consistent with the prior literature and reliable in all cells. Confidence was generally positively associated with accuracy (Koriat, Goldsmith, & Pansky, 2000; Roediger & Desoto, 2014; Roediger, Wixted, & Desoto, 2012) as indexed by gamma. Furthermore, response latency was inversely associated with accuracy. Finally, response latency and confidence were inversely related (Ratcliff & Murdock, 1976; Robinson, Johnson, & Herndon, 1997). Thus rapid confident responses tended towards success and slow uncertain responses tended towards errors.

To sum up, Model confidence, response latency, and performance were all reliably associated in the expected manner. This in turn indicates that it would be beneficial if others could glean this confidence-linked information from the videotaped responses, since it is in fact diagnostic of trial-wise accuracy. Thus we next examined whether a second set of subjects could detect this information when directly asked to do so.

Rater Performance The first analysis examined the correspondence between the raters' estimates and the actual keyed confidence of the Models using the gamma statistic. Again, the rater could not see the key press of the Models and hence any correspondence must be driven by non-verbal information in the video clip.

The average gamma of 40 raters was .52 (*SD* = .17) and reliable for 35 out of 40 raters via permutation tests with 5000 replications (*p*s < .05, 2-tailed). Thus observers are sensitive to the confidence of others, even during brief, single utterances of a recognition response ("old" or "new"). Furthermore, the gamma relationship is on par with that of the Models when using their own confidence to predict accuracy. That is, the metacognitive accuracy of the Models (with respect to their own accuracy – Table 1) was comparable to the social metacognitive accuracy of the raters (with respect to the Model's confidence).

Does the rater's sensitivity depend on the types of responses?

Based on prior literature demonstrating greater metacognitive awareness for old versus new judgments (Weber & Brewer, 2004; 2006), we also were interested in whether the raters' sensitivity differed for the old versus new responses of the Models. Consistent with the prior literature, the Models themselves demonstrated superior metacognition for their old responses than new responses. Average gamma for old responses across the four Models was .76 (*SD* = .19) and gamma for new responses was .40 (*SD* = .13). This raises the possibility that the Models might be easier to 'read' by the raters when giving 'old' responses because they themselves have greater insight into their own strength of evidence when judging stimuli as old. Figure 1 (left panel) suggests that the gamma values were higher for 'old' than 'new' model responses. We tested the reliability of this difference through bootstrapping. Figure 2 (left panel) shows the bootstrap distributions of the mean gamma differences between old and new response gammas, and the 95%, percentile confidence interval does not contain zero. Thus the results demonstrate that 'old' responses were easier for the raters to 'read' than 'new' responses.

What Model characteristics influenced the raters?

The analysis of gamma scores demonstrates that the raters can extract the confidence of the Models on individual trials, however, it is not clear what information they are using when doing so. First, we focused on the latency to respond, which has been demonstrated to be important indicator of confidence in spoken language (Brennan & Williams, 1995; Pon-Barry, 2008; Swerts & Krahmer, 2005). Critically, Table 1 demonstrated that each Model's response latency was linked to their own confidence and therefore, raters may be sensitive to this correspondence and use perceived speed of responding as one proxy for confidence. Alternatively, they might be sensitive to additional expression or speech prosody information not reliably reflected in latency but nonetheless linked to the Model's self-rated confidence. These possibilities were examined via multi-level modeling in which the confidence estimates of the raters were modeled as a function of three trial-wise characteristics of the Models' recognition responses, namely, the Model's Response (MR; "Old" = 1, "New" = 0), the Model's Verbal response Latency (MVL), and the Model's self-rated Confidence (MC; Low = 1 to High = 3). Again, the MC variable acts to potentially soak up any confidence-linked variance that is not captured by the first two variables (e.g., facial expressive content). The modeling was implemented in the R language using the nlme package (Pinheiro, Bates, DebRoy, Sarkar, R Core Team, 2015). The confidence and the latency factors were centered within each subject to establish a meaningful zero point and to ease interpretation of any interactions that might arise. In this and subsequent analyses we begin with full models containing all possible (or theoretically meaningful) interactions, which are then simplified by removing higher order terms until likelihood ratio tests indicate a significant loss of fit. The intercept of each rater was treated as a random effect.

The three-way interaction between MC, MR, and MVL was not significant. Following its removal, none of the two-way terms was significant (*ps* > .55) and their removal did not degrade the fit $\chi^2(4) = 2.99$, $p = .56$. This resulted in the final main effects model in Table 2. Critically, the Model's confidence and response latency both appear to uniquely predict rater confidence estimates, suggesting that subjects are using the latency of responding as a proxy or heuristic for confidence, as well as gleaning additional information (represented by the confidence variable) from the clips presumably linked to either the Models' expressions and/or prosody of their responses. Additionally, the response type factor was also significant, which demonstrates that old responses were judged as more confident than new responses by the raters. This provides another demonstration of their ability since the Models were in fact more confident for old than new responses [mean confidence across Models: 2.31 (.22) vs. 1.66 (.33) for old and new responses, respectively].

Discussion

Raters were able to reliably extract confidence information from the single utterances and expressions of the Models. The estimates given by the raters were influenced by the Model's response latency, the type of response ('old' versus 'new') and other information not captured by these two variables but which covaried with the actual Model confidence (indicated by key presses outside of the clip frame). The latter suggests that over and above latency cues, other factors in the clips conveyed confidence information (presumably, paralinguistic features, facial expression or perhaps movements). Additionally, there was a main effect of Model response type, with raters giving higher confidence estimates when the Models responded "Old" versus "New", although this effect was considerably smaller than the other two main effects.

Experiment 1 demonstrates that not only are there reliable relationships between confidence, accuracy and response latency for participants (Models) taking recognition tests, but that observers are sensitive to some of these relationships when viewing the simple, single responses of others. This is remarkable given that the average duration of the clips was short as 3.92 seconds and the duration of the actual utterances was shorter than this. Further, the Models did not know that others would be rating their perceived confidence and therefore they were likely not attempting to be overly expressive. These findings motivate the question as to whether or not observers are naturally inclined to use this non-verbal confidence information when instead of directly rating others, they are instead trying to use the Model's video-taped responses to elevate their own recognition performance. We examined this possibility in Experiment 2 providing these video clips as a more social version of the explicit memory cueing paradigm.

Experiment 2

Experiment 2 examined the influence of the video clips described in Experiment 1 when used as recognition cues. Additionally, for comparison we also included a condition that used static text cues as is typically done in the explicit memory cueing paradigm. Because text cues cannot convey trial-wise confidence information, performance in this condition serves as a baseline for evaluating any additional influence of the video cues.

Because observers could detect nonverbal confidence in Experiment 1 when directly asked to, and because this information is predictive of Model accuracy, we anticipated that subjects given access to video cues would benefit more than those given access to the corresponding verbal cues when trying to bolster their own recognition performance. Additionally, we again expected that verbal latency and other confidence-linked information in the video clips would be shown to influence the participants in a trial-wise manner such that clips with non-verbal cues suggesting higher confidence, would bias the recognition reports of the subjects more heavily than clips with cues suggesting lower confidence. This possibility was tested via multi-level modeling.

Methods

Participants 75 undergraduate students (average age = 19.8 years; 39 female) from Washington University in St. Louis participated for the course credit or $10 of payment. All participants provided informed consent in accordance with the university's institutional review board.

Materials and Procedure Participants were randomly assigned to one of the three groups (25 per group) in a cued recognition task. The first two groups were cued and the last was an uncued control group. The cued groups consisted of a video and text cue group. For the video group, clips of Model 1 from Experiment 1 were provided as external memory cues. We chose Model 1 because of the Model's high metacognitive ability (gamma between confidence and accuracy = .67) and expressiveness as indicated by the raters ability to judge his confidence (mean gamma between Rater's rating & Model confidence = .46).

Additionally, the verbal responses of this particular Model did not include any fillers (such as "uh" and "um") that have been suggested to have functional role in signaling utterance confidence (Smith & Clark, 1993). There was 1 trial where this Model changed his/her response before he settled on the final response. This trial was coded according to the final decision. For the text cue group, the text of the Model's response ("Old" or "New") was instead provided as the cue. Finally, the control group did not receive any external cues during the recognition test.

The study materials, test materials, and presentation order were fixed and identical to that of Model 1. This was done to avoid confounds that might arise from the arbitrary pairing of Model responses to trials. For example, subjects might become suspicious if the Model responded confidently on a trial in which they (from the subject's perspective) were clearly incorrect. Additionally, using the identical test raises the ecological validity of the paradigm in the sense that it more closely resembles the type of situation in which one might use a confidant's judgments because one (correctly) assumed they had the same experiences and are engaged in the same testing situation. After study, the test began with instructions regarding the cue use. The video group was correctly informed that the video clips were recorded when the

Model was taking the same recognition test as theirs, and therefore they should use the clips as useful hints, while keeping in mind that the answers could be either right or wrong. They were also accurately told that the Model was about 80% accurate overall (the Model's actual accuracy was 82.5%, thus 17.5% of the cues were invalid.). There was no mention that the Model also indicated his or her confidence during the taping (outside the field of view). The text group was correctly instructed that they would be shown another participant's verbal answers to the same recognition test with the remainder of the instructions being identical to the video group. Lastly, the control group was just told that each test word would be presented right after a "XXX" sign. We included this uncued period in an attempt to equate total test time of the control to the cued groups.

For the cued groups, cues preceded each recognition test probe. The relative onset of the cue and the subsequent memory probe was governed by how long the Model took to respond during his initially recorded test (hence, same as the length of each video clip). For example, if the Model took exceptionally long on a given trial, say 6 seconds, then the time between the onset of the cue and the onset of the test probe was 6 seconds. A test probe was immediately presented after the cue disappeared, and remained on the screen until the participants made a self-paced response. Recognition judgments were entered by via keypress ('1' (old) or '2' (new)). After each judgment, a confidence rating was also collected ('1' (low), '2' (medium) and '3' (high confidence)).

Results

We first tested the net performance of the groups to see if the presence of cues improved performance, and if so, whether this improvement was greater for video than text cues (Table 3).

Recognition accuracy was summarized by subtracting false alarm rates from hit rates. A one-way ANOVA on the group factor (Video, Text and Control) revealed a the main effect of group (*F*(2,72) = 7.31, *MSe* = .03, *p* = .001). Pairwise follow-up tests (Tukey's honestly significant difference: HSD) revealed that both the Video and Text groups were more accurate than the Control group (*p* = .001, *p* = .02, respectively) while not differing from one another (*p* = .65). Although the video group was numerically more accurate than the text group, it was not reliably so and therefore, to the extent the observers in that group were influenced by the confidence it was insufficient to yield net gains relative to the text group. However, there are a host of contingencies that must be satisfied for confidence cues (when correctly perceived) to reliably elevate the performance of recognition test takers in this design, and the lack of a reliable increase in net accuracy for the video versus the text group does **not** necessarily mean that the video group was completely insensitive to Model confidence. We return to this issue more thoroughly in the general discussion, however, below we more directly test whether the video group participants were influenced by the Model's confidence through multi-level modeling.

Is the Model's trial-wise influence moderated by his confidence?

To more directly test for an influence of Model confidence on the video group's trial-wise responses we again turned to multilevel modeling (MLM). Here the dependent variable was the binary response of the observers (1 = 'Old', 0 = 'New') and the goal of the MLM was to see if the influence of the Model's 'Old' and 'New' judgments were moderated by his confidence. In other words, does Model cueing have a more potent effect on trials in which the Model is confident versus uncertain?

Each group was initially considered separately for ease of interpretation. The MLM examined the influence of two core factors on the subject's responses, namely, the Item Type (1 = Old, 0 = New) and the Model Response (1 = Old, 0 = New). With respect to the Item Type factor, observers will more often tend to respond 'old' if the item is actually old (Item Type) and this represents a measure of observer accuracy; that is, the correspondence between the item's status and the observer's judgments. However, this Item Type influence should also be moderated by how easy or distinctive each item is if such a measure were available. In the current design each subject received the exact same test list, in the same order. Thus we were able to code a variable reflecting the tendency of each individual item to be correctly judged (Item Ease). To do so in a way that was statistically independent of the Video and Text groups, we simply tabulated the proportion of Control subjects who correctly identified each item yielding a variable similar to consensuality (Koriat, 2008). Again, because this variable was derived from the Control group, when applied to the Video and Text groups it represents an unbiased, normative estimate of the ease of each item at each serial position in the recognition test and will capture information such as item distinctiveness. Thus the first terms in the model predicting the subjects' Old (1) or New (0) responses are Item Type (IT) + Item Ease (IE), and their interaction (IT*IE). The latter term will capture greater probability of a correct old response for an easy than a hard old item, for example. This type of model is termed a linear probability model, and the coefficients are easily interpretable in the context of recognition memory[2]. For example, in the absence of an interaction, the coefficients of the main effects terms represent the increase in the probability of responding 'Old'. Thus the Item Type effect is, when added to the intercept term, the predicted hit rate for items of average ease, and the intercept term in isolation corresponds to the predicted false alarm rate for items of average ease.

The second key factor of the MLM reflects the influence of the Model's responses on the subject's responding. Here we expect the Model's Response (MR) to robustly influence the subject's tendency to respond old or new (viz., bias the reports). However, this effect should be potentially moderated by two factors. First and foremost, we examined whether the influence of the Model's response was moderated by its latency (MR*MVL). As Experiment 1 demonstrated that subjects used response latency to estimate Model confidence, a significant interaction in the current experiment would indicate that subjects were using confidence to moderate how strongly they relied upon the cue. The second potential moderator was Model's self-rating of confidence, which would capture any influence of confidence not captured by the response latency variable (MR*MC)[3].

Equation 1.

Subject Response = Item Type (IT) + Item Ease (IE) + IT*IE + Model Response (MR) + Model Verbal Response Latency (MVL) + Model Confidence (MC) + MR* MVL + MR*MC.

In the case of the Video group, a confidence influence would be revealed through significant MR*MVL and/or MR*MC interaction terms. When this model is applied to the Text group, these interaction terms should not be significant since text cues cannot convey any indicators of Model's confidence. Finally, in the case of the Control group, the Model Response (MR) term should also not be significant since they do not actually receive any cues during testing. Thus the latter two groups allow us to examine the validity of the model.

For all groups Subject and Item Type were treated as random effects meaning that the intercept of each subject could vary (i.e., the general tendency towards ‘Old’ responses) and the accuracy could vary (i.e., the tendency of responses to correctly track the item type).

Video Cue Group When applied to the video group, two of the three interaction terms were significant, however, the MR*MC was not (p = .134). Thus, there was no reliable evidence that confidence-cues beyond verbal response latency reliably moderated how influential the Model’s responses were upon the subject’s responding. Removal of this interaction term resulted in a model in which all terms except the main effect of Model confidence were reliable, and removal of the Model confidence factor altogether resulted in the reduced model in Table 4 in which all main effects and interactions are reliable. Comparison of this restricted and the full model demonstrate no reliable loss in fit, $\chi^2(2) = 2.69$, $p = .26$.

The IT*IE interaction confirms that observers are sensitive to the normative difficulty (established in an independent group) of the items such that the link between the item type and observer’s responses was stronger for easy as opposed to normatively difficult items. In other words, subjects are more accurate for normatively easier items. The MR*MVL interaction demonstrates that the influence of an ‘Old’ response by the Model was greater for quick than slow Model responses (hence the negative sign). This is consistent with the use of latency as a signal of confidence such that subjects were more influenced by responses that were more likely confident.

Text Cue Group The model in Equation 1 was also applied to the group who received cues in the form of static text (‘Old’ and ‘New’). Because the static text cues for this group cannot possibly convey trial-wise confidence information, fitting this model serves as a type of validation check on the MLM approach.

When applied to the Text group only the IT*IE effect was reliable (all others $p > .086$). Removal of the non-significant MR*MC and MR*MVL interaction terms yielded a model in which the main effect of Model confidence (MC) and the main effect of Model response latency (MVL) were not reliable (*ps* > .42). Removal of these factors altogether yielded the model in Table 5 in which all main effects and interactions are reliable. This restricted model did not differ reliably in fit compared to the initial full model, $\chi^2(4) = 4.30$, $p = .37$.

Thus the MLM confirms that while the observers in the Text group were robustly influenced by the text cues, these cues were not modulated by indicators of the Model's confidence, which stands to reason since the text cues cannot convey such information. Had such terms been reliable, it would have suggested the model was mis-specified.

Validation with Control Group A further validation of the MLM can be achieved with the Control group. Here, there were no cues provided, and hence neither the Model response factor, nor any of its moderators should be reliable. This was the case as none of the cue terms were significant, and their complete removal did not harm the model fit ($\chi^2(5) = .837$, $p = .975$). Thus the only significant predictors for this uncued group were Item Type, Item Ease, and the interaction between these two[4] (Table 6).

Group Comparison The MLMs suggest different influences operating on the Video and Text groups, and in particular that the influence of the cues in the Video group is moderated by their latency. In contrast, there was no reliable evidence for this effect in the Text group.

However, a firm conclusion that this effect differs requires direct statistical contrast of the groups using a common model. To do this, we included both groups into the final model obtained with the Video group. The key question was whether this Group factor (GRP) influenced the interaction of Model response and Model response latency (GRP*MR*MVL). If so, it would demonstrate a statistically reliable difference across the groups. However, it did not ($p > .197$) and so as we discuss below, the case for the use of confidence as a moderating influence in the Video group is equivocal.

Discussion

Experiment 2 provided equivocal support for the hypothesis that observers would use the non-verbal confidence information of the video cues when incorporating these cues into their recognition judgments. Since confidence is linked to the accuracy of the Model, the active use of such information would provide a greater net benefit than simply having access to the Model's responses without confidence-linked information (e.g., the Text group). However, comparison of net accuracy measures demonstrated that the Video group was not reliably superior to the Text group, although it was numerically so. When a more sensitive MLM approach was applied separately to the groups, it was clear that both Video and Text groups were influenced by the cues, but only the Video group showed a reliable interaction between the Model's response and the latency of that response, such that they were more influenced for quick than slow Model cues. While this, in conjunction with the numerically higher net accuracy, suggests the Video group was gleaning additional confidence information from the cues that was obviously not available to the Text group, direct comparison of this effect across the groups was not significant. Thus, to the extent the Video group was using confidence information, it was not sufficiently robust to drive reliable group differences.

One possibility is that the Video group in Experiment 2 is not gleaning all of the confidence information potentially available in the cues because it requires additional effort and/or they are unaware of their ability to do so. Alternatively, it may be that some individuals choose to try to glean this information during cue integration whereas others do not. In either case, this leads to the conclusion that the registration of confidence-linked information in these brief cues is not obligatory or automatic. If so, then forcing participants to actually rate the confidence of each of the Model's responses before integrating them into their own recognition judgments should maximize the use of confidence cues during cue integration.

Experiment 3

In Experiment 3 we combined the rating procedure of Experiment 1 with the cue integration procedure of Experiment 2, based on the prediction that this would maximize the use of confidence cues by the participants when making recognition decisions.

Methods

Participants 27 undergraduate students (average age = 19.07 years; 17 female) from Washington University in St. Louis participated for the course credit or $10 of payment. All participants provided informed consent in accordance with the university's institutional review board. Data from 1 participant was excluded before the analysis due to failure to conform the instruction (distraction during the task), and data from 2 other participants was also excluded in the reports due to poor performance (Hit-FA rates of .17 and -.03, respectively, see figure 5 for extremeness of these outliers), leaving 24 participants for the further analyses.

Materials and Procedure Participants' task in Experiment 3 was identical to the Video cue group in Experiment 2, except that they made an explicit confidence assessment of the Model's response prior to each cued recognition test trial. As before, participants were correctly told that the video clips were recorded during the same test of another person. Following the presentation of the clip, participants first rated how confident they believed the Model (Model 1 of Experiment 1) to be using a 3-point scale. After estimating the Model's confidence, the test item was displayed for recognition judgment. Again, they were instructed to use the response of the person in the video as useful hint for their memory judgment, and the approximated accuracy of the Model was also informed. As in Experiment 2, we made no mention of using the Model's confidence to moderate his influence although we assumed that the initial rating would spur or enforce this tendency. After making their recognition judgment, the participants reported their own confidence regarding the recognition judgment using a 3-point scale.

Results

Rating Model 1's confidence. As in Experiment 1, we calculated the gamma correlation relating the participants' estimates of confidence to the Model's actual self-reported confidence. The average gamma of 24 subjects was .47 (*SD* = .16) and reliable for 23 out of 24 raters via permutation test (*p*s < .05, 2-tailed), providing converging evidence along with Experiment 1 that people are quite sensitive to non-verbal cues of confidence even in these brief clips.

We again tested if observers' sensitivity to Model confidence differed across 'Old' and 'New' judgments. Both gammas were reliably greater than zero (Figure 1, right panel), but unlike Experiment 1, there was not reliable evidence that the observers 'read' the confidence of 'Old' reports better than that of 'New' reports, indicated by the confidence interval of bootstrapped distribution of difference score containing zero (Figure 2, right panel). Thus, despite the fact that the Model 1 himself was better resolved or metacognitively aware for 'Old' and 'New' responses (gamma of .75 vs. .57), we couldn't replicate the better sensitivity of the observers for old responses than new responses that we demonstrated in Experiment1. However, the numerical direction was the same; namely, the average gamma statistic was higher for 'Old' reports than 'New' reports. The data overall illustrate the considerable sensitivity of the observers when explicitly tasked to estimate the confidence of another.

In order to more directly estimate the cues observers were using to rate confidence, we again turned to MLM using the approach of Experiment 1 in which confidence was modeled as a function of three trial-wise characteristics of the models' responses, namely, the Model's Response (MR; "Old" = 1, "New" = 0), the Model's Verbal Response Latency (MVL), and the Model's self-rated Confidence (MC; Low = 1 to High = 3), in addition to the potential interactions among these factors. Here again, the MC variable acts to potentially soak up any confidence-linked information that is not captured by the Model's response latency. The 3-way term was not reliable in the full model ($p > .059$), and its removal yielded a restricted model in which two of the three, 2-way interaction terms were not reliable (MR*MC and MR*MVL). Their removal yielded the restricted model in Table 7 in which all main effects and interactions are reliable, and comparison of this model to the initial full model demonstrated no reliable loss of fit, $\chi^2(4) = 3.62$, $p = .31$.

The main effect of Model response again demonstrates that subjects rate the Model's 'Old' responses as higher in confidence than his 'New' responses, which is consistent with the actual reports given by this Model. The interaction of Model confidence and response latency on the ratings is illustrated in Figure 3, which indicates while the correspondence between Model's confidence and participant's rating still remained, such relationship was the strongest for the slowest responses, because the faster responses were rated as medium to high confidence in general, suggesting the dominant effect of latency on perceived confidence. Nonetheless, the MLM reinforced the findings of Experiment 1. The subjects are highly sensitive to response latency and rate quicker responses as higher in confidence (MVL effect). Aside from this effect, they appear to be sensitive to other confidence-linked information as well (MC effect) although this residual confidence effect is much smaller than the latency effect.

Testing for confidence effects during cue integration. The analysis above demonstrates that the observers are sensitive to and registering the confidence of the Model. This in turn, suggests they should be able to more efficiently incorporate his responses into their own recognition judgments. We begin by comparing the net accuracy (Hits-FAs) of the current subjects to those of the Text group in Experiment 2, to see if forcing the participants to initially rate the Model's confidence did indeed improve their overall benefit from the cues he provided. The direct comparison of the groups was reliable, *t*(47) = 2.38, *p* = .02, suggesting that subjects gleaned more useful information from the video clips than the corresponding static text cues. Thus it appears that performing the initial rating of other's confidence, yields better cue integration than the Video group of Experiment 2 who failed to reliably exceed the Text group in terms of net performance (see Figure 5 for group comparison). Another evidence for cue integration to better net performance relationship can be demonstrated by the fact that there was a marginally significant correlation between participant's rating performance (gamma between rating of other's confidence and Model's self-confidence) and the participant's cued recognition performance (Hits-FAs) for Explicit rating group (*r* = .40, *p* = 052), suggesting that people who are sensitive to the variation of other's confidence might have been able to use the memory cues more efficiently.

We next applied the model in Equation 1 to the current subjects to directly test whether the Model's confidence moderated the influence of his reports. Replicating the Video group of Experiment 2, in the full model, the interaction of Model response and Model confidence was not reliable (MC*MR, $p > .44$). Its removal resulted in a model in which all main effects and interactions were reliable except the main effect of Model confidence (MC, $p > .83$). The removal of this factor yielded the restricted model in Table 8 in which all main effects and interactions are reliable. Comparison of this restricted model to the full model demonstrated no reliable loss of fit, $\chi^2(2) = .86$, $p = .65$.

The final MLM here, closely resembles that obtained during Experiment 2 although the (absolute) t-value of the MR*MVL interaction is somewhat larger suggesting a greater moderating effect of latency on the influence of the Model's response during recognition judgments. The steeper slopes of Explicit rating group than that of Video group in Figure 4 also indicates the larger moderation effect. The conclusions however are the same; the subjects did not reliably use non-latency confidence information (measured by self-rated Model confidence) over and above the latency information when incorporating the Model's responses into their own recognition decisions. This is somewhat surprising as these same subjects were able to extract this information when directly rating the Model. Nonetheless, it appears that their sensitivity to this content is sufficiently small that it is not reflected in the final recognition decisions. Instead, the subjects are heavily influenced by the Model's response, and the degree of influence is moderated by the speed at which the Model responds.

Discussion

Experiment 3 provided converging evidence that people are sensitive to the confidence of individual recognition judgments and that this rests primarily on the use of the latency of the verbal response. Although observers also seem to glean non-latency confidence information over and above response latency when explicitly rating another, this information is less robust and does not exert a detectable influence in terms of moderating how observers use these responses when attempting to bolster their own recognition. Nonetheless, it appears that by making the participants rate the confidence of the Model prior to rendering their own recognition decisions, we were able to amplify the moderating effect of non-verbal confidence during recognition such that the moderating effect of the Model's response latency was more robust than that of the Video group of Experiment 2 and the net accuracy higher than the Text group.

General Discussion

Experiment 1 demonstrated that people are quite sensitive to variations in another's confidence levels, even if conveyed through extremely brief video clips of single 'Old' or 'New' recognition judgments. This is all the more remarkable given that neither personal history with the Model nor personal interaction between the Model and the rater were needed to yield reliable confidence estimation. Critically, this skill appeared to rely upon at least two, sources of information, namely, the latency of the Model's response, and additional confidence cues not carried via latency (e.g., speech prosody and facial expression). This is the first report demonstrating that observers can extract such a considerable amount of confidence-linked information from such brief, non-interactive, social cues.

Experiment 2 tested whether observers use this confidence-linked information over and above the direction of response, when attempting to incorporate the recognition reports of others (Models) into their own recognition judgments. The benefit for doing so seemed quite high as Experiment 1 demonstrated that not only could participants 'read' the confidence of Models from these simple reports, but importantly, that the confidence of the Models did in fact track their own trial-wise accuracy. Thus we anticipated that non-verbal confidence would clearly influence the degree to which the observers in Experiment 2 were biased by the reports of the Models during recognition testing. However, the influence of confidence information in Experiment 2 was more modest than anticipated. First, the net accuracy of the observers receiving video cues was not elevated above that of those receiving the corresponding text cues (although it was numerically higher). This however is a fairly insensitive measure of the presence or absence of confidence influences because, as discussed below, there are a series of contingencies that must be met before Model confidence can reliably elevate the accuracy of participants using those Models to bias their own judgments. Additionally, a net accuracy increase likely also depends upon the 'headroom' available for improvement over and above a simple fixed strategy that ignores trial-wise confidence variation. To address this, we turned to a more powerful approach of MLM of individual trial responses, which when applied to the Video group in Experiment 2 suggested that observers did moderate their reliance on the Model's reports as a function of the latency of the Model's responses. They did not however, use additional confidence-linked cues beyond latency.

Although there was some evidence for a moderating influence of confidence-linked latency cues, this effect did not survive direct comparison of the Video and Text groups. Based on these findings we concluded the confidence effects were modest in Experiment 2, which begs the question as to why. If the Model's confidence tracks his own accuracy, and the observers are quite capable of assessing this non-verbal confidence when explicitly asked to do so, why did they not use this information more robustly during Experiment 2? Part of the answer may lie in the highly indirect path that confidence must take to ultimately tip an observer's response in the correct direction during this design. More specifically, if we treat the gamma statistics in Experiment 1 as roughly analogous to correlation coefficients, this indirectness becomes apparent. In order for a statistically reliable confidence-linked influence to emerge during recognition testing, a subject has to a) have insight into the weakness or strength of his or her own memorial information (metacognitive awareness), b) have insight into the recognition confidence of the Model (social metacognitive awareness), and finally, c) be in situations where his or her judgments and confidence diverge with that of the Model. The last requirement merely means that if, in the absence of any confidence information, one would have already agreed with the Model's judgment, then clearly confidence cannot further modulate the 'Old' or 'New' response probabilities.

Instead, it is presumably only on those trials in which the observers experience a subjective sense of no reliable memorial information (subjective guessing) or very modest evidence in opposition to the response of the Model, that the Model's confidence becomes important for swaying the outcomes in his direction. Additionally, because both the Model and subject participated in the same test with the same items, these disagreements will likely be fairly rare because both are presumably performing well above baseline (indeed the Model was 80% correct) and because item effects will further enhance subject & Model decision congruence (e.g., both are very unlikely to err on the easy items). Of course, one could view the use of identical tests as an unnecessary draw on the power of the design (given it results in a high natural judgment correspondence), or as a positive feature reflecting a realistic situation in which one draws on the reports of another who has had the same initial experiences prior to the current memory demands. Additionally, as noted in the methods, using a matched test also likely minimizes the likelihood that participants will view the cues as misleading, and it has the added benefit of being a truthful representation of the testing situation. The above considerations led to the possibility that subjects in Experiment 2 were not extracting as much information from the cues as possible when incorporating them into their own recognition judgments, and specifically that they were inconsistently considering the actual confidence of the Models during the recognition decisions.

Experiment 3 tested this possibility by having participants explicitly rate the confidence of the Model's before using their responses as hints during their own recognition judgments. This increased the influence of Model verbal response latency in moderating their use of the Model's reports and it elevated their overall performance above that of the Text group of Experiment 2 who did not have access to confidence-linked information. Despite this, there was no evidence for an influence over and above response latency during the recognition judgments even though the observers effectively used more than simple latency when initially rating the Model's confidence before each recognition judgment.

Overall, these data suggest a remarkable skill in extracting non-verbal confidence from the brief reports of others. Joint consideration of Experiments 2 and 3 suggest that the dominant cue which observers use when incorporating another's recognition reports in these situations is the latency of those reports. In neither case did information over and above this cue moderate the influence of the Model on the observers' recognition decisions.

However, it is premature to conclude that subjects cannot use other signals of confidence during these types of cueing tasks because the relative reliance on different types of confidence cues may depend heavily upon their salience and availability. For example, it is currently unknown whether other expressive cues would emerge as reliable moderators if latency information was obscured or unavailable. In the current paradigm, observers viewed the actual delay between stimulus appearance and the Model's judgments. This information is diagnostic of the Model's accuracy and indeed if there is a high correspondence between latency and other expressive cues of confidence, there may not be much additional benefit to using these additional sources, which also may vary more in their utility across individuals. One way to begin to address these possibilities would be to strip the latency information from the video clips and repeat Experiment 3. If observers were unable to incorporate non-latency confidence information into their recognition decisions it would suggest that there is a fundamental difference between assessing the confidence of another and using the confidence of another to moderate the degree to which people rely on other’s answers when rendering their own judgments. Alternatively, if the moderating effects of confidence were largely preserved, this would mean that observers likely shift to the next most salient (and perhaps next most reliable) indicator of confidence available during these tasks which would indicate that they flexibly adopt heuristics that are far better than not using non-verbal confidence at all, but somewhere short of ideal.

Aside from questions of flexibility and cue precedence, a second important facet of the current data is the social metacognitive ability itself, which as noted, seems fairly remarkable given the brevity and non-interactive nature of the cues. In the current paradigm, participants in Experiments 1 and 3 were deliberately attempting to read the confidence of peers. Although our goal was not to investigate broader social issues of class, gender, age, and race, the paradigm clearly could be used to address these. For example, is there a reliable tendency to rate one gender as generally less confident than another, or do individuals have greater difficulty reading the trial-wise confidence cues of their own versus another gender, or their own versus another age group? Additionally, given the declines in many cognitive and perceptual skills in aging, the use of others for external cues presumably increases in importance, which in turn, leads to the question of whether this skill increases, declines, or remains stable with aging. Of course, addressing these questions would require considerably larger designs with large pools of Models, and again, these particular questions were not directly relevant to the basic question of how or if confidence cues are dynamically incorporated into recognition biases that was addressed in the current report. Finally, another interesting avenue of research might be to examine the correspondence between one's own metacognitive abilities, the ability to glean confidence from the simple reports of others, and the degree to which one's own non-verbal cues are easily read by others. These would seem to be very different cognitive/perceptual demands but the principles of embodied cognition may surprisingly extend to the self-assessment, display, and social assessment of non-verbal memorial confidence[5].

In summary, the current experiments demonstrate considerable skill in the assessment of the confidence of others' simple, binary recognition memory judgments. This skill appears to rely on the detection of latency cues as well as non-latency prosody and/or expressive information, all of which are quickly extracted from brief, single responses. However, the information is not fully used, when observers are attempting to integrate the judgments of others into their own recognition memory reports. In these situations, response latency appears to be the dominant signal of confidence used by the observers. The degree of confidence-cue use is amplified by requiring observers to first explicitly rate the confidence of the Model before using their report to bias their own judgments, however, it remains the case that it is still latency information that appears to be the fundamental cue that is used to moderate reliance on the external source. The consequences of rendering response latency non-diagnostic in these situations remains to be seen, but doing so tests whether it is merely the salience of this information that leads to its dominance, or whether it is the only reliable cue that observers can effectively use in such situations.

Footnote

1. We used upper case M for “Model” to refer a person in the video clips, and lower case m to refer “regression models” to avoid any confusion.
2. The logistic model yielded the same conclusions, but here we report the linear model for the easier interpretation of the parameter estimates.
3. We also considered the possibility that the influence of the model’s response would be moderated by the ease of the items such that it would be more influential for items that were more difficult (MR*IE term). However, this approach is inappropriate because item ease presumably influenced the model’s responses directly when they were initially recorded. That is, to the extent the model’s latency to respond and other non-verbal cues in the clips represent his confidence, then these cues also reflect the normative ease of the items and thus item ease is already captured by the MVL and MC terms.
4. The Item Ease Variable is not independent when applied to the Control group since the variable was derived from the raw data of this group. Thus the absence of other effects in the model should be interpreted cautiously as the Item Ease variable in this case may be overfitting the data.
5. We have some indirect evidence of the correspondence between self-assessment and social-assessment. In Experiment 3, across the participants, the correlation between meta-memory measure (gamma between recognition accuracy and confidence rating) and social-assessment measure (gamma between participant’s rating of Model’s confidence

and Model's self-rating) was reliable (*r* = .52, *p* < .01). However, this is indirect because the meta-memory measure is for cued performance, not the baseline uncued performance.

Table 1. Correspondence among recognition accuracy, response latency, and confidence of each Model

Model	**Accuracy & Confidence (Gamma)**	**Verbal Response Latency & Accuracy (*r*)**	**Verbal Response Latency & Confidence (*r*)**
1	.67***	-.31***	-.37***
2	.67***	-.26**	-.43***
3	.38**	-.20*	-.37***
4	.55***	-.24**	-.33***

* *p* < .05, ** *p* < .01, *** *p* < .001 in 2-tailed tests.

Table 2. Multi-level model predicting rating of other’s confidence (Experiment 1)

Random Effect : ~1|Subjects

	Intercept	**Residual**
Standard Deviation	.24	.61

Fixed Effects:
Rater’s Rating = 1 + Model Confidence + Model Response + Model Verbal response Latency

	PE	***SE***	***DF***	***t-value***	***p-value***
(Intercept)	2.20	.04	4707	55.33	< .0001
Model Confidence	.18	.01	4707	12.78	< .0001
Model Response	.10	.02	4707	4.88	< .0001
Model Verbal response Latency	-.35	.01	4707	-25.70	< .0001

Table 3. Recognition Accuracy for each group (Experiment 2 and 3)

Experiment	Group	Proportion Correct	Hits	False Alarms	Hit-FA rates
Exp. 2	Video	.81 (.06)	.83 (.10)	.21 (.11)	.62 (.13)
	Text	.79 (.08)	.82 (.09)	.24 (.13)	.57 (.16)
	Control	.71 (.12)	.72 (.16)	.30 (.18)	.43 (.24)
Exp. 3	Explicit	.84 (.07)	.83 (.09)	.16 (.08)	.68 (.13)

Table 4. Multi-level model for the Video cue group predicting participant’s response (Experiment 2)

Random Effect:					
Formula : ~1+Item | Subject					
	SD	**Corr**			
(Intercept)	.10	(Intr)			
Item	.11	-.62			
Residual	.35				
Fixed Effects: Response = Item Type (IT) + Item Ease (IE) + Model Response (MR) + Model Verbal response Latency (MVL) + IT * IE + MR * MVL					
	PE	***SE***	***DF***	***t-value***	***p-value***
(Intercept)	.15	.02	2944	6.56	< .0001
Item Type (IT)	.40	.03	2944	13.78	< .0001
Item Ease (IE)	-.36	.08	2944	-4.36	< .0001
Model Response (MR)	.31	.02	2944	16.60	< .0001
Model Verbal response Latency (MVL)	.06	.02	2944	2.88	.0040
IT * IE	.74	.13	2944	5.60	< .0001
MR * MVL	-.07	.02	2944	-2.76	.0059

Table 5. Multi-level model for the Text cue group predicting participant's response (Experiment 2)

Random Effect:		
Formula : ~1+Item | Subject		
	SD	**Corr**
(Intercept)	.12	(Intr)
Item	.15	-.89
Residual	.38	

Fixed Effects:
Response = Item Type (IT) + Item Ease (IE) + Model Response (MR) + IT * IE

	PE	***SE***	***DF***	***t-value***	***p-value***
(Intercept)	.19	.03	2946	7.53	< .0001
Item Type (IT)	.44	.04	2946	12.26	< .0001
Item Ease (IE)	-.55	.09	2946	-6.29	< .0001
Model Response (MR)	.21	.02	2946	10.40	< .0001
IT * IE	1.13	.24	2946	8.03	< .0001

Table 6. Multi-level model for the Control group predicting participant’s response (Experiment 2)

Random Effect:					
Formula : ~1+Item | Subject					
	SD	**Corr**			
(Intercept)	.15	(Intr)			
Item	.22	-.71			
Residual	.41				

Fixed Effects:
Response = Item Type (IT) + Item Ease (IE) + IT * IE

	PE	***SE***	***DF***	***t-value***	***p-value***
(Intercept)	.26	.03	2829	8.01	< .0001
Item Type (IT)	.45	.05	2829	9.63	< .0001
Item Ease (IE)	-1.05	.09	2829	-11.68	< .0001
IT * IE	2.09	.15	2829	14.13	< .0001

Table 7. Multi-level model for the Explicit rating group predicting rating of other's confidence (Experiment 3)

Random Effect : ~1|Subjects

	Intercept	**Residual**
Standard Deviation	.21	.55

Fixed Effects:
Rating = Model Response (MR) + Model Confidence (MC) + Model Voice response Latency (MVL) + MC * MVL

	PE	***SE***	***DF***	***t-value***	***p-value***
(Intercept)	2.10	0.05	2828	44.55	< .0001
Model Response (MR)	0.10	0.02	2828	4.68	< .0001
Model Confidence (MC)	0.14	0.02	2828	7.39	< .0001
Model Voice response Latency (MVL)	-0.48	0.02	2828	-22.10	< .0001
MC * MVL	-0.09	0.03	2828	-3.06	< .0001

Table 8. The multi-level model for the Explicit rating group predicting participant's response (Experiment 3)

Random Effect:					
Formula : ~1+Item | Subject					
	SD	**Corr**			
(Intercept)	.07	(Intr)			
Item	.11	-.78			
Residual	.34				
Fixed Effects: Response = Item Type (IT) + Item Ease (IE) + Model Response (MR) + Model Voice response Latency (MVL) + IT * IE + MR * MVL					
	PE	***SE***	***DF***	***t-value***	***p-value***
(Intercept)	.11	.02	2826	6.29	< .0001
Item Type (IT)	.50	.03	2826	16.65	< .0001
Item Ease (IE)	-.48	.08	2826	-5.96	< .0001
Model Response (MR)	.24	.02	2826	13.09	< .0001
Model Voice response Latency (MVL)	.06	.02	2826	2.77	.0057
IT * IE	.93	.13	2826	7.11	< .0001
MR * MVL	-.08	.02	2826	-3.16	.0016

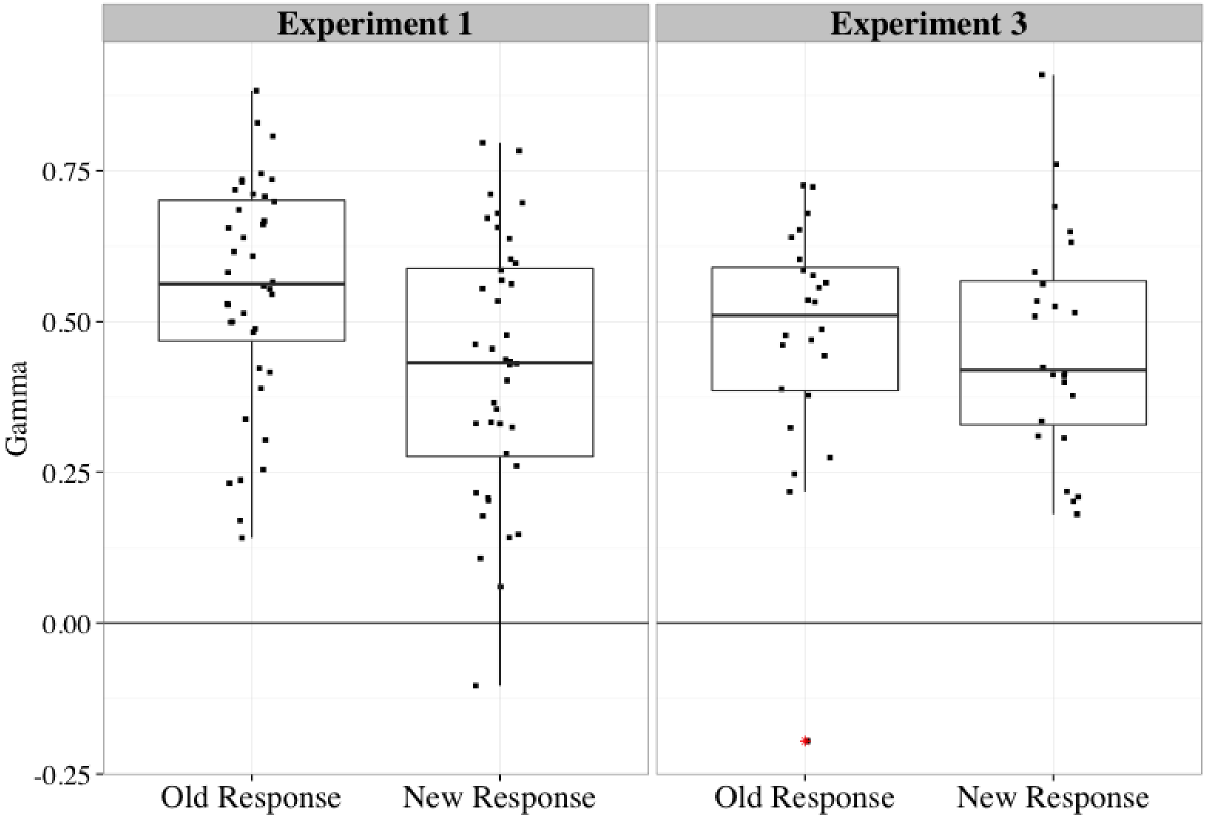

Figure 1. Gamma correlation between rater's rating and model's confidence for 'old' and 'new' responses. The boxes represent inter-quartile range and the thick horizontal lines represent median. The whiskers represent values within 1.5 * the inter-quartile range. Data beyond the end of the whiskers are outliers and marked with asterisks next to the points (as specified by Tukey).

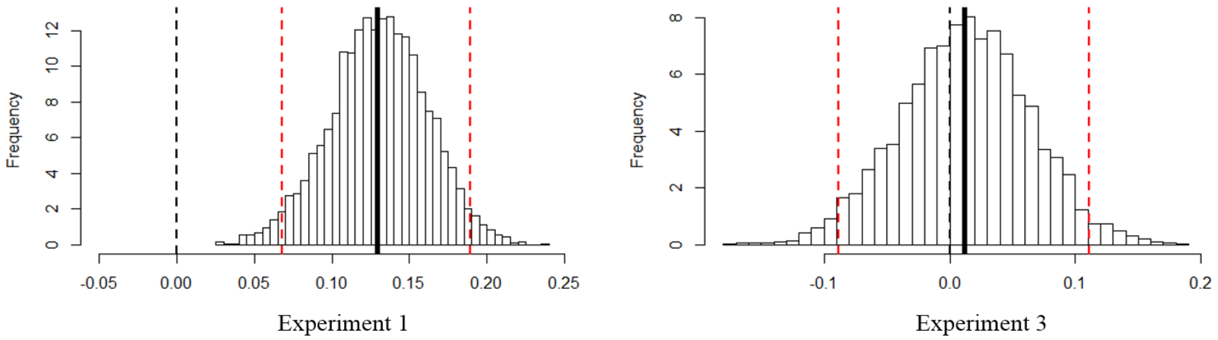

Figure 2. Bootstrapped sampling distribution of difference score between gamma for ‘old’ responses and gamma for ‘new’ responses. Two dashed red lines indicate bootstrap percentile confidence intervals (95%). The thin dashed black line marks the zero point and the thicker black line indicates the observed difference score.

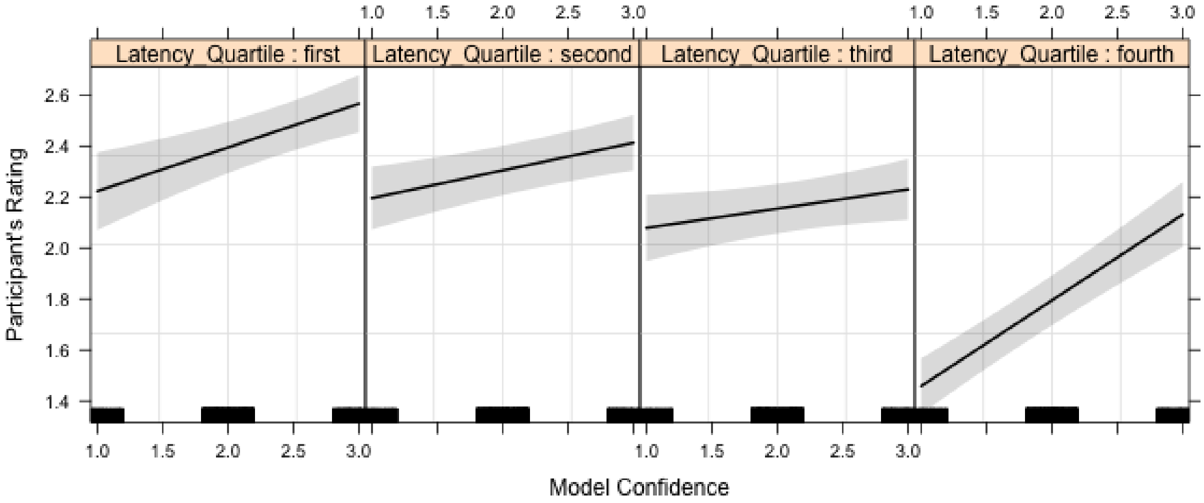

Figure 3. The interaction between Model Confidence and Model Voice Latency predicting participant’s Rating (Experiment 3). Each cell represents the Model Confidence and participant’s rating correspondence for 1st through 4th quartile of Model’s latency, starting from the fastest to slowest responses.

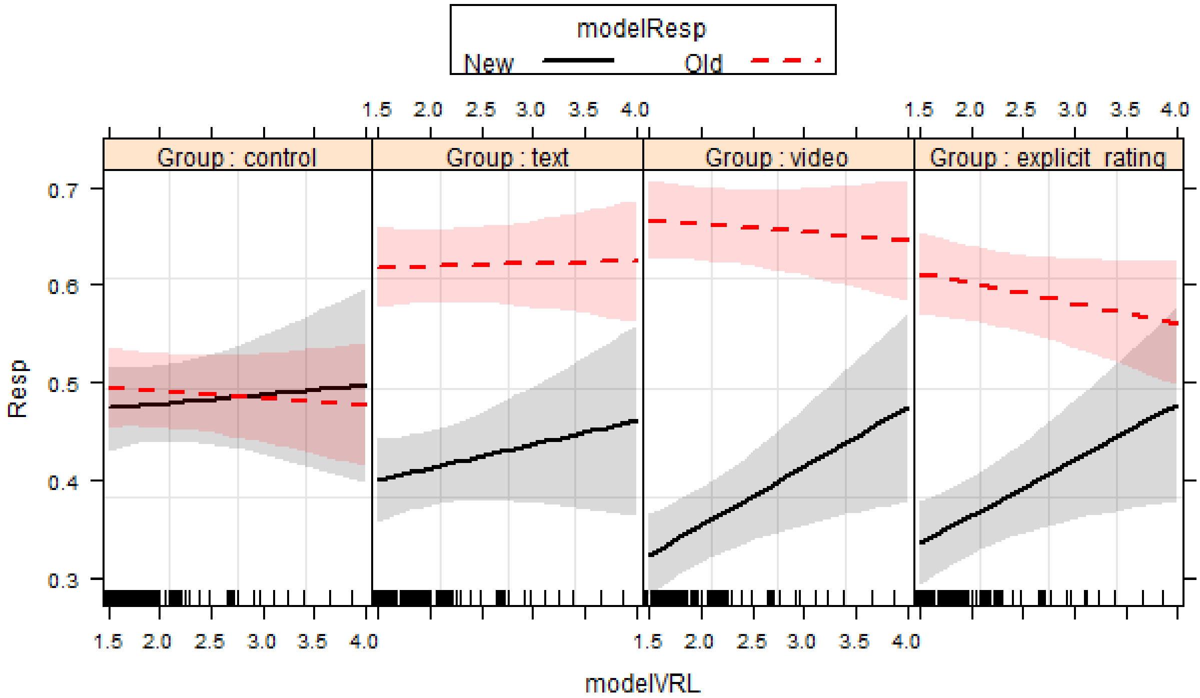

Figure 4. The moderation effect on memory cueing by Model Voice Latency. The y-axis indicates probability of "old" response of participants. The red and black line represents this probability given old and new Model response respectively and the distance between two lines indicates the size of the memory cueing effect, the degree to which the participants followed the Model's response. The slope of each line indicate how much this cueing effect was moderated by Model's voice response latency.

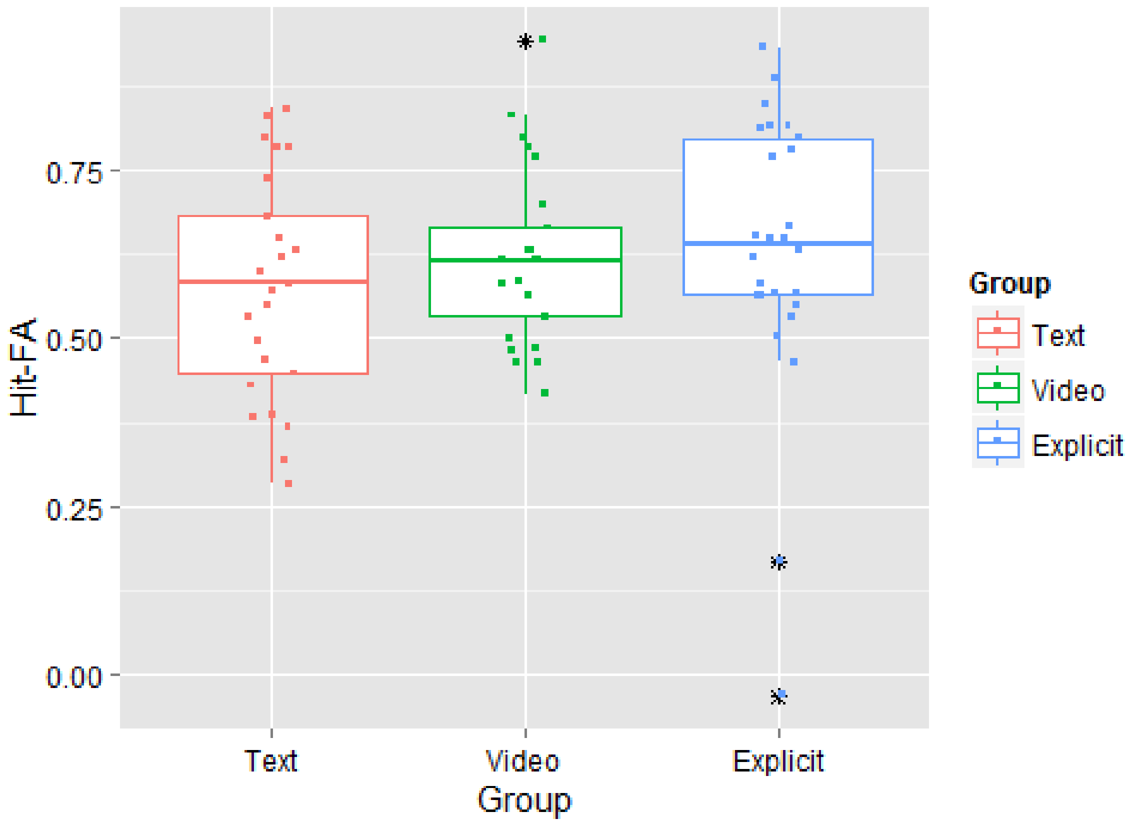

Figure 5. Mean level comparison of accuracy (Hits-FA rate) between cued groups. The boxes represent inter-quartile range and the thick horizontal lines represent median. The whiskers represent values within 1.5 * the inter-quartile range. Data beyond the end of the whiskers are outliers (as specified by Tukey) and marked with asterisks.

References

Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior: Judgmental accuracy from thin slices of the behavioral stream. *Advances in Experimental Social …*, *32*, 201–271. doi:10.1016/s0065-2601(00)80006-4

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081–1085. doi:10.1126/science.1185718

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., et al. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. doi:10.1163/156856897X00357

Brennan, S. E., & Williams, M. (1995). The Feeling of Another′ s Knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of Memory and Language*, *34*(3), 383–398. doi:10.1006/jmla.1995.1017

Desoto, K. A., & Roediger, H. L. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science*, *25*(3), 781–788. doi:10.1177/0956797613516149

Dobbins, I. G., Jaeger, A., Studer, B., & Simons, J. S. (2012). Use of explicit memory cues following parietal lobe lesions. *Neuropsychologia*, *50*(13), 2992–3003. doi:10.1016/j.neuropsychologia.2012.07.037

Goodman, L. A., & Kruskal, W. H. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, *49*(268), 732. doi:10.2307/2281536

Jaeger, A., Konkel, A., & Dobbins, I. G. (2013). Unexpected novelty and familiarity orienting responses in lateral parietal cortex during recognition judgment. *Neuropsychologia*, *51*(6), 1061–1076. doi:10.1016/j.neuropsychologia.2013.02.018

Jaeger, A., Lauris, P., Selmeczy, D., & Dobbins, I. G. (2012). The costs and benefits of memory conformity. *Memory & Cognition*, *40*(1), 101–112. doi:10.3758/s13421-011-0130-z

Koriat, A. (2008). Subjective confidence in one's answers: the consensuality principle. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 945–959. doi:10.1037/0278-7393.34.4.945

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology*, *51*(1), 481–537. doi:10.1146/annurev.psych.51.1.481

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, *28*(2), 203–208. doi:10.3758/BF03204766

O'Connor, A. R., Han, S., & Dobbins, I. G. (2010). The inferior parietal lobule and recognition memory: expectancy violation or successful retrieval? *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *30*(8), 2924–2934. doi:10.1523/JNEUROSCI.4225-09.2010

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. doi:10.1163/156856897x00366

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team. (2015, February 20). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-120. Retrieved from http://CRAN.R-project.org/package=nlme

Pon-Barry, H. (2008). Prosodic manifestations of confidence and uncertainty in spoken language. *Interspeech*.

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190–214. doi:10.1037/0033-295X.83.3.190

Robinson, M. D., Johnson, J. T., & Herndon, F. (1997). Reaction time and assessments of cognitive effort as predictors of eyewitness memory accuracy and confidence. *The Journal of Applied Psychology*, *82*(3), 416–425.

Roediger, H. L., & Desoto, K. A. (2014). Confidence and memory: assessing positive and negative correlations. *Memory*, *22*(1), 76–91. doi:10.1080/09658211.2013.795974

Roediger, H. L., III, Wixted, J. H., & Desoto, K. A. (2012). *The Curious Complexity between Confidence and Accuracy in Reports from Memory*. *Memory and Law* (pp. 84–117). Oxford University Press. doi:10.1093/acprof:oso/9780199920754.003.0004

Selmeczy, D., & Dobbins, I. G. (2013). Metacognitive awareness and adaptive recognition biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(3), 678–690. doi:10.1037/a0029469

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*(1), 25–38. doi:10.1006/jmla.1993.1002

Schneider, D. M., & Watkins, M. J. (1996). Response conformity in recognition testing. *Psychonomic Bulletin & Review*, *3*(4), 481-485.

Swerts, M., & Krahmer, E. (2005). Audiovisual prosody and feeling of knowing. *Journal of Memory and Language*, *53*(1), 81–94. doi:10.1016/j.jml.2005.02.003

Weber, N., & Brewer, N. (2004). Confidence-accuracy calibration in absolute and relative face recognition judgments. *Journal of Experimental Psychology: Applied*, *10*(3), 156–172. doi:10.1037/1076-898X.10.3.156

Weber, N., & Brewer, N. (2006). Positive versus negative face recognition decisions: confidence, accuracy, and response latency. *Applied Cognitive Psychology*, *20*(1), 17–31. doi:10.1002/acp.1166

Wright, D. B., & Carlucci, M. E. (2011). The response order effect: people believe the first person who remembers an event. *Psychonomic Bulletin & Review*, *18*(4), 805–812. doi:10.3758/s13423-011-0089-6