

Overdistribution Illusions: Categorical Judgments Produce Them, Confidence Ratings Reduce Them

C. J. Brainerd, K. Nakamura, and V. F. Reyna
Cornell University

R. E. Holliday
University of Leicester

Overdistribution is a form of memory distortion in which an event is remembered as belonging to too many episodic states, states that are logically or empirically incompatible with each other. We investigated a response formatting method of suppressing 2 basic types of overdistribution, disjunction and conjunction illusions, which parallel some classic illusions in the judgment and decision making literature. In this method, subjects respond to memory probes by rating their confidence that test cues belong to specific episodic states (e.g., presented on List 1, presented on List 2), rather than by making the usual categorical judgments about those states. The central prediction, which was derived from the task calibration principle of fuzzy-trace theory, was that confidence ratings should reduce overdistribution by diminishing subjects' reliance on noncompensatory gist memories. The data of 3 experiments agreed with that prediction. In Experiment 1, there were reliable disjunction illusions with categorical judgments but not with confidence ratings. In Experiment 2, both response formats produced reliable disjunction illusions, but those for confidence ratings were much smaller than those for categorical judgments. In Experiment 3, there were reliable conjunction illusions with categorical judgments but not with confidence ratings. Apropos of recent controversies over confidence-accuracy correlations in memory, such correlations were positive for hits, negative for correct rejections, and the 2 types of correlations were of equal magnitude.

Keywords: memory overdistribution, disjunction illusions, conjunction illusions, noncompensatory memories

Over the past three decades, false memory has been one of the most widely studied topics in psychology, for both theoretical and practical reasons. Practical motivations have been especially prominent, owing to high-stakes situations in which these errors have quite undesirable consequences (e.g., sworn testimony in courtrooms, eyewitness identifications during police investigations, reports of symptoms during emergency room treatment, reports of battlefield experiences, and interrogation-induced reports of criminal acts). Inevitably, the scientific study of false memories has revealed broader distortion phenomena, of which false memories are examples. This article is concerned with one of them, overdistribution illusions.

Overdistribution illusions measure the tendency to remember events as belonging to too many episodic states. Although overdistribution was first studied in connection with false memories, it is a more encompassing distortion that arises from noncompensatory relations among mutually incompatible ways of remembering an event, and it occurs for true as well as false memories (Brainerd, Wang, Reyna, & Nakamura, 2015). Relations exist among memories of events that are objectively compensatory inasmuch as remembering an event in one way ought to preclude remembering it in other ways, by reason of logical or empirical contradiction. On a history test, for instance, remembering that cancer caused Churchill's death and that Einstein was born in Switzerland, which are both false, should rule out remembering that Churchill died from a stroke and that Einstein was born in Germany. Conversely, remembering that Churchill died from a stroke and that Einstein was born in Germany, which are both true, should rule out remembering that Churchill died from cancer and that Einstein was born in Switzerland. However, the data show that when subjects remember an event in one way, their tendency to remember it in other incompatible ways is not reduced by equivalent amounts—allowing Churchill to die more than once and Einstein to be born in more than one place.

The original examples of overdistribution, disjunction illusions, were detected in conjoint recognition experiments (Brainerd & Reyna, 2008). These are standard false memory designs in which subjects respond to recognition tests that are composed of three types of test cues: old targets (O; e.g., *sofa*), new-similar distractors (NS; e.g., *couch*), and new-dissimilar distractors (ND; e.g.,

C. J. Brainerd, K. Nakamura, and V. F. Reyna, Institute of Human Neuroscience, Cornell University; R. E. Holliday, Department of Psychology, University of Leicester.

This research was supported by a Department of Agriculture Grant (NIFA 1003856) to the C. J. Brainerd, a National Institutes of Health Grant (1RC1AG036915) to the C. J. Brainerd and V. F. Reyna, a National Institute of Nursing Research Grant (R01NR014368-01) and a National Science Foundation Grant (SES1536238) to the V. F. Reyna. Some of the results in this article were presented at the 56th Annual Meeting of the Psychonomic Society, Chicago, IL, November, 2015. We thank David Kellen for his comments on a draft of this article.

Correspondence concerning this article should be addressed to C. J. Brainerd, Institute of Human Neuroscience, Cornell University, MVR Hall, Ithaca, NY 14853, E-mail: cb299@cornell.edu

tomato). The novel feature of conjoint recognition is that three types of judgments are factorially crossed with these cues: old? (O?), new-similar? (NS?), and old-or-new-similar? (O-or-NS?). Naturally, subjects exhibit false memories in this paradigm—the probabilities of judging NS cues to be O and of judging O cues to be NS are greater than zero. However, disjunction illusions refer to the fact that the complementary probabilities of true memories (judging NS cues to be NS and O cues to be O) are not then reduced by commensurate amounts; that is, true and false memory are not fully compensatory.

Both findings are shown in Figure 1, for a corpus of 264 sets of conjoint recognition data. The false memory finding can be seen in Panel A, where the mean probabilities of remembering NS cues to be O and O cues to be NS are both well above zero, and the noncompensation finding can be seen in Panel B. Concerning Panel B, if the probability of remembering NS cues to be O reduces the probability of remembering them to be NS by an equivalent amount, then the sum of those two probabilities will equal the probability of remembering them to be O-or-NS, and likewise for O cues. Explicitly, $p(O|NS) + p(NS|NS) = p(O\text{-or-}NS|NS)$, and $p(O|O) + p(NS|O) = p(O\text{-or-}NS|O)$. However, if remembering a cue as belonging to one of these states does not reduce the probability of remembering it as belonging to the other incompatible state by an equivalent amount, then the sum of those two probabilities will *exceed* the probability of remembering it to be O-or-NS; that is, $p(O|NS) + p(NS|NS) > p(O\text{-or-}NS|NS)$, and $p(O|O) + p(NS|O) > p(O\text{-or-}NS|O)$. Panel B shows the latter pattern.

Psychologically, Figure 1 means that item memory exhibits a reality violation that is analogous to a well-known reality violation in physics, quantum superposition (Brainerd, Wang, & Reyna, 2013; Brainerd et al., 2015; Wang & Busmeyer, 2015, in press).

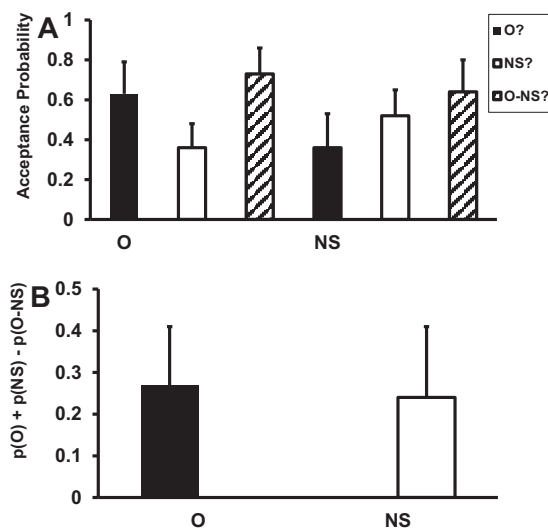


Figure 1. Disjunction illusions in 264 sets of conjoint recognition data. O = old cues (targets) and NS = new-similar distractor cues. O? = the judgment that a cue is old, NS? = the judgment that a cue is new-similar, and O-NS? = the judgment that a cue is either old or new-similar. Panel A plots the mean probabilities of making each one of these judgments for O and NS cues. Panel B plots the mean overdistribution probability $p(O) + p(NS) - p(O-NS)$ for both O and NS cues.

In physics, quantum superposition refers to the fact that particles can occupy mutually incompatible physical states (e.g., spinning up *and* spinning down), whereas in memory, items can occupy mutually incompatible episodic states (e.g., presented *and* not presented in Figure 1). There are deeper commonalities at a mathematical level: The nonadditive relations among response probabilities that demonstrate that items can occupy mutually incompatible episodic states parallel the nonadditive relations that supply classical demonstrations of superposition in physics (see Feynman, Leighton, & Sands, 1965), and both types of relations can be modeled in the same way—namely, as superposed state vectors in a Hilbert space (Brainerd et al., 2013). Theoretical analysis of the types of traces that support memory for a given episodic state shows that some should confer this superposition property, whereas others should produce compensatory relations among incompatible states.

Here, Brainerd et al. (2015) pointed out that in fuzzy-trace theory's (FTT) distinction between verbatim and gist traces, gist traces are noncompensatory while verbatim traces are compensatory. According to that distinction, subjects store and retrieve verbatim traces of targets plus gist traces of their senses, patterns, and meanings in parallel. For instance, in the *sofa-couch* example, a verbatim trace of *sofa's* surface form plus gist traces such as "living room furniture" are stored in parallel during list presentation and retrieved in parallel on memory tests. Relying on gist traces supports noncompensatory responses to both true and false memory probes. With *sofa*, for instance, the gist memory that some of the list words referred to living room furniture is obviously consistent with *sofa* being either a target or a similar distractor, allowing subjects to remember it as old on O? probes *and* as new-similar on NS? probes. With *couch*, the same gist memory is obviously consistent with this cue being either a target or a similar distractor, also allowing subjects to remember it as both old and new-similar. Summing up, the overdistribution pattern in Figure 1 follows if subjects sometimes rely on gist traces of meaning content to make O? and NS? judgments about test cues.

In contrast, as has been widely discussed in research on false memory editing (e.g., Gallo, 2004; Lampinen & Odegard, 2006; Lampinen, Odegard, & Neuschatz, 2004), verbatim traces support compensatory responses to both O and NS cues. With *sofa*, relying on verbatim traces of its earlier presentation supports remembering it as old on O? probes and *not* remembering it as new on NS? probes ("No, *sofa* cannot be new because I clearly remember seeing it on the list."). With *couch*, relying on verbatim traces of the presentation of its corresponding target supports remembering it as new-similar on NS? probes and not remembering it as old on O? probes ("It was *sofa*, not *couch*, that I saw on the list.>").

Recently, overdistribution has also been studied with a source-monitoring paradigm that allows a further example of such distortion to be investigated, conjunction illusions (Brainerd, Holliday, Nakamura, & Reyna, 2014). In source-monitoring experiments (e.g., Dennis et al., 2008; Hicks & Starns, 2006a; Kurilla & Westerman, 2010), subjects usually encode words in one and only one of two (or more) distinct contexts, such as List 1 versus List 2, and then respond to a series of test cues composed of words from each context plus distractors. In most experiments, subjects make an old/new judgment about each cue, followed by a forced-choice source judgment if and only if the cue is judged to be old. In studies of overdistribution, however, the test is modified so that

three judgments are factorially crossed with the three types of cues: presented on List 1? ($L_1?$), presented on List 2? ($L_2?$), and presented on either List 1 or List 2? ($L_1\text{-or-}L_2?$). Disjunction illusions are the twin findings that (a) the probabilities of judging L_1 cues to have been encoded on L_2 and L_2 cues to have been encoded on L_1 are both greater than zero (false memory), and (b) the probabilities of the incompatible judgments that L_1 cues were encoded on L_1 and L_2 cues were encoded on L_2 are not reduced by comparable amounts (noncompensation). Similar to conjoint recognition experiments, then, $p(L_1|L_1) + p(L_2|L_1) > p(L_1\text{-or-}L_2|L_1)$ and $p(L_1|L_2) + p(L_2|L_2) > p(L_1\text{-or-}L_2|L_2)$. Psychologically, this means that source memory displays the same superposition property as item memory.

A key distinction between the conjoint recognition and source-monitoring procedures for studying overdistribution is that O and NS are logically incompatible states, whereas the incompatibility between the L_1 and L_2 states is empirical. (Although no test cues are presented on both lists, they could have been.) Thus, it is possible, without logical contradiction, to request conjunctive judgments (presented on List 1 and List 2? [$L_1\text{-and-}L_2?$]) with the source-monitoring procedure. If the relation between true and false memories of a cue's source is truly noncompensatory, then surprisingly, the probability that the cue is erroneously judged to have been presented *on both lists* should be greater than zero. Brainerd et al. (2014) detected such conjunction illusions in experiments in which they replaced the disjunctive judgments in the aforementioned design with conjunctive ones, while holding other design factors constant. Even more surprising, the subjects in some conditions judged it to be more probable that a cue appeared on all lists than that it appeared on one of the individual lists, which is impossible.

In this article, we report some experiments that dealt with the question of whether disjunction and conjunction illusions can be suppressed with a theoretically motivated manipulation that should diminish reliance on noncompensatory memory information. Our general approach is predicated on the fact that although these illusions have been detected with a variety of materials, they have been measured using memory tests that request *categorical judgments* about test cues (e.g., agree-disagree). Certain lines of research in the judgment and decision making literature (e.g., Kuhberger, Schulte-Mecklenbeck, & Perner, 1999; Kühberger Tanner, 2010; Mills, Reyna, & Estrada, 2008; Reyna et al., 2011) supply theoretical grounds for supposing that categorical judgments allow more latitude for subjects to rely on inherently noncompensatory memories than do other, more differentiated, response formats. The same lines of research suggest that it may be possible to lessen such reliance by shifting from categorical judgments to confidence ratings, thereby reducing or eliminating overdistribution illusions. We discuss the theoretical basis for that prediction below, before presenting the experiments.

Overview of the Research

Brainerd et al. (2015) showed that in source-monitoring designs, noncompensatory relations among incompatible episodic memories also fall out as a prediction of FTTs verbatim-gist principle; that is, if subjects sometimes rely on gist traces when responding to source probes such as $L_1?$ and $L_2?$, source memory will be noncompensatory. For instance, if *trumpet* is a List 1 target, gist

memories such as “musical instrument” can be used interchangeably to correctly accept it on L_1 probes and erroneously accept it on L_2 probes. Consistent with this principle, overdistribution in source-monitoring experiments has been tied to variability in reliance on gist memories: Higher levels of overdistribution are observed in subjects who prefer to rely on gist rather than verbatim memory and in conditions in which gist memories are strengthened by presenting multiple targets that exemplify the same semantic content (Brainerd, Reyna, Holliday, & Nakamura, 2012; Nakamura & Brainerd, 2013).

At a more general level, there is much evidence that subjects rely on gist memories in classic source paradigms. In the Loftus (1975) misinformation procedure, for instance, it has long been known that (a) rates of false memory for suggested events (incorrectly judging them to have occurred during the encoding phase) are higher when they preserve the semantic content of the encoding phase than when they do not (e.g., Bjorklund et al., 2000; for a review see, Titcomb & Reyna, 1995) and that (b) subjects sometimes judge such suggested events to have only occurred during the encoding phase *and* to have only occurred during the misinformation phase (e.g., Thierry, Lamb, Pipe, & Spence, 2010). In another widely used paradigm, process dissociation (Jacoby, 1991), judging that an item was presented with one configuration of contextual details (font, color, and position) and was also presented with another configuration increases as the semantic overlap between the contexts increases (e.g., Brainerd & Reyna, 2008; McBride & Shouder, 2003). More recent evidence of reliance on memory for the semantic content of targets when making erroneous source judgments can be found in a variety of articles, including Arndt (2012) and Ball, DeWitt, Knight, and Hicks (2014).

We assumed as a working hypothesis that reliance on noncompensatory gist memories is at least partially responsible for overdistribution illusions. (We consider another potential contributors in the General Discussion.) If so, it should be possible, while holding other design factors constant, to reduce overdistribution illusions by imposing conditions that diminish reliance on such memories. This brings us to the judgment and decision making literature and a principle called task calibration that FTT uses to account for some surprising effects (e.g., preference reversals) and to predict others (e.g., risk perception reversals, nonnumerical framing illusions).

Task Calibration

FTT is an example of theoretical approaches to judgment and decision making that implement the hypothesis that illusions and biases must somehow be rooted in basic memory processes, such as working memory capacity (e.g., Dougherty & Hunter, 2003; Dougherty & Sprenger, 2006) or selective retrieval (e.g., Johnson, Haubl, & Keinan, 2007; Ting & Wallsten, 2011). FTT explains such phenomena—the Allais paradox, the framing illusion, and hindsight bias, for instance—as by-products of reliance on gist memories (Reyna & Brainerd, 2011). Specifically, subjects store verbatim and gist traces in parallel and retrieve them in parallel, but they prefer to rely on the bottom-line meaning of problem information rather than the verbatim details that ensure logically coherent reasoning. In the gain frame of the classic Asian disease farming problem (Tversky & Kahneman, 1986), for example, the

categorical gist of the two options (A = 200 people will be saved; B = a 1/3 probability that 600 people will be saved and 2/3 probability that no people will be saved) is A = people are saved and B = people are saved and people die (Reyna & Brainerd, 1991). That gist obviously creates a preference for the certain option over the gamble, whereas processing the verbatim numerical details produces indifference (the options have the same expected value). In the loss frame, the categorical gist of the two options (C = 400 people will die; D = a 1/3 probability that nobody will die and 2/3 probability that 600 people will die) is C = people die and D = people live and people die. Now, the gist favors the gamble over the certain option, whereas the verbatim details of the numbers are still indifferent with respect to the two options. Thus, the tendency to rely on categorical gist foment the framing illusion—statistical preferences for certain options in the gain frame but gambles in the loss frame.

The task calibration principle posits that despite the baseline preference for simple gist on reasoning problems, the demands of the response format and the specificity of the cues that are provided in the problem information influence that preference (Corbin, Reyna, Weldon, & Brainerd, 2015). The general rule is that reliance on gist shrinks as response formats and problem cues become increasingly numerical and differentiated (Wolfe & Reyna, 2010). To illustrate, categorical gists, such as those described above, will work when the response format involves choices among discrete options, but not when it involves producing numerical estimates: If you like apartment A more than apartment B, that suffices to choose which you prefer to live in, but not to decide how much more you are willing to pay to live in apartment A (Reyna & Brainerd, 2011). FTT uses task calibration to explain phenomena such as preference reversals (e.g., Slovic & Lichtenstein, 1983), in which reasoning is inconsistent across response formats that differ in specificity. In the standard example, subjects prefer option A over option B when asked to choose between them, but they are willing to pay more for B when asked to specify the dollar amounts that they will pay for each. For instance, this reversal occurs when some subjects choose between options while other subjects specify exact dollar amounts with options such as A = a 3/4 chance of winning \$1.20 and a 1/4 chance of losing \$.10 versus B = a 1/4 chance of winning \$9.20 and a 3/4 chance of losing \$2.00. At the level of categorical gist, subjects treat 10 cents as nothing, so that the gist of A is “winning something or losing nothing” and the gist of B is “winning something or losing something,” favoring A over B for subjects who chose between them (Stone, Yates, & Parker, 1994). That does not suffice for subjects whose task is to specify how much to pay for each option, and now, they are willing to pay more for B than for A.

Beyond explaining existing effects, task calibration has predicted some surprising new ones, such as reversals in personal risk perception (Mills et al., 2008) and nonnumerical framing illusions (Reyna et al., 2011). In the former, which are response format effects, subjects judge the perceived risk of certain behaviors (e.g., unprotected sex) and the perceived frequency with which they engage in them, and negative correlations (the higher perceived risk, the lower the judged frequency) are typical (e.g., Halpern-Felsher, Biehl, Kropp, & Rubinstein, 2004). Mills et al. noted that subjects made categorical judgments (e.g., Are you likely to get pregnant?) in those studies, for which simple gists about personal

behavior suffice. They hypothesized that if graded numerical judgments were made instead (e.g., How likely are you are to get pregnant on a 0–100 scale?), gist reliance would decrease, and the sign of the correlation would change *from negative to positive* (the higher the perceived risk, the higher the judged frequency). Their reasoning was that suppressing gist reliance means that (a) a larger proportion of subjects' responses will be based on verbatim memories of specific instances of a risky behavior (e.g., instances of unprotected sex) and (b) that produces positive correlations because the perceived risk of such behavior will increase as the number of instances increases. Consistent with task calibration, Mills et al. found negative correlations with categorical judgments but positive correlations with graded numerical judgments.

The other example of task calibration predictions, nonnumerical framing illusions, involves the specificity of the cues in problem information. Recall that all of the options in framing problems provide subjects with detailed numerical information, such as a 1/3 probability that 600 people are saved and 2/3 probability that no people are saved. As we saw, FTT assumes that subjects rely more on the categorical gist of these options, which generates framing illusions, than on the verbatim numbers, which works against the illusion. This leads to the prediction that the framing illusion will increase if some or all of the numerical information is stripped out of the options. Consistent with task calibration, it is well established that, indeed, nonnumerical versions of framing problems (e.g., A = people are saved; B = people are saved and people die; C = people die; D = people live and people die) produce more robust illusions than standard numerical problems (e.g., Kühberger & Tanner, 2010; Reyna & Brainerd, 1991; Reyna et al., 2011).

Reducing Overdistribution With Confidence Ratings

This brings us back to episodic memory. If noncompensatory gist memories foment overdistribution illusions, an obvious strategy for reducing them is to exploit the calibration principle to decrease gist reliance. That certain test formats have this effect is a familiar idea in the false memory literature (Brainerd & Reyna, 2005). One example is recall versus recognition. The key finding there is that with NS items, for which gist traces are available but verbatim traces are not, false memory levels are consistently lower with recall (e.g., Seamon et al., 2002). In the experiments that we report, we compared overdistribution illusions in conditions in which subjects made the usual categorical item and source judgments to conditions in which they made ratings of item and source confidence. According to the calibration principle, such graded numerical judgments ought to reduce reliance on noncompensatory gist, relative to categorical judgments, reducing overdistribution illusions if such memories are an important factor in those illusions.

Naturally, confidence ratings have a long history in many domains of psychology (see Busey, Tunnicliff, Loftus, & Loftus, 2000). In memory research, they are widely used to plot the receiver operating characteristic (ROC) in recognition (e.g., Heathcote, 2003; Heathcote, Bora, & Freeman, 2010; Lampinen, Odegaard, Blackshear, & Toglia, 2005; Lampinen, Watkins, & Odegaard, 2006), and to separate the effects of recollection from those of familiarity (e.g., Malmberg, 2008; Parks, Murray, Elfman, & Yonelinas, 2011). In the applied sphere, witnesses to crimes normally provide confidence ratings for categorical judgments about

suspects during identification tests (e.g., Brewer & Wells, 2006; Jones, Williams, & Brewer, 2008; Juslin, Olsson, & Winman, 1996; Wells & Murray, 1984), with the ratings being presented as evidence at trial. In other fields of memory research, confidence ratings are used to measure subjects' prospective perceptions of the difficulty of learning different types of items (e.g., Finn & Metcalfe, 2008; Mueller, Dunlosky, Tauber, & Rhodes, 2014; Thiede & Dunlosky, 1999) and their retrospective perceptions of how well they have learned items that they cannot recall (e.g., Koriat & Levy-Sadot, 2001; Tekcan & Akturk, 2001; Thomas, Bulevich, & Dubois, 2011).

In the present research, confidence ratings function as a theoretically motivated procedure for reducing subjects' reliance on gist memories of the semantic content of list items. FTT assumes that retrieval of verbatim and gist memories is controlled by test cues (e.g., *sofa*, *trumpet*), but the degree to which subjects rely on gist to generate responses is influenced by response format. On episodic memory tasks, such as item or source recognition, verbatim memories trump gist because verbatim traces contain vivid, specific information about an item's presentation (Brainerd & Reyna, 2005). When verbatim traces are not retrieved, subjects may rely on nonspecific gist. Here, the calibration principle specifies that subjects are less likely to rely on gist when making graded numerical responses than when making categorical judgments, and hence, confidence ratings should reduce disjunction and conjunction illusions. Two existing lines of evidence that are congruent with the view that confidence ratings reduce gist reliance are (a) confidence rating data for different types of cues in false memory experiments and (b) correlations between confidence ratings and reports of realistic recollective phenomenology.

Concerning a, consider a standard false memory design in which subjects make categorical (old/new) recognition judgments about O, NS, and ND cues. Verbatim traces are only stored for O cues, so that hits are a mix of verbatim and gist processing and response bias, whereas false alarms to NS cues are a mix of gist processing and response bias. In certain experiments (e.g., Hauschildt, Peters, Jelinek, & Moritz, 2012), subjects provide confidence ratings following hits and false alarms. If it is true that such ratings deemphasize gist reliance, they should be lower for NS false alarms than for O hits because only gist memories support the former, whereas verbatim as well as gist memories support the latter. Lower confidence ratings for NS false alarms than for O hits is a ubiquitous result (Brainerd & Reyna, 2005), and some experiments by DeSoto and Roediger (2014) provide a recent illustration. Subjects studied word lists composed of blocks of exemplars of familiar categories (e.g., birds), with presented and unpresented exemplars serving as O and NS cues, respectively, on recognition tests. Over these experiments, mean confidence ratings (0–100 scale) following old judgments were 84 and 62 for O and NS cues, respectively.

Concerning b, if subjects rely less on gist memory when making confidence ratings than when making categorical judgments, increasing the proportion of O hits that are verbatim-based by default, another obvious prediction is that confidence ratings will correlate positively with reports of vivid, realistic study phase details, which are traditional phenomenological signals of reliance on verbatim memory (Lampinen et al., 2005). In particular, the phenomenology that subjects experience when they assign higher confidence values to old judgments about O cues ought to be richer

in vivid, realistic details than when they assign lower confidence values because higher ratings should reflect higher proportions of verbatim-based hits. Selmecky and Dobbins (2014) found that this was indeed the case when subjects provided extemporaneous descriptions of the phenomenologies that were associated with confidence ratings. Over two experiments, the highest confidence rating produced realistic phenomenological statements 46% of time, whereas lower confidence ratings produced such statements 13% of the time.

In the sections that follow, we report three experiments in which categorical judgments versus confidence ratings supplied the core manipulation. In Experiment 1, we compared the magnitude of disjunction illusions under the two response formats, using the same two-list procedure that originally identified these illusions in source monitoring (Brainerd et al., 2012). A key finding was that although illusions were present with categorical judgments at levels comparable to prior experiments, they were unreliable with confidence ratings. In Experiment 2, we again compared disjunction illusions under the two response formats, but this time, we used a three-list procedure that generates illusions that are far more robust. Now, although disjunction illusions were greatly reduced for confidence ratings relative to categorical judgments, they were statistically reliable in some conditions with confidence ratings. Finally, in Experiment 3, we compared conjunction illusions under the two response formats, again using a three-list procedure that produces especially robust illusions (Brainerd et al., 2014). The effects of response format were dramatic: Conjunction illusions were present in all conditions with categorical judgments but were unreliable in all of those same conditions with confidence ratings.

Experiment 1

This experiment paralleled the design of the original studies of disjunction illusions in source monitoring, with subjects studying two lists of words that were accompanied by distinctive contextual details (different fonts and background colors). The lists were followed by a recognition test on which three types of test cues (List 1 targets, List 2 targets, and distractors) were factorially crossed with two types of source probes ($L_1?$ and $L_2?$) plus an item probe (L_1 -or- $L_2?$). Half of the subjects responded to these probes by making categorical judgments (accept-reject), and half responded by rating their confidence that each probe was true. Although the general prediction, based on the calibration principle, is that confidence ratings should shrink disjunction illusions by decreasing reliance on noncompensatory gist memories, there was also a prediction about how it would affect performance on source versus item probes. Prior research shows that subjects are aware of differences in the inherent memory demands of different types of source tests and of source versus item tests, and that this influences the memory content that they rely on when responding to test probes (e.g., Hicks & Starns, 2006a, 2006b). This suggests that the subjects ought to be more susceptible to the gist-suppression effect of confidence ratings with source probes than with item probes, for the simple reason that relying on gist memories produces errors with the former (false alarms to incorrect source probes) but not with the latter.

As the central theoretical hypothesis is that gist reliance foments overdistribution and confidence judgments reduce such reliance, our experiments included two other manipulations that were in-

tended to produce quantitative differences in the accessibility of the two types of traces, one that elevates verbatim accessibility (list order) and one that elevates gist accessibility (word frequency). Concerning list order, FTT assumes that verbatim traces carry the contextual details that are necessary to make source discriminations, and various findings from misinformation experiments show that memory for such details is more sensitive to retroactive interference (Reyna & Lloyd, 1997). This seems to be a key basis for misinformation effects, wherein subjects falsely remember events that only occurred during the misinformation phase as having occurred earlier, when targets were presented (e.g., Lindsay & Johnson, 1989). In the present paradigm, this simply means that subjects will be more likely to access verbatim memories of contextual details for List 2 words than for List 1 words (Brainerd et al., 2012). Concerning word frequency, it is well established that recognition memory is better for low- than for high-frequency words (Hall, 1979). This appears to be a semantic-processing effect that occurs because the semantic content of low-frequency words receives more processing attention than that of high-frequency words (e.g., Estes & Maddox, 2002; Ozubko & Joordens, 2011). The implication for our research is that subjects will be more likely to access verbatim traces of List 2 targets than List 1 targets, and they will be more likely to access gist traces of low- than of high-frequency targets.

A key reason for including these manipulations was to provide further tests of the task calibration principle's analysis of confidence ratings. As mentioned, task calibration assumes that confidence ratings do not affect the accessibility of verbatim or gist traces on item and source tests, but simply affect subjects' tendency to base responses on the latter. In any given List $\times$ Frequency condition, subjects are assumed to retrieve the same verbatim and gist memories with both response formats but to rely less on gist with confidence ratings. Under that scenario, the list and frequency manipulations ought to have the same qualitative effects under the two response formats because the same verbatim and gist memories are being retrieved.

Method

Subjects. The subjects were 224 introductory psychology students who participated in the experiment to fulfill a course requirement. Individual subjects were randomly assigned to one of two response format conditions: categorical judgments or confidence ratings. The sample sizes for this experiment and for Experiments 2 and 3 were determined by computing estimated statistical power ($\beta > .8$), based on the results of prior experiments on disjunction and conjunction illusions (Brainerd et al., 2012, 2014). Very similar procedures were used in those experiments, except for the present confidence rating condition. Based on those experiments, the present sample sizes would allow even small disjunction illusion effects (e.g., $d = .20$) to be detected with $\beta > .8$.

Materials. A pool of 256 nouns was created, using the Kucera and Francis (1967) frequency norms and the Toglia and Battig (1978) semantic word norms. The pool consisted of two groups of words, each containing 128 items: (a) high-frequency nouns (HF; e.g., *industry*) and (b) low-frequency nouns (LF; e.g., *barnacle*). The mean frequency values (per million in printed text) were 72.4 (HF) and 2.0 (LF). The two study lists that were administered to individual subjects were generated by sampling (without replace-

ment) 48 words from the pool, 24 HF and 24 LF, such that mean word length did not differ for HF versus LF words. The cue words on the test lists that were administered to individual subjects consisted of (a) these 96 presented words, and (b) 96 distractors that were obtained by sampling a further 48 HF words and 48 LF words, subject to the same length constraint, from the words that remained in pool.

Each subject viewed two lists of words, with each list being accompanied by distinctive contextual details that were generated by presenting all of the words on that list in one of several different fonts (e.g., Algerian, Broadway, and Script) against one of several background colors (e.g., yellow, white, and pink). Thus, each list context was distinguished by a specific combination of temporal order, font, and background color details.

During the study phase, 108 words were presented, 54 on List 1 and 54 on List 2. Each list began and ended with a three-word buffer composed of filler words *that did not appear on the memory test*. The 48 focal words (24 HF and 24 LF) comprised the remainder of the list, and they were presented in random order. Thus, over the two lists, subjects were exposed to 48 HF words and 48 LF words, randomly intermixed on their individual lists. During the test phase, 192 probes were administered in random order. There were four types of cue words (HF and LF targets; HF and LF distractors), over which three types of test probes were factorially varied: presented on List 1 ($L_1?$), presented on List 2 ($L_2?$), and presented on List 1 or List 2 ($L_1\text{-or-}L_2?$). In the categorical judgment condition, the subjects were instructed to classify each probe as true or false according to whether it was a correct or incorrect description of the cue word. In the confidence rating condition, the subjects were instructed to rate each probe according to how likely it was that it was a true description of the cue word, using the scale: 0, 20, 40, 60, 80, and 100%. As is standard procedure with confidence ratings, subjects were told to use the entire scale—not just the extreme values.

Procedure. At the start of the experiment, each subject was told that two completely different lists of words would be presented, one after the other, followed by a memory test. The two lists were then presented on a computer screen, with individual words appearing at a 2-s rate, centered on the screen in 72-point bold type. The background color of the screen and the font in which words were printed were different for List 1 versus List 2. There was a 15-s pause between lists, and after List 2 had been presented, the subject received instructions for the memory test, which stated that some of the upcoming test cues would be list words and the rest would be new words (distractors). The three types of probe descriptions were defined and illustrated during the instructions, and examples with accompanying answers were provided, so that the subject understood how to respond to each. Subjects in the categorical judgment condition were instructed to classify a probe as true if they thought it was correct for the cue word, whereas subjects in the confidence rating condition were told to use the confidence scale to rate each probe with respect to how likely it was that it was true of the cue word. The instructions reiterated that the two lists did not overlap and that if subjects could clearly recollect the appearance of a word in one context, it could not have appeared in the other. The 192 test probes were then presented in random order, and the subject responded in a self-paced manner.

Results

The factorial structure of this experiment was 2 (response format: categorical judgment vs. confidence rating) $\times$ 2 (frequency: high vs. low) $\times$ 2 (list: 1 vs. 2) $\times$ 3 (probe type: $L_1?$ vs. $L_2?$ vs. L_1 -or- $L_2?$), with the disjunction illusion metric $DI = p(L_1) + p(L_2) - p(L_1$ -or- $L_2)$ supplying the dependent variable in an initial analysis of variance (ANOVA) and acceptance probabilities for the three types of probes supplying the dependent variable in a second ANOVA. Summary statistics for this experiment appear in Table 1, which displays raw and bias-corrected response probabilities for four variables—namely, the two types of source probes, $L_1?$ and $L_2?$, the item probe, L_1 -or- $L_2?$, and the DI metric. These probabilities are reported separately for the two response formats, and within each of those conditions, they are reported separately for the list-order and word-frequency manipulations. The probabilities that are reported for the confidence scale are simply the averages of the percentage rankings in each condition, after transforming percentages into probabilities; that is, averages that were computed after transforming 0, 20, 40, 60, 80, and 100% to 0, .2, .4, .6, .8, and 1, respectively.

The bias-corrected data for both response formats were generated by the two-high-threshold (2HT) method. There is a well-known measurement theory for applying this method to categorical judgment data, for both item memory tests (see Snodgrass & Corwin, 1988) and source memory tests (see Meiser & Broder, 2002). For item tests, 2HT assumes that each target cue induces one of two memory states, old or uncertain, and distractor cues induce one of two memory states, new or uncertain, with cues in the old, new, and uncertain states being judged to be old with probabilities 1, 0, and $1 > a > 0$, respectively. For source probes such as ours, 2HT assumes that when target cues induce the old

state, they induce one of two source states, correct or uncertain, with the corresponding cues being accepted as correct with probabilities 1 and $1 > g > 0$, respectively. When target cues induce the uncertain state, they are accepted as correct with probability $1 > b > 0$. For distractor cues that induce the new and uncertain states, source probes are accepted as correct with probabilities 0 and $1 > b > 0$, respectively. As applications of 2HT give good empirical fits (e.g., Snodgrass & Corwin, 1988), it has been widely applied in studies of item and source memory—including prior research on overdistribution illusions (Brainerd et al., 2012, 2014).

Recently, Broder et al. (2013) developed a measurement theory that extends 2HT from categorical judgments to confidence ratings, and they showed that this extension delivered good fits to confidence rating data from item recognition experiments. Extended 2HT assumes that the item/source memory states that were just mentioned for categorical judgments also apply to confidence ratings. For our experimental design, 2HT specifies that those states are mapped with confidence ratings as follows. First, whenever the state for a target cue is uncertain on item or source tests, subjects guess a confidence rating, using the entire the scale. Second, whenever the state for a target cue is old on an item test or correct on a source test, subjects select a confidence rating from above the midpoint of the scale with probability 1. Third, whenever the item memory state for a distractor cue is uncertain, subjects guess a confidence rating, using the entire rating scale, on both item and source tests. Fourth, whenever the item memory state for a distractor cue is new, subjects select a confidence rating from above the midpoint of the scale with probability 1, on both item and source tests. 2HT imposes no further assumptions about how item/source states are mapped with confidence ratings, and in particular, it does not restrict the distributions of confidence ratings

Table 1
Raw and Bias-Corrected (in Parentheses) Probabilities of Accepting Probes as True in Experiment 1

Probe type	Categorical judgment		Confidence rating	
	High frequency	Low frequency	High frequency	Low frequency
List 1				
Targets				
$L_1?$	.58 (.25)	.56 (.34)	.51 (.19)	.52 (.27)
$L_2?$	.40 (.12)	.57 (.35)	.44 (.09)	.5 (.28)
L_1 -or- $L_2?$	.62 (.27)	.76 (.52)	.67 (.27)	.77 (.49)
DI	.36 (.10)	.37 (.17)	.28 (.01)	.25 (−.06)
Distractors				
$L_1?$	.33	.22	.32	.27
$L_2?$	.28	.22	.35	.28
L_1 -or- $L_2?$	.35	.24	.40	.30
List 2				
Targets				
$L_1?$	.41 (.08)	.40 (.18)	.35 (.03)	.34 (.07)
$L_2?$	.52 (.24)	.63 (.41)	.48 (.13)	.59 (.31)
L_1 -or- $L_2?$	.69 (.34)	.60 (.36)	.66 (.20)	.61 (.33)
DI	.24 (−.02)	.43 (.23)	.17 (−.04)	.32 (.05)
Distractors				
$L_1?$	.33	.22	.32	.27
$L_2?$	.28	.22	.35	.28
L_1 -or- $L_2?$	.35	.24	.40	.30

Note. $L_1?$ = the cue was presented on List 1, $L_2?$ = the cue was presented on List 2, and L_1 -or- $L_2?$ = the cue was presented on List 1 or List 2. DI = $p(L_1?) + p(L_2?) - p(L_1$ -or- $L_2?)$, which is the disjunction illusion index.

within those states. Owing to this extension of 2HT to confidence ratings, the same method of bias correction can be applied to both categorical judgments and confidence ratings in the experiments that we report in this article.

We computed two ANOVAs to answer the questions of principal interest. The first used the bias-corrected scores for the DI metric as the dependent variable, and the second used the bias-corrected scores for the three types of probes (L_1 , L_2 , and L_1 -or- L_2) as the dependent variable. To answer the central question of whether the disjunction illusions that are observed with categorical judgments are ameliorated by confidence ratings, we computed a 2 (response format) $\times$ 2 (word frequency) $\times$ 2 (list order) ANOVA of the data for the DI metric. Second, to determine how the effects of response format on disjunction illusions arise from differential effects on the three probes, we computed a 2 (response format) $\times$ 2 (word frequency) $\times$ 2 (list order) $\times$ 3 (probe type) ANOVA of the data for the individual probes. We report the two sets of results separately.

Response format effects on disjunction illusions. Recall that disjunction illusions are circumstances in which the sum of $p(L_1)$ and $p(L_2)$ is subadditive with respect to $p(L_1$ -or- $L_2)$. Without the aid of any statistical analysis, a glance at Table 1 reveals that the response format manipulation had the predicted effect: On the one hand, DI was positive in three of the four categorical judgment conditions ($M = .13$) and within the range of the DI values in prior two-list experiments (cf. Brainerd et al., 2012), but on the other hand, this metric only had a small positive value in one of the corresponding conditions for confidence ratings and its mean value was slightly negative (but not reliably different than 0). Because the predicted value of DI is 0 for the null situation in which there is no overdistribution, the appropriate statistical test to determine whether an observed value of DI exhibits reliable subadditivity ($p < .05$) is a one-sample t test that compares that value to a predicted value of zero (Brainerd et al., 2012). For categorical judgments, three of the four tests were reliable: $t(118) = 3.92$, for List 1/HF; $t(118) = 5.34$, for List 1/LF, and $t(118) = 7.24$, for List 2/LF. Thus, categorical judgments produced disjunction illusions under conditions resembling those that have previously produced them, but under those same conditions, disjunction illusions were not reliable with confidence ratings.

The results of the ANOVA of DI values were as follows. (All reported effects were reliable at or beyond the .05 level in this experiment and also in Experiments 2 and 3.) First, it produced main effects for response format, $F(1, 222) = 70.62$, $MSE = 0.13$, $\eta^2 = .24$, word frequency, $F(1, 222) = 29.93$, $MSE = 0.11$, $\eta^2 = .12$, and list, $F(1, 222) = 6.78$, $MSE = 0.06$, $\eta^2 = .03$. As can be seen in Table 1, DI values were lower for confidence ratings than for categorical judgments, for HF targets than for LF targets, and for List 2 than for List 1 targets. Second, response format did not interact with either the list-order or word-frequency manipulations, so that the DI metric reacted in the same way to list order and word frequency under both response formats. This is an instructive result theoretically. It suggests that there must have been strong overlap in the memory content that subjects retrieved in the two response format conditions, as the task calibration principle assumes.

Third, there was an important Word Frequency $\times$ List interaction, $F(1, 222) = 39.14$, $MSE = 0.06$, $\eta^2 = .15$, which qualified the word frequency main effect. As can be seen in Table 1, the mean value of DI for LF was subadditive for both lists, but the

corresponding mean value for HF was only subadditive for List 1. This pattern is consistent with prior two-list experiments in which DI has had larger mean values for List 1 than for List 2, which is consistent with the view that verbatim memories are harder to access for List 1, owing to their sensitivity to retroactive interference (Brainerd et al., 2012, 2014).

Response format effects on source and item probes. The results of the 2 (response format) $\times$ 2 (word frequency) $\times$ 2 (list order) $\times$ 3 (probe type) ANOVA were as follows. First, there were main effects for response format, $F(1, 222) = 13.94$, $MSE = 0.23$, $\eta^2 = .06$, word frequency, $F(1, 222) = 288.19$, $MSE = 0.05$, $\eta^2 = .57$, list, $F(1, 222) = 51.17$, $MSE = 0.03$, $\eta^2 = .19$, and probe type, $F(2, 444) = 180.23$, $MSE = 0.05$, $\eta^2 = .45$. As can be seen in Table 1, response probabilities were lower for confidence ratings than for categorical judgments, for HF targets than for LF targets, and for List 2 than for List 1 probes. In addition, response probabilities were highest for $p(L_1$ -or- $L_2)$, lowest for $p(L_1)$, and intermediate for $p(L_2)$, with paired-samples t tests showing that each pairwise difference was reliable.

Second, there was an important Response Format $\times$ Probe Type interaction, $F(2, 444) = 15.60$, $MSE = 0.05$, $\eta^2 = .07$, which qualified the probe main effect, and an important List $\times$ Frequency $\times$ Probe Type interaction, $F(2, 444) = 38.25$, $MSE = 0.02$, $\eta^2 = .15$, which qualified the list and frequency main effects. The Response Format $\times$ Probe interaction bears on a prediction that we considered earlier—namely, that if subjects are sensitive to the differing memory demands of source versus item probes, the effect of switching to confidence ratings should be more marked for the former. Consistent with that notion, inspection of Table 1 confirms that the mean reduction in response probability for confidence ratings versus categorical judgments was greater for the source probes than it was for the item probe. Indeed, post hoc tests (Tukey's honest significant difference [HSD]) revealed that whereas the reductions for source probes were statistically reliable, they were not for the item probe. Returning to the List $\times$ Frequency $\times$ Probe Type interaction, because this interaction did not involve specific pairwise predictions, we teased it apart with post hoc tests. Those tests showed that the interaction was because of the fact that, naturally, mean response probabilities for the two source probes reversed as function of list because L_1 ? was correct and L_2 ? was incorrect for List 1 but conversely for List 2. In addition, this reversal was more marked for HF cues than for LF cues.

Summary. As predicted by the calibration principle, confidence ratings lowered the DI metric relative to categorical judgments. Indeed, disjunction illusions were no longer reliable, although with categorical judgments, they were reliable at levels that were comparable to prior two-list experiments. Even in experimental conditions that elevate disjunction illusions with categorical judgments (List 1 words, LF words), the DI metric was still not reliably >0 . Thus, the suppressive effects of switching to confidence ratings were quite marked.

The way that confidence ratings reduced disjunction illusions was revealed by examining how the response format manipulation affected acceptance probabilities for source versus item probes. Because the disjunction illusion index is $p(L_1) + p(L_2) - p(L_1$ -or- $L_2)$, any manipulation that lowers $p(L_1)$ and/or $p(L_2)$ more than it lowers $p(L_1$ -or- $L_2)$ necessarily decreases the DI metric. Here, we saw that the response format manipulation low-

ered both $p(L_1)$ and $p(L_2)$, but did not have reliable effects on $p(L_1\text{-or-}L_2)$.

Finally, the results for the word frequency and list order manipulations were consistent with the hypothesis that although subjects are less inclined to rely on noncompensatory gist with confidence ratings, the same the types of traces are being retrieved with the both response formats. This is because these manipulations affected acceptance probabilities in the same way with both formats.

Experiment 2

The aim of this experiment was to conduct a much stronger test of the ability of response format to reduce disjunction illusions. In prior research, the values of the DI metric in three-list designs have been three times larger, on average, than those in two-list designs. That pattern was originally predicted on theoretical grounds, based on differences in the number of possible noncompensatory memory relations with two versus three contexts (see Brainerd et al., 2012). In two-list designs, such as Experiment 1, disjunction illusions are by-products of noncompensatory relations between one pair of contexts (List 1 vs. List 2). In three-list designs, they are by-products of noncompensatory relations between three pairs of contexts (List 1 vs. List 2, List 1 vs. List 3, and List 2 vs. List 3). More noncompensatory relations ought to translate into stronger illusions, and they have in prior experiments (Brainerd et al., 2012, 2104).

Therefore, our aim was to determine whether confidence ratings would also suppress the far more robust illusions that are observed with three contexts. Similar to Experiment 1, the subjects in this experiment studied lists of words that were accompanied by distinctive font and color details followed by a recognition test on which four types of test cues (List 1 targets, List 2 targets, List 3 targets, and distractors) were factorially crossed with three types of source probes ($L_1?$, $L_2?$, and $L_3?$) and an item probe ($L_1\text{-or-}L_2\text{-or-}L_3?$). As in Experiment 1, the subjects in one condition responded by making categorical judgments, and the subjects in the other condition responded by rating their confidence that the probes were true. We saw in Experiment 1 that, as predicted by the calibration principle, this response format manipulation interacted with probe type, affecting response probabilities more on source probes than on item probes. This pattern was also expected in Experiment 2, for the same theoretical reasons (i.e., subjects' awareness that the memory demands of source probes are more exacting than those of item probes).

Finally, we included the same list-order and word frequency manipulations as in the first experiment and for the same reasons. Disjunction illusions have previously been found to be more marked for List 1 and for LF words, and the same was true in Experiment 1 with categorical judgments. In this experiment, there were no qualitative differences in how the DI metric responded to these manipulations with confidence ratings versus categorical judgments, even though this metric was not reliably >0 with confidence ratings. That pattern is consistent with the view that at the level of memory processes, the same types of traces are retrieved with two response formats, although subjects are less inclined to rely on noncompensatory gist when making confidence ratings. Therefore, we included these same manipulations in the present experiment to check whether DI continued to react similarly to them under the two response formats.

Method

Subjects. The subjects were 232 introductory psychology students who participated in the experiment to fulfill a course requirement. Individual subjects were randomly assigned to one of two response format conditions: categorical judgments or confidence ratings.

Materials and procedure. Methodologically, this experiment paralleled Experiment 1, but there were three key design changes. The most critical one was that there were now three presentation contexts—List 1, List 2, and List 3. As before, each list was distinguished by a unique combination of screen background color, letter font, and temporal order cues. The total number of target words that was presented over the three lists (108) was the same as the number that had been presented over the two lists in Experiment 1, and hence, the only increase in memory load was the increase in presentation contexts from two to three. Each list consisted of 36 words, an opening buffer of two words, 32 focal words (16 HF words and 16 LF words), and a closing buffer of two words. The other two design changes involved the source and item probes on the memory test. There was now a third source probe (presented on List 3? [$L_3?$]), and the item probe involved three contexts rather than two (presented on List 1 or List 2 or List 3? [$L_1\text{-or-}L_2\text{-or-}L_3?$]). Thus, although the 192 cue words on the memory test were the same types as in Experiment 1 (i.e., 96 targets and 96 distractors, half HF and half LF), four memory probes rather than three were factorially varied over these cues (i.e., 48 cues per probe type rather than 64).

Finally, the instructions and testing procedures for the two response formats were the same as in Experiment 1. Briefly, the subjects in the categorical judgment condition made true-false judgments about the correctness of individual probes, whereas the subjects in the confidence rating condition rated their confidence that each probe was true of the cue word, using the scale: 0, 20, 40, 60, 80, and 100%.

Results

The factorial structure of this experiment was 2 (response format: categorical judgment vs. confidence rating) $\times$ 2 (frequency: high vs. low) $\times$ 3 (list: 1 vs. 2 vs. 3) $\times$ 4 (probe type: $L_1?$ vs. $L_2?$ vs. $L_3?$ vs. $L_1\text{-or-}L_2\text{-or-}L_3?$), with the disjunction illusion metric $DI = p(L_1) + p(L_2) + p(L_3) - p(L_1\text{-or-}L_2\text{-or-}L_3)$ supplying the dependent variable in an initial ANOVA and target acceptance probabilities for the four types of probes supplying the dependent variable in a second ANOVA. Summary statistics for this experiment appear in Table 2, which displays the raw and 2HT bias-corrected response probabilities for the three types of source probes, the item probe, and the DI metric. These probabilities are reported separately for the two response formats, and within each of those conditions, they are reported separately for the list-order and word-frequency manipulations.

Before reporting the ANOVAs, two important findings are apparent from the means in Table 2. First, as in prior studies of disjunction illusions, the values of DI in the categorical judgment condition were far larger in this experiment than they were in Experiment 1 (grand $M_s = .43$ vs. $.13$). When the six values of DI for the categorical judgment condition were tested for statistical significance, all were reliably greater than zero, with values of the $t(116)$ statistic ranging from 5.34 to 17.54 (all $p_s < .0001$).

Table 2
Raw and Bias-Corrected (in Parentheses) Probabilities of Accepting Probes as True in Experiment 2

Probe type	Categorical judgment		Confidence rating	
	High frequency	Low frequency	High frequency	Low frequency
List 1				
Targets				
L ₁ ?	.63 (.38)	.63 (.45)	.54 (.23)	.52 (.27)
L ₂ ?	.38 (.14)	.62 (.46)	.37 (.06)	.43 (.19)
L ₃ ?	.58 (.33)	.56 (.41)	.44 (.13)	.43 (.18)
L ₁ -or-L ₂ -or-L ₃ ?	.60 (.30)	.75 (.56)	.69 (.25)	.80 (.47)
DI	.99 (.55)	1.07 (.76)	.66 (.21)	.58 (.17)
Distractors				
L ₁ ?	.25	.18	.31	.25
L ₂ ?	.24	.16	.31	.24
L ₃ ?	.25	.15	.31	.25
L ₁ -or-L ₂ -or-L ₃ ?	.30	.19	.44	.33
List 2				
Targets				
L ₁ ?	.34 (.09)	.41 (.23)	.32 (.01)	.32 (.07)
L ₂ ?	.49 (.25)	.57 (.41)	.46 (.15)	.48 (.24)
L ₃ ?	.42 (.17)	.44 (.29)	.39 (.08)	.38 (.13)
L ₁ -or-L ₂ -or-L ₃ ?	.72 (.42)	.68 (.48)	.73 (.29)	.76 (.43)
DI	.53 (.09)	.74 (.45)	.44 (−.05)	.42 (.01)
Distractors				
L ₁ ?	.25	.18	.31	.25
L ₂ ?	.24	.16	.31	.24
L ₃ ?	.25	.15	.31	.25
L ₁ -or-L ₂ -or-L ₃ ?	.30	.19	.44	.33
List 3				
Targets				
L ₁ ?	.35 (.10)	.41 (.23)	.33 (.02)	.34 (.09)
L ₂ ?	.54 (.30)	.57 (.41)	.43 (.12)	.47 (.23)
L ₃ ?	.48 (.23)	.58 (.43)	.47 (.16)	.47 (.22)
L ₁ -or-L ₂ -or-L ₃ ?	.67 (.37)	.71 (.52)	.70 (.26)	.72 (.39)
DI	.70 (.54)	.85 (.55)	.53 (.04)	.56 (.15)
Distractors				
L ₁ ?	.25	.18	.31	.25
L ₂ ?	.24	.16	.31	.24
L ₃ ?	.25	.15	.31	.25
L ₁ -or-L ₂ -or-L ₃ ?	.30	.19	.44	.33

Note. L₁? = the cue was presented on List 1, L₂? = the cue was presented on List 2, L₃? = the cue was presented on List 3, and L₁-or-L₂-or-L₃? = the cue was presented on List 1 or List 2 or List 3. DI = $p(L_1?) + p(L_2?) + p(L_3?) - p(L_1\text{-or-}L_2\text{-or-}L_3?)$, which is the disjunction illusion index.

Second, despite the fact that disjunction illusions were far more robust than before in the categorical judgment condition, the grand mean of the six values of the DI metric for confidence ratings was zero. Nevertheless, inspection of the individual means suggests that reliable disjunction illusions were present in some of the List Order $\times$ Word Frequency cells. In particular, they were present in the cells that produced the largest values of DI in the categorical judgment condition in this experiment and in prior experiments, which are those for List 1. The mean value of the two List 1 cells in the confidence rating condition was .15, which is slightly larger than the corresponding mean value for the *categorical judgment* condition in Experiment 1, and the mean value for the LF cell for List 3 was also substantially greater than zero. When those three values were tested for statistical significance, all were reliably greater than zero: $t(115) = 3.43$ (List 1, HF), 4.36 (List 1, LF), and 2.40 (List 3, LF). Thus, when much stronger disjunction illusions were induced by presenting targets in three encoding contexts,

these illusions could be detected with confidence ratings in some cells of the design.

Two ANOVAs were conducted. The first provided evidence bearing on the central question of whether the disjunction illusions that are observed with categorical judgments are ameliorated by confidence ratings. This was a 2 (response format) $\times$ 2 (word frequency) $\times$ 3 (list order) ANOVA of the DI data. Second, to pinpoint how the effects of response format on DI arise from its differential effects on the four types of probes, we computed a 2 (response format) $\times$ 2 (word frequency) $\times$ 3 (list order) $\times$ 4 (probe type) ANOVA of the data for the individual probes.

Response format effects on disjunction illusions. The results of the ANOVA of DI values were as follows. First, it yielded main effects for response format, $F(1, 227) = 139.30$, $MSE = 0.47$, $\eta^2 = .38$, word frequency, $F(1, 227) = 57.50$, $MSE = 0.27$, $\eta^2 = .20$, and list, $F(2, 454) = 92.03$, $MSE = 0.31$, $\eta^2 = .27$. Concerning the response format, as already noted, DI values were

much higher for categorical judgments than for confidence ratings (grand $M_s = .43$ and 0). With respect to the word frequency and list effects, DI values were higher for LF targets than for HF targets (grand $M_s = .32$ and $.11$), and the order of the DI values for the three lists was List 1 > List 2 > List 3 (grand $M_s = .39$, $.19$, and $.07$; all pairwise differences were reliable by paired-sample t tests). That ordering is consistent with prior experiments and with the theoretical notion that verbatim traces become harder to access as retroactive interference mounts (Brainerd et al., 2012).

Next, as in Experiment 1, there was a Response Format $\times$ Word Frequency $\times$ List interaction, $F(2, 454) = 32.95$, $MSE = 0.21$, $\eta^2 = .13$. Post hoc analysis produced three effects. First and most important, as in Experiment 1, the *qualitative* effects of word frequency and list order on the DI metric were the same for categorical judgments and confidence ratings: The ordering of DI values by list was the same in both conditions, and DI values were higher for LF than for HF words in both conditions. The fact that there were no qualitative differences in how disjunction illusions were influenced by these manipulations again suggests that the same memory content was being retrieved under both response formats, although subjects were less inclined to rely on noncompensatory gist when making confidence ratings.

Second, the absolute magnitude of the word-frequency effect was larger with categorical judgments than with confidence ratings: The average difference in DI for HF versus LF words was $.27$ with categorical judgments versus $.15$ with confidence ratings. Third, the absolute magnitude of the list-order effect, on the other hand, was larger with confidence ratings than with categorical judgments: The average difference in DI between specific pairs of lists was $.33$ with confidence ratings and $.26$ with categorical judgments. Notice that, together, these two effects are consistent with the calibration hypothesis that confidence ratings reduce reliance on noncompensatory gist. As previously mentioned, gist memories should be more highly accessible with LF than with HF words, and hence the LF-HF difference in the DI metric ought to be smaller with confidence ratings than with categorical judgments, if confidence ratings target gist reliance. Also as previously mentioned, the list-order effect is assumed to be primarily a verbatim effect, with traces of source details being less accessible for earlier than for later lists. If confidence ratings target gist reliance, manipulations that primarily affect verbatim memory ought to have larger effects with confidence ratings than with categorical judgments because verbatim memory makes proportionately larger contributions to performance with confidence ratings.

Readers will recall that neither of these two effects—larger word-frequency effects for categorical judgments but larger list effects with confidence ratings—were observed in Experiment 1. This may be because of a difference in statistical power caused by the fact that disjunction illusions were far more robust in Experiment 2 than in Experiment 1.

Response format effects on source and item probes. The results of the 2 (response format) $\times$ 2 (word frequency) $\times$ 3 (list order) $\times$ 4 (probe type) ANOVA were as follows. First, there were main effects for response format, $F(1, 227) = 59.16$, $MSE = 0.39$, $\eta^2 = .23$, word frequency, $F(1, 227) = 287.63$, $MSE = 0.09$, $\eta^2 = .56$, list, $F(2, 454) = 31.06$, $MSE = 0.05$, $\eta^2 = .13$, and probe type, $F(3, 691) = 185.81$, $MSE = 0.08$, $\eta^2 = .45$. As can be seen in Table 2, acceptance probabilities were lower for confidence

ratings than for categorical judgments, and for HF targets than for LF targets. The order of acceptance probabilities for the three lists was List 1 > List 2 > List 3, with paired-samples t tests showing that probabilities were reliably higher for List 1 than for the other two lists, but List 2 and List 3 did not differ reliably. With respect to the probe type main effect, the order of response probabilities was $L_1\text{-or-}L_2\text{-or-}L_3? > L_2? > L_3? > L_1?$. Paired-samples t tests revealed that probabilities (a) were higher for the item probe than for any of the three source probes and (b) did not differ reliably for the three source probes.

Second, as in Experiment 1, there was an important Response Format $\times$ Probe Type interaction, $F(3, 691) = 12.96$, $MSE = 0.25$, $\eta^2 = .05$, and an important List $\times$ Frequency $\times$ Probe Type interaction, $F(6, 1362) = 10.55$, $MSE = 0.04$, $\eta^2 = .04$. The reasons were also the same as in Experiment 1. Concerning the Response Format $\times$ Probe Type interaction, the effects of switching from categorical judgments to confidence ratings were more marked for the source probes than for the item probe, which confirms our earlier prediction that this should happen because subjects are sensitive to the differing memory demands of source versus item probes. A new finding that emerged from the analysis of this interaction is that unlike Experiment 1, confidence ratings produced a reliable reduction in response probabilities for item probes as well as source probes. Concerning the List $\times$ Frequency $\times$ Probe Type interaction, it was again because of the fact that (a) naturally, mean acceptance probabilities for the three source probes reversed as function of list because $L_1?$ was correct for List 1, $L_2?$ was correct for List 2, and $L_3?$ was correct for List 3, and (b) these reversals in source probe acceptance probabilities as a function of which probe was correct were more marked for HF targets than for LF targets.

Summary. The most informative outcome is that although the three-list design yielded more than a threefold increase in the strength of disjunction illusions with categorical judgments, confidence ratings continued to suppress those illusions. Indeed, once again, the mean value of DI in the confidence rating condition over all cells of the design was not reliably >0 . Nevertheless, there were three cells in which DI was reliably >0 zero with confidence ratings, and crucially, those were also the cells in which categorical judgments produced the highest values of DI. Thus, the most reasonable conclusion is that confidence ratings produce small disjunction illusions in conditions in which those illusions are particularly robust; some residual noncompensation remains with confidence ratings, under favorable conditions.

A second informative finding also concurred with the results of Experiment 1—namely, the manner in which confidence ratings reduced disjunction illusions, as revealed by the Response Format $\times$ Probe Type interaction. Because the DI metric is $p(L_1) + p(L_2) + p(L_3) - p(L_1\text{-or-}L_2\text{-or-}L_3)$, any manipulation that lowers $p(L_1)$ and/or $p(L_2)$ and/or $p(L_3)$ more than it lowers $p(L_1\text{-or-}L_2\text{-or-}L_3)$ must decrease DI. Analysis of the interaction showed that confidence ratings reduced response probabilities for the three source probes by roughly equal amounts ($.16$ on average) and reduced them more than for the item probe ($.08$ on average). The latter reduction was reliable, unlike Experiment 1. The indicated conclusion, then, is that confidence ratings reduce disjunction illusions because their memory effects are more pronounced for source than for item probes, but they affect item probes, too.

Experiment 3

So far, confidence ratings have dramatically reduced one of the two indexes of overdistribution: Using the DI metric, confidence ratings produced no reliable evidence of disjunction illusions in Experiment 1, and only modest evidence of them in Experiment 2. According to that same metric, the effects of confidence ratings were sensitive to the greater memorial precision that source probes demand because they suppressed target acceptance probabilities more for source probes than for item probes.

We turn now to the other index of overdistribution, conjunction illusions. For the sake of comparability with the results reported up to this point, Experiment 3 preserved major design features of the disjunction illusion experiments, the chief alteration being that the item probes of earlier experiments were replaced with conjunctive source probes of the form L_1 -and- L_2 -and- L_3 ?, with $p(L_1$ -and- L_2 -and- L_3 ?) being the conjunction illusion metric. In prior experiments in which the target cues on memory tests had *never* been presented on more than one list, this metric was nevertheless substantially > 0 (.20 on average in Brainerd et al., 2014). Further, in some conditions $p(L_1$ -and- L_2 -and- L_3) was reliably greater than $p(L_1)$, $p(L_2)$, or $p(L_3)$, which is quite counterintuitive because it cannot be more probable that a target appeared on all three lists than that it appeared on any one of them.

There was reason to expect, based on the first two experiments, that confidence ratings would dramatically affect conjunction illusions. Note in that respect that L_1 -and- L_2 -and- L_3 ? is a source probe rather than an item probe, and in the first two experiments the suppressive effects of confidence ratings were most pronounced for source probes. We included the same list-order and word-frequency manipulations as in the disjunction experiments because in prior conjunction illusion experiments, $p(L_1$ -and- L_2 -and- L_3) was also affected by them. By including those manipulations, we were able to determine whether the data in the categorical judgment condition behaved as they have in prior experiments and to determine whether confidence ratings were similarly affected by list order and word frequency. Recall that in first two experiments, these manipulations had the same qualitative effects on the DI metric in the categorical judgment and confidence rating conditions, even though confidence ratings dramatically suppressed this metric.

In Experiments 1 and 2, we tested the prediction that response format would affect acceptance probabilities more for source than for item probes because subjects (a) are aware of differences in the memory demands of these probes and (b) this affects how they process retrieved memory content (e.g., Hicks & Starns, 2006a, 2006b). In the present experiment, this leads to a new prediction about how the response format manipulation should affect response probabilities on different types of probes. Although all of the test probes in this experiment were source probes, the memory demands of the conjunctive probe are obviously greater than those of standard probes. To accept a conjunctive probe for a given target cue, subjects should, in theory, retrieve some of the unique contextual details for each list. To accept any of the standard probes (say, L_1 ?), however, subjects need only retrieve some of the unique contextual details for that one list. Assuming that subjects are sensitive to such differences in memory demands, the response

format manipulation ought to affect target acceptance probabilities more for conjunctive probes than for standard ones.

Method

Subjects. The subjects were 232 introductory psychology students who participated in the experiment to fulfill a course requirement. Individual subjects were randomly assigned to one of two response format conditions: categorical judgments or confidence ratings.

Materials and procedure. The word pool was the same as in the first two experiments. The details of how the study lists for individual subjects were constructed and presented were the same as Experiment 2, except for one change. In Experiment 2, none of the targets on a given list was repeated on either of the other lists. Brainerd et al. (2014) pointed out that during list presentation, some subjects might notice this and adopt a metacognitive rejection strategy on conjunctive probes during the test phase, rather than following instructions to base their responses on information that is retrieved from memory. That, in turn, would produce underestimates of $p(L_1$ -and- L_2 -and- L_3). To control for this possibility, Brainerd et al. lengthened each list slightly to include eight filler items that were presented in random positions on each of the three study lists but that did not appear on any of test lists (i.e., no test cue had ever appeared on more than one list). We also implemented this modification in the present experiment.

The details of the test phase—the instructions and the construction of the test lists for individual subjects—were the same as in Experiment 2, except for three changes. First, remember that the instructions in the first two experiments stated that no cue on the test list had appeared on more than one study list. Those statements were removed from the present instructions. Second, the instructions in Experiment 2 provided examples of item probes, whereas in the present experiment, they were replaced with examples of conjunctive source probes. Third, on the test list, all of the L_1 -or- L_2 -or- L_3 ? probes in Experiment 2 were replaced with L_1 -and- L_2 -and- L_3 ? probes.

Results

The factorial structure of this experiment was 2 (response format: categorical judgment vs. confidence rating) $\times$ 2 (frequency: high vs. low) $\times$ 3 (list: 1 vs. 2 vs. 3) $\times$ 4 (probe type: L_1 ? vs. L_2 ? vs. L_3 ? vs. L_1 -and- L_2 -and- L_3 ?), with the conjunction illusion metric $p(L_1$ -and- L_2 -and- L_3) supplying the dependent variable in an initial ANOVA and acceptance probabilities for the four types of probes supplying the dependent variable in a second ANOVA. Summary statistics for this experiment appear in Table 3, which displays raw and 2HT bias-corrected response probabilities for the three types of standard source probes and the conjunctive source probe. As in prior experiments, these probabilities are reported separately for the two response formats, and for the word-frequency and list-order manipulations.

Before reporting the ANOVA, two instructive findings can be extracted from the means in Table 3. First, consistent with earlier studies of disjunction illusions, the values of $p(L_1$ -and- L_2 -and- L_3) were noticeably > 0 in all cells with categorical judgments. Thus, there was robust evidence of conjunction illusions; subjects routinely judged targets to have been presented on all three

Table 3
Raw and Bias-Corrected (in Parentheses) Probabilities of Accepting Probes as True in Experiment 3

Probe type	Categorical judgment		Confidence rating	
	High frequency	Low frequency	High frequency	Low frequency
List 1				
Targets				
L ₁ ?	.61 (.28)	.60 (.32)	.57 (.22)	.54 (.26)
L ₂ ?	.57 (.20)	.58 (.24)	.46 (.05)	.52 (.18)
L ₃ ?	.54 (.26)	.58 (.37)	.50 (.23)	.45 (.44)
L ₁ -and-L ₂ -and-L ₃ ?	.36 (.14)	.42 (.26)	.27 (.07)	.21 (.01)
Distractors				
L ₁ ?	.33	.28	.35	.28
L ₂ ?	.37	.34	.41	.34
L ₃ ?	.28	.21	.27	.21
L ₁ -and-L ₂ -and-L ₃ ?	.22	.16	.34	.22
List 2				
Targets				
L ₁ ?	.41 (.08)	.46 (.18)	.41 (.06)	.35 (.07)
L ₂ ?	.46 (.09)	.53 (.19)	.49 (.08)	.54 (.20)
L ₃ ?	.53 (.25)	.54 (.33)	.47 (.20)	.44 (.23)
L ₁ -and-L ₂ -and-L ₃ ?	.38 (.10)	.34 (.18)	.26 (−.08)	.22 (.00)
Distractors				
L ₁ ?	.33	.28	.35	.28
L ₂ ?	.37	.34	.41	.34
L ₃ ?	.28	.21	.27	.21
L ₁ -and-L ₂ -and-L ₃ ?	.22	.16	.34	.22
List 3				
Targets				
L ₁ ?	.51 (.14)	.30 (.02)	.38 (.03)	.40 (.12)
L ₂ ?	.54 (.17)	.48 (.14)	.49(.08)	.37 (.03)
L ₃ ?	.52 (.24)	.47 (.26)	.53 (.26)	.50 (.29)
L ₁ -and-L ₂ -and-L ₃ ?	.49 (.25)	.34 (.18)	.29 (−.05)	.22 (.00)
Distractors				
L ₁ ?	.33	.28	.35	.28
L ₂ ?	.37	.34	.41	.34
L ₃ ?	.28	.21	.27	.21
L ₁ -and-L ₂ -and-L ₃ ?	.22	.16	.34	.22

Note. L₁? = the cue was presented on List 1, L₂? = the cue was presented on List 2, L₃? = the cue was presented on List 3, and L₁-and-L₂-and-L₃? = the cue was presented on List 1 and List 2 and List 3. $p(L_1\text{-and-}L_2\text{-and-}L_3?)$ is the conjunction illusion index.

lists when they never were. When the six mean values of $p(L_1\text{-and-}L_2\text{-and-}L_3?)$ in the categorical judgment condition were tested for statistical significance, all were reliable—with values of the $t(111)$ test statistic ranging from a low of 8.28 (List 3, LF) to a high of 9.11 (List 1, LF; all $ps < .0001$). The mean acceptance probability for conjunctive source probes was reliably smaller than the mean probability for standard source probes that were false (grand $Ms = .19$ and $.32$), although for List 3, that pattern was reversed, with the mean probability being larger for conjunctive than for standard source probes that were false (grand $Ms = .22$ and $.12$). This reversal was previously observed by Brainerd et al. (2014) and is by-product of the fact that the false alarm rate for false standard probes is much lower for List 3 targets than it is for List 1 or List 2 targets.

The other finding that emerges from Table 3 is that conjunction illusions were not reliably >0 in the confidence rating condition. The mean value of $p(L_1\text{-and-}L_2\text{-and-}L_3?)$ in that condition across the Word Frequency $\times$ List Order cells was slightly negative, $-.01$, whereas the corresponding value with categorical judgments was positive, $.19$. When the negative mean value in the

confidence rating condition was tested for statistical significance, it did not differ reliably from zero.

Response format effects on conjunction illusions. Moving to ANOVA results, the design of this experiment was 2 (response format: categorical judgment vs. confidence rating) $\times$ 2 (word frequency: HF vs. LF) $\times$ 3 (list order: 1 vs. 2 vs. 3) $\times$ 4 (probe type: L₁? vs. L₂? vs. L₃? vs. L₁-and-L₂-and-L₃?). As in earlier experiments, we computed a preliminary 2 (response format) $\times$ 2 (word frequency) $\times$ 3 (list) ANOVA of the overdistribution metric $p(L_1\text{-and-}L_2\text{-and-}L_3?)$, to focus on the question of whether it reacted similarly to the word-frequency and list-order manipulations under both response formats. It did. To begin, there were main effects for response format, $F(1, 226) = 105.76$, $MSE = 0.12$, $\eta^2 = .32$, word frequency, $F(1, 226) = 34.82$, $MSE = 0.05$, $\eta^2 = .13$, and list, $F(2, 452) = 7.22$, $MSE = 0.03$, $\eta^2 = .03$. Concerning those effects, mean values of $p(L_1\text{-and-}L_2\text{-and-}L_3?)$ were lower with confidence ratings than categorical judgments, they were lower for HF than for LF words, and the mean values followed the same list ordering as the ordering of DI values in Experiment 2 (i.e., List 1 $>$ List

2 > List 3). Post hoc analysis of the list-order effect indicated that the mean values for the first two lists were both reliably larger than the mean value for List 3 but did not differ reliably from each other.

Next, with respect to interactions, there was a Response Format $\times$ Word Frequency $\times$ List Order interaction, $F(2, 452) = 22.06$, $MSE = 0.06$, $\eta^2 = .09$. Post hoc analysis yielded the following pattern. First, as in Experiment 1, the frequency effect was smaller with confidence ratings than with categorical judgments. Second, that difference in the frequency effect was only reliable for List 1 targets.

Response format effects on all source probes. We computed a 2 (response format) $\times$ 2 (word frequency) $\times$ 3 (list order) $\times$ 4 (probe type) ANOVA, the results of which were as follows. First, there were main effects for response format, $F(1, 226) = 50.22$, $MSE = 0.19$, $\eta^2 = .18$, word frequency, $F(1, 226) = 39.47$, $MSE = 0.10$, $\eta^2 = .15$, list, $F(2, 452) = 74.40$, $MSE = 0.09$, $\eta^2 = .25$, and probe type, $F(3, 678) = 98.07$, $MSE = 0.08$, $\eta^2 = .31$. As can be seen in Table 3, mean acceptance probabilities were lower for confidence ratings than for categorical judgments and for HF targets than for LF targets. The order of acceptance probabilities for the three lists was List 1 > List 2 > List 3, with paired-samples t tests showing that they were reliably higher for List 1 than for the other two lists but that List 2 and List 3 did not differ reliably (as in Experiment 2). With respect to the probe type main effect, the order of probabilities was $L_3? > L_1? > L_2? > L_1\text{-and-}L_2\text{-and-}L_3?$, with paired-samples t tests showing that probabilities were reliably higher for $L_3?$ than for any of the other probes and were reliably higher for $L_1?$ and $L_2?$ than for $L_1\text{-and-}L_2\text{-and-}L_3?$.

Second, there was a supervening Response Format $\times$ Word Frequency $\times$ List Order $\times$ Probe Type interaction, $F(6, 1356) = 14.14$, $MSE = 0.04$, $\eta^2 = .06$. Post hoc analysis produced two interesting component effects. The first was concerned with how standard versus conjunctive source probes interacted with response format, word frequency, and list order. We just saw that the conjunctive probes interacted simultaneously with all three manipulations. Standard probes did, too, but the nature of the interaction was somewhat different, owing to the fact that whereas some standard probes were true and some were false, all conjunctive probes were false. For standard probes, (a) response probabilities reversed as function of list, naturally, because which probe was correct depended on list, (b) those changes were larger for LF than for HF targets, and (c) those changes were the same in the two response format conditions. Recall that this is the same pattern that was observed for standard source probes in Experiment 2. When these results are combined with those for Experiments 1 and 2, the overriding patterns are that, qualitatively speaking, standard probes always reacted similarly to the experimental manipulations in both response format conditions, and the disjunction and conjunction measures of overdistribution also reacted similarly to the experimental manipulations in both conditions.

The second important component effect was that as predicted on the basis of the different memory demands of standard and conjunctive source probes, the response format manipulation had stronger effects on the latter than on the former. With conjunctive probes, the mean difference in response probabilities for categorical judgments versus confidence ratings was .19, whereas the corresponding difference for standard probes was .07, a highly reliable difference. Note, too, in connection with standard probes,

that the reduction in response probabilities that was caused by switching from categorical judgments to confidence ratings was smaller in this experiment than in Experiment 2 (.07 vs. .16). That was *not* because replacement of item probes (Experiment 2) with more demanding conjunctive probes (Experiment 3) produced an overall reduction in response probabilities for standard probes. Rather, it was because it simultaneously drove up those probabilities in the categorical judgment condition (from .23 to .29), and drove them down in the confidence rating condition (from .16 to .13). (We tested both effects for statistical significance by computing one-sample t tests on the probabilities for standard probes in this experiment, and we found that relative to Experiment 2, both the increase in the categorical judgment condition and the decrease in the confidence rating condition were reliable.) With standard probes, then, the contrast in memory demands between them and conjunctive probes may increase reliance on noncompensatory memories with categorical judgments but reduce such reliance with confidence ratings.

Summary. Substituting conjunction illusion measures of overdistribution for the disjunction illusion measures of earlier experiments did not temper the ability of confidence ratings to reduce overdistribution. On the contrary, the effects of switching from categorical judgments to confidence ratings on conjunction illusions were analogous to the effects of this manipulation on disjunction illusions in Experiment 1. In both instances, illusions that were highly reliable with categorical judgments were not reliable with confidence ratings.

Other important findings are concerned with how the conjunction illusion metric reacted to experimental manipulations in the two response format conditions. Here, remember that in prior experiments, the DI metric responded in a similar manner to the word-frequency and list-order manipulations in both conditions. In this experiment, the same was true of the conjunction illusion metric. List ordering was the same under both response formats, although pairwise differences were only reliable with categorical judgments, and those values were larger for LF than for HF words under both formats. In short, regardless of whether disjunction or conjunction illusions provide the measure of overdistribution and whether subjects respond to probes with categorical judgments or confidence ratings, overdistribution is always strongest for List 1 targets and LF targets.

General Discussion

The significance of overdistribution illusions lies in their demonstration that memory is in contradictory minds about experience. Based on accumulated evidence, we remember events as belonging to too many episodic states as a matter of course. This includes states that are logically or empirically incompatible with one another, such as O versus NS in conjoint recognition and List 1 versus List 2 in source monitoring. The current theoretical explanation of overdistribution turns on the notion that some of the memory traces that subjects rely on are noncompensatory with respect to events' episodic states. Although it seems counterintuitive, some traces that are stored when events are encoded (a) support remembering cues to be both O and NS in conjoint recognition and (b) support remembering cues that were presented in a single context as having been presented in multiple contexts. In particular, FTT's notion of gist traces of the meaning of prior

events has those properties: Relying on gist (e.g., “living room furniture”) supports acceptance of both O? and NS? probes for given cues (e.g., *sofa, couch*) in conjoint recognition, and likewise, it supports acceptance of L1? and L2? probes for cues that were presented in only one of those contexts, as well as conjunctive source probes for those cues.

Two indexes of overdistribution have been studied in prior experiments, disjunction and conjunction illusions. Although both are memory distortions, they have rich histories in judgment and decision making research, where they are classic examples of illusions and biases in reasoning (Tversky & Kahneman, 1983; Tversky & Koehler, 1994). In that literature, these illusions have also been explained as by-products of relying on noncompensatory gist (e.g., Reyna & Brainerd, 2011), an explanation that has been evaluated with various manipulations that encourage or discourage such reliance during reasoning (e.g., Kuhberger et al., 1999; Kuhberger & Tanner, 2010; Wolfe et al., 2013; Wolfe & Reyna, 2010; for a review, see Reyna & Brainerd, 2011). The bulk of those manipulations are specific to judgment and decision making tasks and are not easily adapted to memory research. However, specificity of response format is one that is readily applicable in both domains—the manipulation being whether subjects respond by making categorical judgments or graded ratings, such as levels of confidence or preference. The task calibration principle supplies the theoretical basis for this manipulation, according to which response formats and retrieval cues that are more graded reduce subjects’ tendency to base responses on gist. The distinction between categorical judgments and confidence ratings is a special case of this principle that has been studied with illusions and biases in reasoning (e.g., Mills et al., 2008; Reyna et al., 2011).

We implemented that distinction by comparing both indexes of overdistribution when the response format was categorical judgments versus confidence ratings. Our core hypothesis was that if disjunction and conjunction illusions are at least partly due to noncompensatory gist memories, those illusions would shrink when confidence ratings replace categorical judgments because a more graded response format reduces reliance on such memories. The results of our experiments were in line with that prediction. Relative to categorical judgments, confidence ratings completely restored compensation among incompatible episodic states for targets in Experiments 1 and 3 and mostly restored it in Experiment 2. One of the strongest findings indicating that, indeed, this effect was due to reduced reliance on noncompensatory gist memories is that the effect was never observed with distractors, for which such memories were presumably not available.

To conclude this article, we briefly discuss three questions arising from our experiments, each of which is of some significance when it comes to the theoretical interpretation of overdistribution and the influence of response format. The first is the details of how confidence ratings suppress overdistribution, which are somewhat different for conjunction illusions than for disjunction illusions. The second is a question with broader implications for the use of categorical judgments and confidence ratings in applied contexts, such as eyewitness identification of criminal suspects—namely, relations among confidence ratings, categorical judgments, and accuracy. The third is an alternative theoretical account of disjunction illusions that postulates a different process mechanism, source guessing.

How Confidence Ratings Suppress Overdistribution

In disjunction illusion experiments, the manner in which confidence judgments must affect source and item probes to reduce overdistribution falls out of the expression $DI = p(L_1) + p(L_2) - p(L_1\text{-or-}L_2)$. Confidence ratings must decrease $p(L_1) + p(L_2)$ more than $p(L_1\text{-or-}L_2)$. Confidence ratings could also reduce DI by increasing $p(L_1\text{-or-}L_2)$ more than $p(L_1) + p(L_2)$, but the calibration principle predicts that they will reduce target acceptance probabilities, increasingly so as probes’ perceived memory demands increase. In line with that prediction, confidence ratings did not increase target acceptance probabilities in any condition of any of the experiments. Instead, these probabilities always decreased, as would be expected if confidence judgments reduce reliance on gist memories, and more so for source than for item probes.

In conjunction illusion experiments, on the other hand, a direct measure of overdistribution is available—namely, $p(L_1\text{-and-}L_2\text{-and-}L_3)$. Thus, the idea that confidence ratings reduce overdistribution by decreasing reliance on noncompensatory gist memories simply predicts that $p(L_1\text{-and-}L_2\text{-and-}L_3)$ for targets will shrink, relative to the values that are obtained with categorical judgments under otherwise identical conditions. They did. Task calibration makes the further prediction that if, as the source monitoring literature suggests, subjects are sensitive to variations in the memory demands of different source tests (Hicks & Starns, 2006a, 2006b), the shrinkage induced by confidence judgments ought to be greater for conjunctive probes than for standard ones. That effect was observed, too.

Another important finding that bears on the task calibration principle concerns the effects of confidence ratings on distractor acceptance probabilities. Suppose that confidence ratings do not reduce reliance on gist, and instead, they simply make subjects more conservative, increasingly so for more demanding types of probes. If so, target acceptance probabilities will decline, and they will decline more for probes with higher perceived memory demands, as we observed. However, distractor acceptance probabilities will also decline, but that effect was never observed. On the contrary, it can be seen in Tables 1–3 that distractor acceptance probabilities usually increased slightly when subjects made confidence ratings.

Categorical Judgments, Confidence Ratings, and Accuracy

Next, we consider three questions about the relation between the two response formats and accuracy in our experiments: First, do confidence ratings enjoy a global accuracy advantage over categorical judgments? Second, do confidence ratings have a net accuracy advantage over categorical judgments in the particular conditions of our experiments? Third, do confidence ratings predict accuracy in our experiments, and if so, is the relation positive or negative?

Global accuracy advantage. Overdistribution illusions are reality distortions; test cues should not be remembered as belonging to mutually incompatible episodic states. As confidence ratings suppressed this type of distortion so thoroughly, it seems natural to suppose that they must convey global accuracy advantages, relative to categorical judgments, on memory tests. Actually, however, the relation between response format and accuracy is more com-

plex because it depends upon the nature of the memory test and, more especially, on whether gist reliance is good or bad for accuracy. Across our experiments, three distinct types of tests were administered: item probes (Experiments 1 and 2), standard source probes (all experiments), and conjunctive source probes (Experiment 3). For target cues, item probes were always true, conjunctive source probes were always false, and standard source probes were sometimes true and sometimes false.

As discussed, prior research suggests that gist memories contribute to subjects' performance on all three of these tests. Because prior research also indicates that confidence ratings lessen reliance on gist, we expected that they would lower response probabilities in all conditions for all three types of tests, which they did. Notice that such reductions both impair and enhance the accuracy of subjects' responses, depending on the test. With conjunctive source probes and standard source probes that are incorrect (e.g., $L_2?$ for List 1 targets), performance is necessarily more accurate for confidence ratings than for categorical judgments. With item probes and standard source probes that are correct (e.g., $L_1?$ for List 1 targets), on the other hand, performance is necessarily less accurate for confidence ratings than for categorical judgments (cf. Benjamin, Tullis, & Lee, 2013, for a related discussion of item probes).

Net accuracy advantage. Although, qualitatively, confidence ratings cannot confer a global accuracy advantage if they reduce response probabilities on all memory tests, some of the theoretical principles that figured in our research predict that, quantitatively, they should yield a net accuracy advantage across the various conditions of the experiments. For instance, consider the two tests for which accepting probes as true is either uniformly correct or uniformly incorrect—namely, item probes and conjunctive source probes. It follows from the principle that subjects are sensitive to the different memory demands of these tests that confidence ratings should lower acceptances more for the latter (enhancing accuracy) than for the former (impairing accuracy). Because the last two experiments were similar in design, a test of that prediction can be obtained by examining response probabilities under the two response formats for item probes in Experiment 2 versus conjunctive source probes in Experiment 3. It can be seen in Tables 2 and 3 that, as predicted, the reduction in the confidence rating condition was more pronounced for conjunctive probes, and indeed, confidence ratings did not produce reliable reductions in acceptance of item probes in Experiment 2.

A parallel prediction can be made about the effect of confidence ratings on standard source probes that were correct (e.g., $L_1?$ for List 1 targets) versus incorrect (e.g., $L_2?$ for List 1 targets). Theoretically, subjects rely on a mix of verbatim and gist traces as a basis for accepting the former as true, while they rely on gist but not verbatim traces as a basis for accepting the latter as true (Titcomb & Reyna, 1995). It follows that reducing reliance on gist should lower response probabilities more for incorrect than for correct probes, and that prediction was confirmed. Within each List $\times$ Word Frequency cell of each experiment, both correct and incorrect source probes were administered, so that the effects of switching from categorical judgments to confidence ratings can be compared as a function of whether a probe was correct or incorrect, with all other factors constant. Across experiments and conditions, 16 such comparisons were possible for correct probes, and 28 were possible for incorrect probes. When we computed the

average reduction in acceptance probability that was produced by switching from categorical judgments to confidence ratings, the reduction was 50% greater for incorrect than for correct probes.

Thus, although confidence ratings cannot confer a global accuracy advantage on all episodic memory tests, relative to categorical judgments, they produced a net accuracy advantage, which could be predicted on theoretical grounds. However, it is also possible to predict, with other configurations of memory tests, that confidence ratings will produce net accuracy *disadvantages*. To illustrate, consider a familiar type of design from the false memory literature in which (a) the test cues are O, NS, and ND words, and (b) the test probe for each word is either O? or NS-or-O? (Brainerd & Reyna, 2002; Koutstaal, 2003; Lampinen et al., 2005). Prior research suggests that subjects rely on verbatim and gist traces to correctly accept both types of probes for O cues and to correctly accept NS-or-O? probes for NS? cues, but they rely on gist traces to incorrectly accept O? probes for NS cues and verbatim traces to correctly reject them (Brainerd, Stein, & Reyna, 1998). If confidence ratings were compared to categorical judgments, the predictable result is that confidence ratings' tendency to lessen reliance on gist memory would yield net reductions in accuracy: Acceptance rates would fall for all four types of probes, but acceptance is the correct response for three of them.

Summing up, the overall picture of how response format influences accuracy reduces to two conclusions. First, if confidence ratings lower reliance on gist, whether they enhance or impair accuracy depends on whether such memories support correct or incorrect responses. Second, it is possible to anticipate what the net influence of confidence ratings on accuracy should be in our particular designs by considering two factors: the configuration of memory tests and whether the effects of confidence ratings are more marked for some tests than for others.

Confidence-accuracy relations. A perennial question about confidence ratings in memory research is how well they predict accuracy (Brewer & Wells, 2006; Busey et al., 2000; Jones et al., 2008; Juslin et al., 1996; Wells & Murray, 1984). That question has been studied for over a century (e.g., Dallenbach, 1913) and, in recent decades, most extensively in connection with eyewitness identification of criminal suspects. The motivation there is that confidence ratings are used to gauge the accuracy of eyewitness identifications (Technical Working Group for Eyewitness Evidence, 1999), and when faulty identifications are accompanied by high confidence ratings, they are known to stimulate false convictions (Connors, Lundregan, Miller, & McEwan, 1996). The conventional conclusion is that there is a moderate positive correlation between accuracy and confidence that does not rise to a forensic standard; that is, confidence increases as accuracy increases but not enough to preclude unacceptable levels of false identification (Connors et al., 1996). Although that conclusion is widely accepted in forensic psychology, it has recently been challenged on the ground that certain eyewitness identification tests yield strong positive correlations (Wixted, Mickes, Clark, Gronlund, & Roediger, 2015). Further, Roediger and associates have documented *negative* confidence-accuracy correlations for certain types of tests (DeSoto & Roediger, 2014; Roediger & DeSoto, 2014; Roediger, Wixted, & DeSoto, 2012).

Our data also bear on the confidence-accuracy relation, and we briefly illustrate that fact. First, it is important to note that eyewitness identification is a form of source memory inasmuch as sub-

jects' task is to decide whether cues (suspects' faces) were seen in specific contexts (e.g., a bank robbery), and as in our experiments, a common error consists of falsely accepting cues that were actually seen in other contexts (e.g., parks, stores, and sporting events). Traditionally, subjects make categorical accept-reject judgments about individual cues, followed by a numerical rating of confidence in the accuracy of each judgment, which means that confidence-accuracy is ultimately a question about the correlation between categorical judgments and confidence ratings. In the literature, that correlation has been computed in three ways (see DeSoto & Roediger, 2014): (a) between-events, which is the correlation between categorical judgments and confidence ratings across groups of cues (conditions) that produce different levels of acceptance for categorical judgments (e.g., HF vs. LF in our experiments); (b) between-subjects, which is the correlation between individual subjects' mean accuracy scores for categorical judgments and their mean confidence ratings; and (c) within-subjects, which is the correlation, for individual subjects, between whether given categorical judgments are correct and the magnitude of the corresponding confidence rating. Note that the first type of correlation can be computed with our data because the experimental manipulations produced groups of cues whose acceptance probabilities varied substantially in the two response format conditions.

Before reporting confidence-accuracy results, we consider whether confidence ratings generally tracked categorical judgments over conditions that affected acceptance probabilities. The most incisive findings are provided by simple source probes, which were administered in all conditions of all experiments and, as we know, could be either correct or incorrect. There were two pertinent findings. The first is shown in Panel A of Figure 2. If subjects have accurate memory for the contexts in which target cues were presented, categorical judgments will produce higher

response probabilities for correct than for incorrect source probes across conditions and experiments. It can be seen that this was indeed the case. Then, if confidence ratings track categorical judgments, mean probabilities in the confidence rating condition should also be higher for correct than for incorrect source probes, and it can be seen that they were. Further, confidence ratings produce better separation between correct and incorrect source probes because the difference in mean probabilities was larger than it was for categorical judgments. Readers will have noticed that the latter result is a necessary consequence of the aforementioned finding that confidence judgments reduced response probabilities more for incorrect than for correct source probes.

Returning to confidence-accuracy, those results appear in Panels B and C of Figure 2, where mean correct categorical judgments about probes are plotted against mean confidence that those probes were correct. Although confidence ratings for source probes tracked categorical judgments over the List $\times$ Word Frequency cells of our experiments, how strong was that relation and was it positive or negative? We regressed confidence ratings on mean correct categorical response probabilities separately for the 16 types of correct source probes (Panel B) and the 28 types of incorrect source probes (Panel C). This produced correlations of similar strength for correct and incorrect probes, accounting for 22 and 26% of the variance, respectively.

The answer to the second question is that the confidence-accuracy relation was neither exclusively positive nor exclusively negative, but rather, its direction depended on whether source probes were correct or incorrect. It can be seen in Figure 2B that the best-fitting regression line had a positive slope for correct source probes, but it can be seen in Figure 2C that the best-fitting line for incorrect source probes had a negative slope. Thus, there were *opposite* confidence-accuracy relations with source probes for hits versus correct rejections; confidence strengthened as the hit rate increased but weakened as the correct rejection rate increased.

Overall, then, comparisons of categorical judgments and confidence ratings over conditions for correct versus incorrect source probes yielded three conclusions. First, confidence ratings tracked categorical judgments' ability to discriminate between correct and incorrect source probes. Second, confidence ratings were better discriminators of performance on correct versus incorrect source probes than categorical judgments were. Third, the confidence-accuracy relation was positive for correct source probes but negative for incorrect source probes. More important, this finding of opposite confidence-accuracy correlations for hits versus correct rejections echoes recent findings that Roediger and associates have reported for item recognition in semantic false memory tasks (DeSoto & Roediger, 2014; Roediger & Desoto, 2014; Roediger et al., 2012). Based on the latter evidence and the data of our experiments, it appears that confidence-accuracy correlations are positive for true memory and negative for false memory in the two most widely used false-memory paradigms.

Source Guessing Explanation of Disjunction Illusions

To conclude this article, we examine an alternative account of the disjunction illusion form of overdistribution, which was proposed by Kellen, Singmann, and Klauer (2014) and was also considered by Brainerd et al. (2012). The study of disjunction

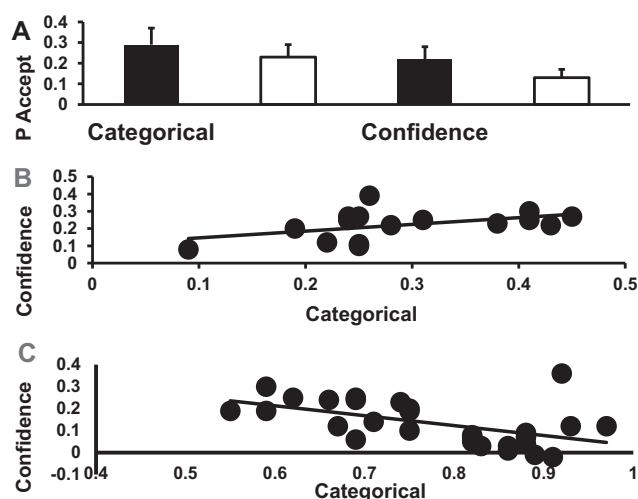


Figure 2. Relations between categorical judgments and confidence ratings, pooled over the conditions of Experiments 1–3. Panel A = the separation between response probabilities for correct (black bars) versus incorrect (white bars) source probes that is produced by categorical judgments and confidence ratings. Panel B = the best-fitting regression line with correct source probes, for categorical judgments versus confidence ratings. Panel C = the best-fitting regression line with incorrect source probes, for categorical judgments versus confidence ratings.

illusions began with *item* memory (see Figure 1), because these phenomena were predicted in conjoint recognition designs by the verbatim-gist principle (Brainerd & Reyna, 2008): If subjects rely on gist memory when responding to O and NS cues, overdistribution will occur because gist is noncompensatory across O? and NS? probes. Because gist is also noncompensatory across simple source probes ($L_1?$ and $L_2?$), the same theoretical principle predicts disjunction illusions in source designs.

Kellen et al. (2014) proposed that source disjunction illusions could also result from a guessing process, which is derived from Batchelder-type models of source monitoring (e.g., Batchelder & Riefer, 1990; Bayen, Murnane, & Erdfelder, 1996). To explain, consider the Batchelder-type model for Experiment 1, which is shown in Table 4. Models of this family assume that when a source probe is presented for a target cue (Was *bagpipe* on List 1?), one of three memory states is induced: M_1 = the cue is recognized as old and its source is recalled (with probability Dd), M_2 = the cue is recognized as old but its source is not recalled (with probability $D(1 - d)$), and M_3 = the cue is not recognized as old (with probability $1 - D$). When the state is M_2 or M_3 , subjects are uncertain as to whether the source that is indicated in the probe (e.g., List 1) is correct, and they are said to guess that it is correct with probabilities g and b , respectively. In this model, disjunction illusions are tied to the values of these guessing parameters, which can be estimated from the data if the model fits.

To see how that falls out, consider as an example the first three lines of Table 4, which contain the model expressions for $p(L_1|L_1)$, $p(L_1|L_2)$, and $p(L_1\text{-or-}L_2|L_1)$. This is a 2HT model inasmuch as there is a detect-old threshold for targets, which is measured by D_1 in this example, and a detect-new threshold for distractors, which is measured by D_N (see the last three lines of Table 4). As Snodgrass and Corwin (1988) pointed out, it is traditionally assumed that the high and low thresholds are equal; $D_1 = D_N$. Under that assumption, the expressions in our example simplify as $p(L_1?|L_1) = D_1d_1 + D_1(1 - d_1)g$, $p(L_1?|L_2) = D_1(1 - d_1)g$, and $p(L_1\text{-or-}L_2|L_1) = D_1$. Because $DI = p(L_1?|L_1) +$

$p(L_1?|L_2) - p(L_1\text{-or-}L_2|L_1)$, whether $DI > 0$ obviously depends on whether $[D_1d_1 + 2D_1(1 - d_1)g] > D_1$. Notice that $D_1d_1 + 2D_1(1 - d_1)g$ can be rearranged to yield $D_1[d_1 - 2d_1g + 2g]$, so that the question of whether $[D_1d_1 + 2D_1(1 - d_1)g] > D_1$ reduces to whether $[d_1 - 2d_1g + 2g] > 1$. That, in turn, is controlled by the value of g because it is easy to see that $[d_1 - 2d_1g + 2g] > 1$ if and only if $g > .5$, that $[d_1 - 2d_1g + 2g] < 1$ if and only if $g < .5$, and that $[d_1 - 2d_1g + 2g] = 1$ if and only if $g = .5$. Therefore, if the model fits the data and subjects guess that a proffered source is correct with probability $> .5$ when the induced state is M_2 , $DI > 0$. If the assumption that the old and new thresholds are equal is not used, the equations do not simplify in this manner. Now, although DI still depends on the value of the g parameter, it also depends on the value of the b parameter. This is conceptually important because g is a source guessing parameter but b is an item guessing parameter. Thus, in the first situation DI depends on how liberal subjects are when they guess the source of an item that they know is old, but in the second situation, DI also depends on how liberal subjects are when they guess the source of an item that they do not know is old.

In the present experiments, the source guessing explanation of how response format affects disjunction illusions is that confidence ratings make subjects more conservative when they are in states of uncertainty (M_2 and M_3) with respect to whether indicated source probes are correct. In other words, confidence ratings reduce the values of the g and b parameters, which Kellen et al. (2014) suggested could be done if subjects more carefully match their guesses to test cues' base rates. That is how the model in Table 4 explains the fact that confidence ratings reduce target DI values. However, note that this explanation makes the further prediction that confidence ratings will reduce acceptance probabilities for distractors as well as targets because the parameter b appears in the expressions for distractors: $p(L_1?|D) = (1 - D_N)b$ and $p(L_2?|D) = (1 - D_N)b$. With respect to that prediction, recall that confidence ratings never reduced distractor acceptance probabilities in any of our experiments, and hence, the distractor data did not support the source guessing explanation.

At a more general level, we concluded in prior work (Brainerd et al., 2012, 2015) that the verbatim-gist account has some theoretical advantages over the source guessing account of disjunction illusions. The main ones are (a) parsimony, (b) breadth of empirical support, and (c) predictive power. Concerning a, the verbatim-gist account explains overdistribution in both item and source memory with a single idea, whereas source guessing only explains disjunction illusions in source memory. As we have seen, there is substantial evidence of overdistribution in item memory as well as in source memory, and actually, the item evidence is more extensive (see Figure 1). Consequently, if we adopt the guessing account for source memory, the verbatim-gist account (or other theoretical principles) is still needed to explain item disjunction illusions.

Turning to the second advantage, the empirical basis for the distinction between verbatim and gist memory is broader than that for source guessing. The former distinction has been widely implemented in both the judgment and decision making and memory literatures (for reviews, see Brainerd & Reyna, 2005; Reyna & Brainerd, 2011). It has been used to explain classic effects in each literature and to predict new ones—including counterintuitive effects such as developmental reversals in false memory and in reasoning illusions. Various manipulations have been identified that shift memory in a verbatim or gist

Table 4
Processes That Are Measured in Batchelder and Riefer's (1990) and Bayen et al.'s (1996) Models When They Are Implemented in Full Factorial Source Designs

Observable probability	Source model expression
$p(L_1? L_1)$	$D_1d_1 + D_1(1 - d_1)g + (1 - D_1)b$
$p(L_2? L_1)$	$D_1(1 - d_1)g + (1 - D_1)b$
$p(L_1UL_2? L_1)$	$D_1 + (1 - D_1)a$
$p(L_1? L_2)$	$D_2(1 - d_2)g + (1 - D_2)b$
$p(L_2? L_2)$	$D_2d_2 + D_2(1 - d_2)g + (1 - D_2)b$
$p(L_1UL_2? L_2)$	$D_2 + (1 - D_2)a$
$p(L_1? D)$	$(1 - D_N)b$
$p(L_2? D)$	$(1 - D_N)b$
$p(L_1UL_2? D)$	$(1 - D_N)a$

Note. With respect to the proof in the text that disjunction illusions in source memory depend on the value of the g parameter, D_1d_1 is the probability that a List 1 target induces state M_1 = the target is recognized as old and its source is recalled; $D_1(1 - d_1)$ is the probability that a List 1 target induces state M_2 = the target is recognized as old and its source is not recalled; and $1 - D_1$ is the probability that a List 1 target induces state M_3 = the subject does not recognize that a target is old and does not recall its source.

direction on reasoning problems or memory tests, with the present response format manipulation being only one example. Empirical support for the source guessing process is far more limited. In the literature on Batchelder-type models, research has focused chiefly on manipulations that are designed to affect the item memory and source discrimination parameters, rather than manipulations that are designed to affect guessing parameters (for an exception, see Kellen et al., 2014).

The third advantage is that the verbatim-gist account predicts as well as explains overdistribution illusions. It predicted these phenomena in item memory before they were observed, on the ground that gist memory is noncompensatory—that it supports remembering O and NS cues as both old and new. Evidence confirming that prediction and extending it to source designs was reported before the source guessing explanation was proposed. In contrast, as Brainerd et al. (2015) discussed, source guessing cannot *predict* any specific relation between $p(L_1|L_1) + p(L_1|L_2)$ and $p(L\text{-or-}L_2|L_1)$. Rather, it only *explains* observed relations *ex post facto* because, as we saw, it permits all possible relations between $p(L_1|L_1) + p(L_1|L_2)$ and $p(L\text{-or-}L_2|L_1)$, depending on the estimated value of g . With our response format manipulation, source guessing makes no advance prediction that confidence ratings will suppress the DI metric. There is only the mathematical constraint that if that happens and if the model in Table 4 fits the data, the g parameter will have smaller values in the confidence rating condition. Thus, the source guessing explanation depends on disjunction illusions, rather than disjunction illusions depending on the source guessing explanation. Data must exist *before* models can be fit and parameters estimated, from which it follows that whether disjunction illusions are present in an experiment is not controlled by estimates of g . Instead, estimates of g are controlled by whether the relation among source and item acceptance probabilities is $[p(L_1|L_1) + p(L_1|L_2)] > p(L\text{-or-}L_2|L_1)$.

Summing up, there are currently two theoretical accounts of overdistribution illusions. One posits that overdistribution, like various other forms of distortion, is a by-product of gist memory, and it predicts disjunction and conjunction illusions in item and source memory as consequences of that principle. The other does not posit such a general principle and, instead, focuses on explaining disjunction illusions in source memory via a guessing process that figures in source models. As things stand, the second explanation faces a series of empirical and theoretical challenges. However, little experimentation has as yet been reported on this second explanation, and some of these challenges may recede as data accumulate.

References

- Arndt, J. (2012). False recollection: Empirical findings and their theoretical implications. *Psychology of Learning and Motivation*, 56, 81–124. <http://dx.doi.org/10.1016/B978-0-12-394393-4.00003-0>
- Ball, B. H., DeWitt, M. R., Knight, J. B., & Hicks, J. L. (2014). Encoding and retrieval processes involved in the access of source information in the absence of item memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 1271–1286. <http://dx.doi.org/10.1037/a0037204>
- Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, 97, 548–564. <http://dx.doi.org/10.1037/0033-295X.97.4.548>
- Bayen, U. J., Murnane, K., & Erdfelder, E. (1996). Source discrimination, item detection, and multinomial models of source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 197–215. <http://dx.doi.org/10.1037/0278-7393.22.1.197>
- Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: Evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 1601–1608. <http://dx.doi.org/10.1037/a0031849>
- Bjorklund, D. F., Cassel, W. S., Bjorklund, B. R., Brown, R. D., Park, C. L., Ernst, K., & Owen, F. A. (2000). Social demand characteristics in children's and adults' eyewitness memory and suggestibility: The effect of different interviewers on free recall and recognition. *Applied Cognitive Psychology*, 14, 421–433. [http://dx.doi.org/10.1002/1099-0720\(200009\)14:5<421::AID-ACPP659>3.0.CO;2-4](http://dx.doi.org/10.1002/1099-0720(200009)14:5<421::AID-ACPP659>3.0.CO;2-4)
- Brainerd, C. J., Holliday, R. E., Nakamura, K., & Reyna, V. F. (2014). Conjunction illusions and conjunction fallacies in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 1610–1623. <http://dx.doi.org/10.1037/xlm0000017>
- Brainerd, C. J., & Reyna, V. F. (2002). Recollection rejection: How children edit their false memories. *Developmental Psychology*, 38, 156–172. <http://dx.doi.org/10.1037/0012-1649.38.1.156>
- Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. New York, NY: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780195154054.001.0001>
- Brainerd, C. J., & Reyna, V. F. (2008). Episodic over-distribution: A signature effect of familiarity without recollection. *Journal of Memory and Language*, 58, 765–786. <http://dx.doi.org/10.1016/j.jml.2007.08.006>
- Brainerd, C. J., Reyna, V. F., Holliday, R. E., & Nakamura, K. (2012). Overdistribution in source memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38, 413–439. <http://dx.doi.org/10.1037/a0025645>
- Brainerd, C. J., Stein, L. M., & Reyna, V. F. (1998). On the development of conscious and unconscious memory. *Developmental Psychology*, 34, 342–357. <http://dx.doi.org/10.1037/0012-1649.34.2.342>
- Brainerd, C. J., Wang, Z., & Reyna, V. F. (2013). Superposition of episodic memories: Overdistribution and quantum models. *Topics in Cognitive Science*, 5, 773–799.
- Brainerd, C. J., Wang, Z., Reyna, V. F., & Nakamura, K. (2015). Episodic memory does not add up: Verbatim-gist superposition predicts violations of the additive law of probability. *Journal of Memory and Language*, 84, 224–245. <http://dx.doi.org/10.1016/j.jml.2015.06.006>
- Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: Effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied*, 12, 11–30. <http://dx.doi.org/10.1037/1076-898X.12.1.11>
- Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, 21, 916–944. <http://dx.doi.org/10.1080/09658211.2013.767348>
- Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic Bulletin & Review*, 7, 26–48. <http://dx.doi.org/10.3758/BF03210724>
- Connors, E., Lundregan, T., Miller, N., & McEwan, T. (1996). *Convicted by juries, exonerated by science: Case studies in the use of DNA evidence to establish innocence after trial*. Alexandria, VA: National Institute of Justice.
- Corbin, J. C., Reyna, V. F., Weldon, R. B., & Brainerd, C. J. (2015). How reasoning, judgment, and decision making are colored by gist-based intuition: A fuzzy-trace theory approach. *Journal of Applied Research in Memory & Cognition*, 4, 344–355.
- Dallenbach, K. M. (1913). The relation of memory error to time interval. *Psychological Review*, 20, 323–337. <http://dx.doi.org/10.1037/h0076103>
- Dennis, N. A., Hayes, S. M., Prince, S. E., Madden, D. J., Huettel, S. A., & Cabeza, R. (2008). Effects of aging on the neural correlates of

- successful item and source memory encoding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 791–808. <http://dx.doi.org/10.1037/0278-7393.34.4.791>
- DeSoto, K. A., & Roediger, H. L., III. (2014). Positive and negative correlations between confidence and accuracy for the same events in recognition of categorized lists. *Psychological Science*, 25, 781–788. <http://dx.doi.org/10.1177/0956797613516149>
- Dougherty, M. R. P., & Hunter, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, 31, 968–982. <http://dx.doi.org/10.3758/BF03196449>
- Dougherty, M. R., & Sprenger, A. (2006). The influence of improper sets of information on judgment: How irrelevant information can bias judged probability. *Journal of Experimental Psychology: General*, 135, 262–281. <http://dx.doi.org/10.1037/0096-3445.135.2.262>
- Estes, W. K., & Maddox, W. T. (2002). On the processes underlying stimulus-familiarity effects in recognition of words and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 1003–1018. <http://dx.doi.org/10.1037/0278-7393.28.6.1003>
- Feynman, R. P., Leighton, R. B., & Sands, M. (1965). *The Feynman lectures on physics* (Vol. 3). Reading, MA: Addison Wesley.
- Finn, B., & Metcalfe, J. (2008). Judgments of learning are influenced by memory for past test. *Journal of Memory and Language*, 58, 19–34. <http://dx.doi.org/10.1016/j.jml.2007.03.006>
- Gallo, D. A. (2004). Using recall to reduce false recognition: Diagnostic and disqualifying monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 120–128. <http://dx.doi.org/10.1037/0278-7393.30.1.120>
- Hall, J. F. (1979). Recognition as a function of word frequency. *The American Journal of Psychology*, 92, 497–505. <http://dx.doi.org/10.2307/1421568>
- Halpern-Felsher, B. L., Biehl, M., Kropp, R. Y., & Rubinstein, M. L. (2004). Perceived risks and benefits of smoking: Differences among adolescents with different smoking experiences and intentions. *Preventive Medicine*, 39, 559–567. <http://dx.doi.org/10.1016/j.ypmed.2004.02.017>
- Hauschildt, M., Peters, M. J. V., Jelinek, L., & Moritz, S. (2012). Veridical and false memory for scenic material in posttraumatic stress disorder. *Consciousness and Cognition*, 21, 80–89. <http://dx.doi.org/10.1016/j.concog.2011.10.013>
- Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29, 1210–1230. <http://dx.doi.org/10.1037/0278-7393.29.6.1210>
- Heathcote, A., Bora, B., & Freeman, E. (2010). Recollection and confidence in two-alternative forced choice episodic recognition. *Journal of Memory and Language*, 62, 183–203. <http://dx.doi.org/10.1016/j.jml.2009.11.003>
- Hicks, J. L., & Starns, J. J. (2006a). Remembering source evidence from associatively related items: Explanations from a global matching model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32, 1164–1173. <http://dx.doi.org/10.1037/0278-7393.32.5.1164>
- Hicks, J. L., & Starns, J. J. (2006b). The roles of associative strength and source memorability in the contextualization of false memory. *Journal of Memory and Language*, 54, 39–53. <http://dx.doi.org/10.1016/j.jml.2005.09.004>
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541. [http://dx.doi.org/10.1016/0749-596X\(91\)90025-F](http://dx.doi.org/10.1016/0749-596X(91)90025-F)
- Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33, 461–474. <http://dx.doi.org/10.1037/0278-7393.33.3.461>
- Jones, E. E., Williams, K. D., & Brewer, N. (2008). “I had a confidence epiphany!”: Obstacles to combating post-identification confidence inflation. *Law and Human Behavior*, 32, 164–176. <http://dx.doi.org/10.1007/s10979-007-9101-0>
- Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence-accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1304–1316. <http://dx.doi.org/10.1037/0278-7393.22.5.1304>
- Kellen, D., Singmann, H., & Klauer, K. C. (2014). Modeling source-memory overdistribution. *Journal of Memory and Language*, 76, 216–236. <http://dx.doi.org/10.1016/j.jml.2014.07.001>
- Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 34–53. <http://dx.doi.org/10.1037/0278-7393.27.1.34>
- Koutstaal, W. (2003). Older adults encode—But do not always use—Perceptual details: Intentional versus unintentional effects of detail on memory judgments. *Psychological Science*, 14, 189–193. <http://dx.doi.org/10.1111/1467-9280.01441>
- Kučera, H., & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.
- Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (1999). The effects of framing, reflection, probability, and payoff on risk preference in choice tasks. *Organizational Behavior and Human Decision Processes*, 78, 204–231. <http://dx.doi.org/10.1006/obhd.1999.2830>
- Kühberger, A., & Tanner, C. (2010). Risky choice framing: Task versions and a comparison of prospect-theory and fuzzy-trace theory. *Journal of Behavioral Decision Making*, 23, 314–329. <http://dx.doi.org/10.1002/bdm.656>
- Kurilla, B. P., & Westerman, D. L. (2010). Source memory for unidentified stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36, 398–410. <http://dx.doi.org/10.1037/a0018279>
- Lampinen, J. M., & Odegard, T. N. (2006). Memory editing mechanisms. *Memory*, 14, 649–654. <http://dx.doi.org/10.1080/09658210600648407>
- Lampinen, J. M., Odegard, T. N., Blackshear, E., & Toglia, M. P. (2005). Phantom ROC. In D. T. Rosen (Ed.), *Trends in experimental psychology research* (pp. 235–267). Hauppauge, NY: NOVA Science Publishers.
- Lampinen, J. M., Odegard, T. N., & Neuschatz, J. S. (2004). Robust recollection rejection in the memory conjunction paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 30, 332–342. <http://dx.doi.org/10.1037/0278-7393.30.2.332>
- Lampinen, J. M., Watkins, K. N., & Odegard, T. N. (2006). Phantom ROC: Recollection rejection in a hybrid conjoint recognition signal detection model. *Memory*, 14, 655–671. <http://dx.doi.org/10.1080/09658210600648431>
- Lindsay, D. S., & Johnson, M. K. (1989). The eyewitness suggestibility effect and memory for source. *Memory & Cognition*, 17, 349–358. <http://dx.doi.org/10.3758/BF03198473>
- Loftus, E. F. (1975). Leading questions and eyewitness report. *Cognitive Psychology*, 7, 560–572. [http://dx.doi.org/10.1016/0010-0285\(75\)90023-7](http://dx.doi.org/10.1016/0010-0285(75)90023-7)
- Malmberg, K. J. (2008). Recognition memory: A review of the critical findings and an integrated theory for relating them. *Cognitive Psychology*, 57, 335–384. <http://dx.doi.org/10.1016/j.cogpsych.2008.02.004>
- Mcbride, D. M., & Shouder, H. (2003). Conceptual processing effects on automatic memory. *Memory & Cognition*, 31, 393–400. <http://dx.doi.org/10.3758/BF03194397>
- Meiser, T., & Broder, A. (2002). Memory for multidimensional source information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 116–137. <http://dx.doi.org/10.1037/0278-7393.28.1.116>
- Mills, B., Reyna, V. F., & Estrada, S. (2008). Explaining contradictory relations between risk perception and risk taking. *Psychological Science*, 19, 429–433. <http://dx.doi.org/10.1111/j.1467-9280.2008.02104.x>
- Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency

- effects or reflect people's beliefs about memory? *Journal of Memory and Language*, 70, 1–12. <http://dx.doi.org/10.1016/j.jml.2013.09.007>
- Nakamura, K., & Brainerd, C. J. (2013, November). *Disjunction fallacies in episodic memory*. Paper presented at the Psychonomic Society, Toronto, Ontario, Canada.
- Ozubko, J. D., & Joordens, S. (2011). The similarities (and familiarities) of pseudowords and extremely high-frequency words: Examining a familiarity-based explanation of the pseudoword effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 123–139. <http://dx.doi.org/10.1037/a0021099>
- Parks, C. M., Murray, L. J., Elfman, K., & Yonelinas, A. P. (2011). Variations in recollection: The effects of complexity on source recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 861–873. <http://dx.doi.org/10.1037/a0022798>
- Reyna, V. F., & Brainerd, C. J. (1991). Fuzzy-trace theory and framing effects in choice: Gist extraction, truncation, and conversion. *Journal of Behavioral Decision Making*, 4, 249–262. <http://dx.doi.org/10.1002/bdm.3960040403>
- Reyna, V. F., & Brainerd, C. J. (2011). Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. *Developmental Review*, 31, 180–206.
- Reyna, V. F., Estrada, S. M., DeMarinis, J. A., Myers, R. M., Stanis, J. M., & Mills, B. A. (2011). Neurobiological and memory models of risky decision making in adolescents versus young adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1125–1142. <http://dx.doi.org/10.1037/a0023943>
- Reyna, V. F., & Lloyd, F. (1997). Theories of false memory in children and adults. *Learning and Individual Differences*, 9, 95–123. [http://dx.doi.org/10.1016/S1041-6080\(97\)90002-9](http://dx.doi.org/10.1016/S1041-6080(97)90002-9)
- Roediger, H. L., III, & DeSoto, K. A. (2014). Confidence and memory: Assessing positive and negative correlations. *Memory*, 22, 76–91. <http://dx.doi.org/10.1080/09658211.2013.795974>
- Roediger, H. L., III, Wixted, J. T., & DeSoto, K. A. (2012). The curious complexity between confidence and accuracy in reports from memory. In L. Nadel & W. Sinnott-Armstrong (Eds.), *Memory and law* (pp. 84–117). Oxford, England: Oxford University Press. <http://dx.doi.org/10.1093/acprof:oso/9780199920754.003.0004>
- Seamon, J. G., Luo, C. R., Kopecky, J. J., Price, C. A., Rothschild, L., Fung, N. S., & Schwartz, M. A. (2002). Are false memories more difficult to forget than accurate memories? The effect of retention interval on recall and recognition. *Memory & Cognition*, 30, 1054–1064. <http://dx.doi.org/10.3758/BF03194323>
- Selmeczy, D., & Dobbins, I. G. (2014). Relating the content and confidence of recognition judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 40, 66–85. <http://dx.doi.org/10.1037/a0034059>
- Slovic, P., & Lichtenstein, S. (1983). Preference reversals: A broader perspective. *The American Economic Review*, 73, 596–605.
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117, 34–50. <http://dx.doi.org/10.1037/0096-3445.117.1.34>
- Stone, E. R., Yates, J. F., & Parker, A. M. (1994). Risk communication: Absolute versus relative expressions of low-probability risks. *Organizational Behavior and Human Decision Processes*, 60, 387–408. <http://dx.doi.org/10.1006/obhd.1994.1091>
- Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement*. Washington, DC: United States Department of Justice.
- Tekcan, A. I., & Aktürk, M. (2001). Are you sure you forgot? Feeling of knowing in directed forgetting. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 27, 1487–1490. <http://dx.doi.org/10.1037/0278-7393.27.6.1487>
- Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 1024–1037. <http://dx.doi.org/10.1037/0278-7393.25.4.1024>
- Thierry, K. L., Lamb, M. E., Pipe, M.-E., & Spence, M. J. (2010). The flexibility of source-monitoring training: Reducing young children's source confusions. *Applied Cognitive Psychology*, 24, 626–644. <http://dx.doi.org/10.1002/acp.1574>
- Thomas, A. K., Bulevich, J. B., & Dubois, S. J. (2011). Context affects feeling-of-knowing accuracy in younger and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 96–108. <http://dx.doi.org/10.1037/a0021612>
- Ting, H., & Wallsten, T. S. (2011). A query theory account of the effect of memory retrieval on the sunk cost bias. *Psychonomic Bulletin & Review*, 18, 767–773. <http://dx.doi.org/10.3758/s13423-011-0099-4>
- Titcomb, A. L., & Reyna, V. F. (1995). Memory interference and misinformation effects. In F. N. Dempster & C. J. Brainerd (Eds.), *Interference and inhibition in cognition* (pp. 263–294). San Diego, CA: Academic Press. <http://dx.doi.org/10.1016/B978-012208930-5/50009-X>
- Toglia, M. P., & Battig, W. F. (1978). *Handbook of semantic word norms*. Hillsdale, NJ: Erlbaum.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293–315. <http://dx.doi.org/10.1037/0033-295X.90.4.293>
- Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *The Journal of Business*, 59, S251–S278. <http://dx.doi.org/10.1086/296365>
- Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101, 547–567. <http://dx.doi.org/10.1037/0033-295X.101.4.547>
- Wang, Z., & Busemeyer, J. (2015). Reintroducing the concept of complementarity into psychology. *Frontiers in Psychology*, 6, 1822. <http://dx.doi.org/10.3389/fpsyg.2015.01822>
- Wang, Z., & Busemeyer, J. R. (in press). Comparing quantum versus Markov random walk models of judgments measured by rating scales. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*.
- Wells, G. L., & Murray, D. M. (1984). Eyewitness confidence. In G. L. Wells & E. F. Loftus (Eds.), *Eyewitness testimony: Psychological perspectives* (pp. 155–170). New York, NY: Cambridge University Press.
- Wixted, J. T., Mickes, L., Clark, S. E., Gronlund, S. D., & Roediger, H. L., III. (2015). Initial eyewitness confidence reliably predicts eyewitness identification accuracy. *American Psychologist*, 70, 515–526. <http://dx.doi.org/10.1037/a0039510>
- Wolfe, C. R., & Fisher, C. R. (2013). Individual differences in base rate neglect: A fuzzy processing preference index. *Learning and Individual Differences*, 25, 1–11. <http://dx.doi.org/10.1016/j.lindif.2013.03.003>
- Wolfe, C. R., & Reyna, V. F. (2010). Semantic coherence and fallacies in estimating joint probabilities. *Journal of Behavioral Decision Making*, 23, 203–223. <http://dx.doi.org/10.1002/bdm.650>

Received February 6, 2016

Revision received September 11, 2016

Accepted September 15, 2016 ■