Journal of Experimental Psychology: General

© 2024 American Psychological Association
ISSN: 0096-3445

2024, Vol. 153, No. 4, 1017–1037
https://doi.org/10.1037/xge0001535

Deep Distortions in Everyday Memory: Fact Memory Is Illogical, Too

Charles J. Brainerd, Daniel M. Bialer, Minyu Chang, and Xinya Liu
Department of Psychology and Institute for Human Neuroscience, Cornell University

A distinction has recently been drawn between surface distortions and deep distortions in false memory, where the former are conventional errors of commission and the latter are illogical relations among multiple memories of items. The deep distortions that have been studied to date are violations of the logical rules that govern incompatibility relations, such as additivity and countable additivity. Because that work is confined to laboratory word-list tasks, it is subject to the ecological validity criticism that memory for everyday facts may not exhibit such phenomena. We report evidence that memory for everyday facts displays the same deep distortions as laboratory tasks. We developed a version of the conjoint-recognition paradigm that measures memory for incompatible general knowledge facts, similar to those found on the quiz program Jeopardy! In experiments with university participants, four deep distortions were detected (violations of the additivity, countable additivity, universal set, and compensation rules), with participants consistently remembering more than what is logically possible. The distortions were more robust than in laboratory experiments, and memories of incompatible facts (e.g., Jupiter and Saturn cannot both be the largest planet in the solar system) did not suppress each other. These patterns were replicated in subsequent experiments with older and more diverse participant samples. Consistent with the notion that deep distortions are by-products of gist memory, conjoint-recognition modeling analyses revealed that memory for everyday facts was even more reliant on gist than memory for word lists, and that verbatim memory was near-floor.

Public Significance Statement
When events are physically incompatible, they obey some inviolate logical constraints: You cannot simultaneously stand on your head and on your feet, nor can you simultaneously turn left and turn right when you arrive at an intersection. Do our memories of incompatible events obey these logical rules? Fuzzy-trace theory predicts that the answer is no because reliance on semantic gist causes events to be over remembered. That prediction has been confirmed in laboratory word-list experiments, but does it hold for real-world memories of meaningful facts? We investigated this question in memory for everyday facts that appear on quiz programs such as Jeopardy! Logical rules such as additivity, countable additivity, universal event, and compensation were all violated by wide margins—the reason being that memory was even more reliant on semantic gist than in word-list experiments.

Keywords: deep distortions, gist memory, incompatibility, memory for everyday facts, fuzzy-trace theory

Supplemental materials: https://doi.org/10.1037/xge0001535.supp

Although the study of memory distortion is a staple of mainstream memory research, the findings of such work have a remarkably wide scope of application in psychology and cognate fields. This can be seen to good advantage in high-stakes applications, where memory distortion can have very unfortunate consequences. Familiar examples are eyewitness identification of criminal suspects (e.g., Lampinen et al., 2019), recovery of repressed memories of trauma (e.g., McNally et al., 2001), criminal interrogations (e.g., Kassin,

This article was published Online First February 1, 2024.

Charles J. Brainerd https://orcid.org/0000-0002-8358-7497

The research reported in this article was supported by United States Department of Agriculture Grant (NYC-321444). The deep distortion in everyday memory instrument and the raw data of the experiments are provided in the online supplemental materials. The research that we report has not been previously disseminated.

Charles J. Brainerd served as lead for conceptualization, formal analysis, funding acquisition, resources, and writing–original draft. Daniel M. Bialer served as lead for data curation and investigation. Minyu Chang served in a supporting role for investigation, methodology, project administration, and software. Xinya Liu served in a supporting role for data curation, formal analysis, investigation, software, and supervision. Charles J. Brainerd and Daniel M. Bialer contributed equally to methodology and supervision. Charles J. Brainerd, Daniel M. Bialer, Minyu Chang, and Xinya Liu contributed equally to writing–review and editing.

Correspondence concerning this article should be addressed to Charles J. Brainerd, Department of Psychology and Institute for Human Neuroscience, Cornell University, G331 MVR Hall, Ithaca, NY 14853, United States. Email: cb299@cornell.edu

2005), medical histories of severe injuries (e.g., Nachson & Slavutskay-Tsukerman, 2010), and diagnosis of neurocognitive diseases (e.g., Brainerd et al., 2014).

Forgetting and false memory are customarily regarded as the prime examples of memory distortion. Both are vast fields of inquiry, with their own distinct paradigms and effects, but they share a common feature: Each is a failure in memory for individual items (their content and contexts), which we simply refer to as facts, rather than a failure in memory for relations that connect facts. More explicitly, they are errors of omission or commission in memory for the content of specific facts, such as eating a salad and filling the car with gas, and memory for the contexts in which they occurred, such as lunch and on the way home from work.

In the past decade, a third form of distortion has been studied in which memory fails to preserve the relations that connect facts (for a review, see Brainerd, 2021). This is the realm of deep distortions. The examples that have been investigated to date involve a very basic class of relations—namely, incompatibility relations. An incompatibility relation is a concept of some sort (e.g., spatial direction, death, birth) that specifies a set of facts whose members cannot exist simultaneously in the real world. For instance, consider the spatial maneuvers that vehicles can execute upon arriving at an intersection. Any specific vehicle, say vehicle A, cannot simultaneously turn left, turn right, and drive straight ahead because each maneuver is incompatible with the other two—the fact of vehicle A turning left is incompatible with the fact of it simultaneously turning right and the fact of it driving straight ahead. Similarly, consider accidental causes of mortality. Any specific person, say person A, cannot simultaneously die from asphyxiation, a gunshot wound to the head, and a crushed spine from a car crash; each rules out the others.

As incompatible facts cannot coexist, their real-world probabilities are perfectly compensatory. Across all vehicles arriving at an intersection, the probability of a left turn increases only to the extent that the probability of other incompatible maneuvers decreases in direct proportion, and conversely. Across all accidental causes of mortality, the probability of asphyxiation increases only to the extent that the probability of other incompatible causes of mortality decreases in direct proportion, and conversely. Hence, any set of incompatible facts forms a tradeoff structure in which increases in the probability of any one of them are exactly compensated by decreases in the probability of some of the others.

This feature of incompatibility relations means that they necessarily exhibit certain logical properties, which are captured in the familiar axioms of probability theory. Three such axioms—additivity, countable additivity, and universal event—and two of their corollaries—monotonicity and null event—are displayed in Table 1. Even a cursory inspection of Table 1 reveals that any real-world incompatibility relation must obey all three axioms and, therefore, the corollaries as well. With respect to the axioms, it is obvious that any record of vehicle maneuvers at an intersection will show that (a) the proportion of vehicles that turned left plus the proportion that turned right will exactly equal the proportion that turned left or right (additivity); (b) the proportion of vehicles that turned left plus the proportion that turned right plus the proportion that continued straight ahead will exactly equal the proportion that turned left or right or continued straight ahead (countable additivity); (c) the proportion of vehicles that turned left plus the proportion that turned right plus the proportion that continued straight ahead will be <1 (universal event), as vehicles can make other incompatible maneuvers, such as a U turn.

What about memory? Does it preserve such logical properties? If it does, the probabilities of remembering incompatible facts will conform to the rules in Table 1. In a review of several data sets in which one or more of those rules could be tested, memories of incompatible facts did not conform to any of them (Brainerd, 2021). Specifically, (a) the sum of the probabilities of remembering that each of two incompatible facts was true was often greater than the probability of remembering that one or the other was true (violating additivity); (b) the sum of the probabilities of remembering that each of three incompatible facts was true was always greater than the probability of remembering that one or the other of them was true (violating countable additivity); (c) the sum of the probabilities of remembering that each of three incompatible facts was true was always >1 (violating universal event); (d) the probability of remembering that a fact was true was sometimes greater than the probability of remembering that either that fact or some other incompatible fact was true (violating monotonicity); and (e) the probability of remembering that two incompatible facts were both true was >0 (violating null set).

Although such findings seem to make a strong case for the existence of deep distortions, the literature review identified two limitations of the experiments that produced the findings, which raise doubts about their generality. First, violations of the logical rules that govern incompatibility relations have only been detected with artificially created facts about word lists. Second, the property of incompatibility relations on which all the others depend, has not been extensively tested. Concerning the first limitation, Brainerd (2021) noted that extant studies of deep distortions provide no evidence that these phenomena extend beyond laboratory experiments with word lists to the important domain of memory for everyday facts. That domain encompasses the information in our shared human knowledge base, and it includes memory for socially and historically notable facts (e.g., the birthplace of Jeanne d'Arc, the date of Gandhi's assassination, the first emperor of Rome) and popular culture facts (e.g., the world's youngest billionaire, the first food eaten in space, the last winner of the World Cup). Concerning the other limitation, deep distortion studies have only occasionally investigated the property from which the other properties in Table 1 follow—namely, that the occurrence probabilities of incompatible facts are mutually compensatory—and only with laboratory word list data (e.g., Brainerd et al., 2015; Brainerd & Reyna, 2018). For memory to preserve that property, there must be strong negative correlations among the probabilities of remembering individual incompatible facts.

The experiments that we report were designed to address both these limitations. To determine whether deep distortions are more than laboratory curiosities, we investigated whether memory for everyday facts that are targeted in board games and quiz programs such as Jeopardy! violates any of the three axioms in Table 1. To evaluate the mutual compensation rule, we investigated correlations among the probabilities of remembering incompatible facts (e.g., that Einstein was born in Germany vs. Austria vs. Switzerland). Before reporting the experiments, we provide some necessary background by briefly sketching the theoretical impetus for deep distortion research, the paradigm that has been implemented in such research, and some illustrative findings.

Table 1

Some Laws of Probability Theory

Property	Definition
Axioms	
Additivity	If A and B are incompatible facts, the sum of their respective probabilities equals the probability of their disjunction.
Countable additivity	If A, B, C, …, N are a series of incompatible facts, the sum of their respective probabilities equals the probability of their disjunction.
Universal event	If U is a universal event that contains all the events in some space of incompatible facts, the sum of the probabilities of any subset of those facts is ≤ 1 and the sum of all their probabilities is exactly 1.
Corollaries	
Monotonicity	If the fact A contains the fact B, the probability of B approaches the probability of A as a limit. This law depends on the countable additivity axiom; that is, monotonicity does not hold if countable additivity is violated.
Null event	If Ø is a null fact, its probability is exactly zero. This corollary depends on the countable additivity axiom; that is, Ø $\neq$ 0 if countable additivity is violated.

Deep Distortion Research

Theoretical Motivation

Deep distortion experiments were stimulated by fuzzy-trace theory's (FTT) opponent-process account of false memory (Brainerd & Reyna, 2005), which posits that people store dissociated episodic traces of the literal form of experience (verbatim traces) and of some of the salient properties it shares with other experiences (gist traces). Thus, if people encode the words mad, fear, hate, temper, rest, awake, tired, dream, table, couch, desk, and sofa, they are assumed to store episodic traces of the presentations of individual items (verbatim) and episodic traces of some shared properties, such as "unpleasant," "sleep," and "furniture" (gist). The two types of traces sometimes reinforce each other and sometimes oppose each other. For example, on memory tests, both point to tired and desk being on the list. If the test words are snooze and seat, however, gist traces point to them being on the list, but verbatim traces point to them being new.

Brainerd et al. (2010) proposed that the notion of gist traces implies that episodic memory will not invariably obey the core compensation rule of incompatibility relations—for instance, falsely remembering that snooze and seat are old may not be incompatible with remembering that they are new or correctly remembering that tired and desk are old may not be incompatible with remembering that they are new. Here, note that gists such as "sleep" and "furniture" are consistent with these incompatible memories because the gists encompass all facts that exemplify their meaning, regardless of whether the facts are incompatible—as when "sleep" encompasses both tired (old) and snooze (new) and "furniture" encompasses both desk (old) and seat (new). It is this feature of gist that predicts the deep distortions that have been observed in laboratory experiments, and it obviously predicts that their robustness will vary as a function of how strongly episodic memory relies on gist.

Conjoint Recognition

Deep distortions have been measured with the conjoint-recognition paradigm—a modification of conventional false memory designs in which participants make three types of memory judgments about test items rather than only making old judgments. The judgments are old (previously encoded on a study list), similar (new but resembling encoded items), and either old or similar. These three judgments are factorially crossed with three types of test items: old, similar, and new. This procedure has been widely used to track the types of traces that are retrieved on various memory and reasoning tasks. Among the questions that have been investigated are the representational content of working memory (e.g., Abadie & Camos, 2019; Abadie et al., 2013), the representational content that controls reasoning illusions (e.g., Abadie et al., 2017; Didyk & Nieznański, in press), how people suppress false memories (e.g., Lampinen et al., 2005, 2006), the representational content of associative memory (e.g., Greene et al., 2022; Greene & Naveh-Benjamin, 2022), loss and sparing of different representations during aging (Greene & Naveh-Benjamin, 2020, 2023), how emotion influences memory (Bookbinder & Brainerd, 2017; Gong et al., 2016), the representational content of linguistic inferences (Singer & Remillard, 2008; Singer & Spear, 2015), and how retrieval is affected by encoding and ability variables (Nieznański et al., 2019; Obidziński, 2021; Obidziński & Nieznański, 2017). The features of conjoint recognition that allow it to determine whether memory obeys the logical rules in Table 1 are that it creates a set of mutually incompatible facts about items on word lists (each can be old, similar, or new), and it includes disjunctive judgments about those facts (old-or-similar judgments).

Because the three types of items are factorially crossed with the three types of judgments, the quantities p(old), p(similar), and p(old-or-similar) can be separately calculated for each of the three types of items. That supplies the necessary degrees of freedom to test the additivity property for the incompatible facts "old" and "similar," by evaluating whether the equality $p(\text{old}) + p(\text{similar}) = p(\text{old-or-similar})$ is always satisfied. Next, the universal event property is evaluated via a small modification to the trio of memory judgments, in which old-or-similar judgments are replaced by judgments that ask whether test items are new (Brainerd et al., 2015). Now, the quantities p(old), p(similar), and p(new) are computed, and the rule $p(\text{old}) + p(\text{similar}) + p(\text{new}) = 1$ is evaluated.

Next, the null set rule can be evaluated by administering "old" and "similar" tests for individual items and determining whether the probability of judging that both descriptions are true is >0 (Brainerd et al., 2020). The monotonicity rule can be tested by simply determining whether the probability of judging an item to be old or the probability of judging an item to be similar is greater than the probability of judging it to be old-or-similar (Brainerd et al., 2019). Finally, the countable additivity rule can be studied by adjusting the basic design so that there are three or more incompatible facts for each test item (Brainerd et al., 2012).

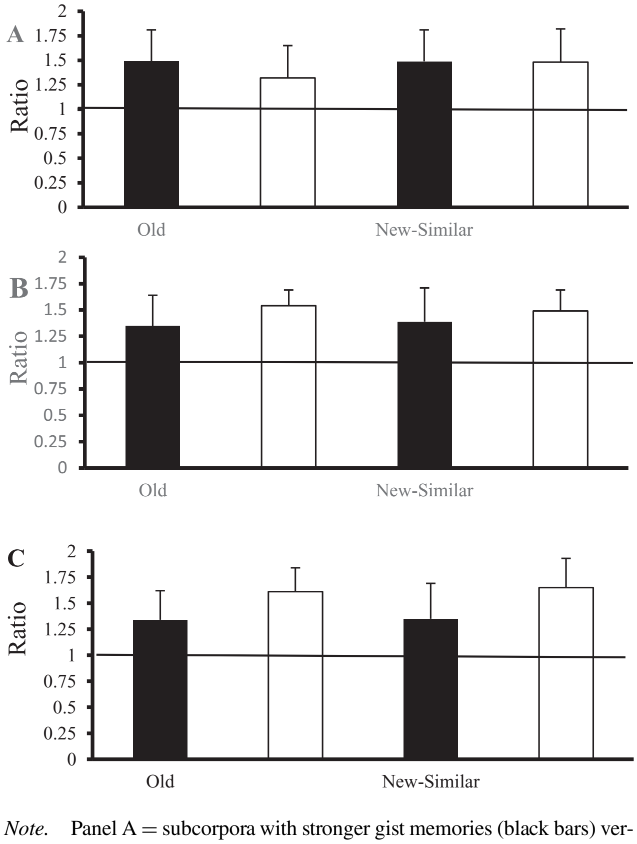
Principal Findings

The picture that has emerged from conjoint-recognition experiments is that episodic memory does not consistently obey any of the rules in Table 1. Taking additivity first, overall results for tests of $p(\text{old}) + p(\text{similar}) = p(\text{old-or-similar})$, which can be conveniently expressed as the ratio $[p(\text{old}) + p(\text{similar})] \div p(\text{old-or-similar})$, are displayed in Figure 1 for a corpus of 614 data sets. This ratio must equal 1 if additivity holds, which it clearly does not. This corpus can be partitioned in three ways, to yield a pair of subcorpora that each contain >150 data sets. Regardless of how the corpus is partitioned, the pattern in Figure 1 is the same: Average ratios are all substantially greater than 1. Mathematically, then, $p(\text{old}) + p(\text{similar})$ is subadditive, relative to $p(\text{old-or-similar})$. Psychologically, participants remembered more than logic permits when making incompatible memory judgments about old and similar items.

Turning to the universal event rule, when new judgments replaced old-or-similar judgments, memory also failed to obey this rule. For average values of $p(\text{old})$, $p(\text{similar})$, and $p(\text{new})$ for old and similar items, the pattern was $p(\text{old}) + p(\text{similar}) + p(\text{new}) > 1$. For the data sets in Brainerd's (2021) review, the mean sums were 1.27 and 1.32 for old and similar items, respectively. As with tests of the additivity rule, then, participants remembered more than what is logically possible.

Figure 1

Violations of the Additivity Property by Old and Similar Items in a Corpus of 614 Data Sets



Note. Panel A = subcorpora with stronger gist memories (black bars) versus weaker gist memories (white bars). Panel B = subcorpora with immediate tests (black bars) versus delayed tests (white bars). Panel C = subcorpora with between- (black bars) versus within-participant (white bars) variation of conjoint-recognition judgments. Ratio = $[p(\text{old}) + p(\text{similar})] \div p(\text{old-or-similar})$, for both old and new-similar items.

Continuing to countable additivity, in experiments with four episodic states (old1, old2, old3, and new), overall results for tests of $p(\text{old1}) + p(\text{old2}) + p(\text{old3}) = p(\text{old1-or-old2-or-old3})$ can be expressed as the ratio $[p(\text{old1}) + p(\text{old2}) + p(\text{old3})] \div p(\text{old1-or-old2-or-old3})$. Although this ratio must be 1 for countable additivity to be satisfied, its average value was 2.40 for the data sets that Brainerd (2021) reviewed. Psychologically, participants once again remembered more than what is logically possible, given that the different judgments tapped memory for incompatible facts.

The two corollaries in Table 1 have also been violated in conjoint-recognition experiments. The overall pattern for the monotonicity rule is displayed in Figure 2. There, the normalized frequency distributions of the statistic $p(\text{old}) - p(\text{old-or-new})$ for old items and the statistic $p(\text{similar}) - p(\text{old-or-similar})$ for new-similar items have been plotted for the complete recognition corpus. Monotonicity requires that $p(\text{old}) \leq p(\text{old-or-similar})$ and $p(\text{similar}) \leq p(\text{old-or-similar})$, so that both curves must fall to the left of the 0 point. Instead, notice that more than half the old item distribution falls to the right of 0 [the $p(\text{old}) > p(\text{old-or-new})$ region], and so does roughly 20% of the new item distribution [the $p(\text{similar}) > p(\text{old-or-new})$ region].

Finally, with respect to the null event rule, an item cannot simultaneously be both old and similar—that is an impossible conjunction of facts. However, this property has been violated in experiments in which participants made both old and similar judgments about individual old and similar test items. Across the data sets reviewed by Brainerd (2021), 23% of old items were judged to be simultaneously old and similar, and 19% of similar items were judged to be simultaneously old and similar.

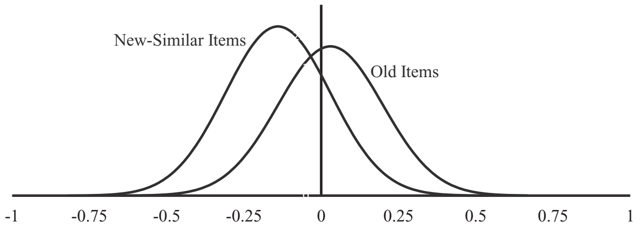
The Present Research

The experiments we report below were focused on the previously noted limitations of deep distortion studies—namely, that they rely on artificially created incompatible facts about word lists and that the key principle of mutual compensation among memories of incompatible facts has rarely been investigated. Before reporting the experiments, we briefly comment on the significance of each limitation and how it was addressed.

The significance of the first limitation is that deep distortion findings are susceptible to ecological validity arguments. Memory violations of logical rules have only been detected with artificial facts about word lists, and hence, they may be mere laboratory curiosities. By comparison, memory for everyday facts may obey all the logical

Figure 2

Normalized Frequency Distributions of the Statistic p*(old) −* p*(old-or-new) for Old Items and the Statistic* p*(similar) −* p*(old-or-similar) for New-Similar Items in the Conjoint-Recognition Corpus*



rules of real-world incompatibility relations. As discussed elsewhere (Brainerd, 2021), that possibility is consistent with ecological validity approaches to memory (see Banaji & Crowder, 1989; Diamond et al., 2020; Unsworth et al., 2013; White, 2021). Such approaches emphasize that the content of memories of everyday facts and the conditions under which they are formed are fundamentally different than the corresponding content and conditions of laboratory memories of facts about word lists. For those reasons, ecological validity approaches stress that memories of everyday facts can fail to exhibit the same phenomena as laboratory memories of word lists. Hence, such memories may preserve the logic of incompatibility relations, despite evidence to the contrary from word-list experiments.

There is theoretical disagreement on this point, however, because FTT expects that memory for everyday facts will also fail to obey the logic of incompatibility relations. FTT's basis for predicting deep distortions in the first place is that people often respond to memory probes by relying on gist, which we saw is noncompensatory. That includes probes for memories of everyday facts (see Liberali et al., 2012; Singer & Remillard, 2008; Singer & Spear, 2015). Thus, there are theoretical grounds for expecting deep distortions in memory for everyday facts as well as in laboratory word-list tasks, although the strength of these phenomena will differ between the two spheres if the degree of reliance on gist memories differs.

To remove the first limitation, our experiments were concerned with memory for everyday facts. They were analogues of traditional conjoint-recognition designs, in which we measured the extent to which memory obeyed the additivity, countable additivity, and universal event rules for 100 sets of incompatible facts generated by concepts that are used regularly in questions on quiz programs such as Jeopardy!

The significance of the second limitation of prior deep distortion studies is that the untested compensation rule is more fundamental than the other logical rules in Table 1: Compensation among the occurrence probabilities of incompatible facts is a necessary condition for all these rules. It is especially important, then, to build a solid data base on compensation. To do that, we analyzed correlations among the remembering probabilities of the sets of incompatible facts that figured in our experiments, across participants and across the sets of facts.

The overall design of the research runs as follows. We conducted four experiments in which the additivity, countable additivity, universal event, and compensation properties were evaluated in memory for sets of incompatible everyday facts. The first two experiments involved undergraduate participant samples, whereas the last two experiments involved participants from older and more diverse populations. The aim of Experiment 1 was to determine (a) whether memories of everyday facts obey the three logical rules that have been most often violated with word lists (additivity, countable additivity, and universal event), and (b) whether those memories exhibit strong mutual compensation. Experiment 2 was concerned with the same questions, but it evaluated whether conformity to these logical rules can be fomented by a manipulation that is known to do so in word-list experiments. There, it has been found that making confidence judgments as part of recognition tests shifts retrieval away from gist memory toward verbatim memory and virtually eliminates violations of additivity, countable additivity, and universal event. Consequently, confidence judgments figured in tests of memory for everyday facts in Experiment 2.

As Experiments 1 and 2 were the first to measure deep distortions in memory for everyday facts, it seemed advisable to replicate all aspects of their results with larger and more diverse participant samples. Experiment 3 was a replication of Experiment 1 and Experiment 4 was a replication of Experiment 2, but with very different participant samples. The aim was simply to determine whether the patterns that were obtained in the first two experiments were also present in participants whose age, background, and ability were different than our undergraduate samples.

Finally, after the results of the individual experiments are presented, all the data are analyzed with the conjoint-recognition model. This model provides quantitative estimates of the relative contributions of verbatim retrieval, gist retrieval, and response bias to memory for incompatible everyday facts. We shall see that differences in the values of the model's parameters deliver tests of process explanations of deep distortions.

Experiment 1

Method

Participants

The participants in Experiment 1 were 116 undergraduates who enrolled to fulfill a course requirement, and who received extra credit for participating. Their mean age was 19.85 years (range = 18–22), and the sample was 80% female, 19% male, and 1% nonbinary. A sensitivity analysis of the data of prior experiments (Brainerd, 2021) indicated that a sample size of 90 provided sufficient power ($1 - \beta = .80$) to detect medium-size effects for the variables of interest to us. However, all individuals who enrolled for each experiment were allowed to participate—hence, the eventual sample size of 116.

Materials

The materials for each experiment were 100 sets of incompatible facts from categories that are routinely used in Jeopardy! questions—including historical events, geography, nutrition, popular culture, science, sports, literature, film, and television. They were assembled in a structured assessment, the distortion in everyday memory (DEM) instrument. Each DEM item consists of a concept that maps with some significant fact, plus three candidates for that fact. To conform to the notation in Table 1, the candidates are denoted A, B, and C, where A is the target fact and B and C are plausible alternatives that participate in a core concept that specifies an incompatibility relation among the three alternatives. Six different memory probes are administered for each of the 100 items: One conventional probe for each of the three candidates (Is A true? Is B true? Is C true?) and one disjunctive probe for each of the three possible pairwise disjunctions of the candidates (Is either A or B true? Is either B or C true? Is either C or A true?)

Examples of DEM items from the popular culture, historical events, and science categories are displayed in Table 2. The full instrument is provided in the online supplemental materials that accompanies this article.

Procedure

The general conjoint-recognition procedure, in which participants' memories are tested for individual candidates and for disjunctive combinations of pairs of candidates, was used in all four experiments. In Experiment 1, the participant was seated at a

Table 2
Illustrative Items From the Distortion in Everyday Memory Instrument

Item	Content details
Item 1	
Popular culture	The city in which Starbuck's opened its first store
Candidates	A = Seattle; B = Portland; C = Los Angeles
Memory probes	A? = Starbucks opened its first store in Seattle
	B? = Starbucks opened its first store in Portland
	C? = Starbucks opened its first store in Los Angeles
	A or B? = Starbucks opened its first store in Seattle or Portland
	B or C? = Starbucks opened its first store in Portland or Los Angeles
	C or A? = Starbucks opened its first store in Los Angeles or Seattle
Item 2	
Historical events	The first U.S. president to face impeachment
Candidates	A = Andrew Johnson; B = Richard Nixon; C = Bill Clinton
Memory probes	A? = The first U.S. president to face impeachment was Andrew Johnson
	B? = The first U.S. president to face impeachment was Richard Nixon
	C? = The first U.S. president to face impeachment was Bill Clinton
	A or B? = The first U.S. president to face impeachment was Andrew Johnson or Richard Nixon
	B or C? = The first U.S. president to face impeachment was Richard Nixon or Bill Clinton
	C or A? = The first U.S. president to face impeachment was Bill Clinton or Andrew Johnson
Item 3	
Science	The largest planet in the solar system
Candidates	A = Jupiter; B = Earth; C = Saturn
Memory probes	A? = Jupiter is the largest planet in our solar system
	B? = Earth is the largest planet in our solar system
	C? = Saturn is the largest planet in our solar system
	A or B? = Jupiter or Earth is the largest planet in our solar system
	B or C? = Earth or Saturn is the largest planet in our solar system
	C or A? = Saturn or Jupiter is the largest planet in our solar system

computer and began by reading generic memory instructions that appeared on the screen. The instructions explained that the participant's memory for a series of everyday facts would be tested and that the task was simply to indicate whether they remembered each stated fact as being correct. This was followed by a short series of example items.

Next, the 100 DEM items were individually presented in random order. For each item, one of the six memory probes was randomly selected and presented. Only one probe was presented per item because otherwise, a participant's memory for the response to an earlier probe (e.g., responding yes to "Saturn is the largest planet in our solar system") could influence the response to a later probe (e.g., encouraging no to "Jupiter is the largest planet in our solar system"). This method of measuring memory for Jeopardy! facts was common to all four experiments. Because there are 100 DEM items and six memory probes, individual participants could not be assigned an equal number of each type of probe. Instead, individual participants responded to 17 instances of four probe types (total = 68 probes) and 16 instances of the remaining two probe types (total = 32 probes). Which probe types were represented 17 times and which were represented 16 times were varied across participants.

Participants responded to all 100 DEM items, with the factorial design being 2 (type of probe: conventional vs. disjunctive) × 3 (candidate fact: A, B, or C). Each participant was randomly assigned to one of six random orderings of the 100 DEM items and random configurations of the six memory probes. All 100 probes were presented individually, centered on the computer screen, and printed in black font. Two radio buttons appeared below the probe, which the subject used to indicate whether the stated fact was remembered as correct. The participant was given 10 s to respond to each probe.

Analyses

To avoid repetition, some brief comments are necessary about the statistical analyses that were used in all experiments to detect violations of the additivity, countable additivity, and universal event rules. For the candidate facts A, B, and C, recall that A denotes the correct candidate, and B and C denote incorrect candidates. Memory for these Jeopardy! facts obeys additivity if $p(\text{A}) + p(\text{B}) = p(\text{A-or-B})$, $p(\text{B}) + p(\text{C}) = p(\text{B-or-C})$, and $p(\text{C}) + p(\text{A}) = p(\text{C-or-A})$. Algebraically, these equalities are equivalent to $[p(\text{A}) + p(\text{B})] \div p(\text{A-or-B}) = [p(\text{B}) + p(\text{C})] \div p(\text{B-or-C}) = [p(\text{C}) + p(\text{A})] \div p(\text{C-or-A}) = 1$. As ratios are more reliable measures than difference scores (e.g., Uhl et al., 2022), we evaluated whether memory for these Jeopardy! facts preserved the additivity property by simply testing the null hypothesis that none of these ratios differed from 1. Turning to countable additivity, memory preserves it if the following equality is satisfied: $p(\text{A}) + p(\text{B}) + p(\text{C}) = p(\text{A-or-B-or-C})$. Algebraically, this equality is equivalent to $p(\text{A}) + p(\text{B}) + p(\text{C}) = p(\text{A-or-B}) + p(\text{B-or-C}) - p(\text{B}) = p(\text{A-or-B}) + p(\text{C-or-A}) - p(\text{A}) = p(\text{C-or-A}) + p(\text{B-or-C}) - p(\text{C})$, from which the following ratios must hold: $[p(\text{A}) + p(\text{B}) + p(\text{C})] \div [p(\text{A-or-B}) + p(\text{B-or-C}) - p(\text{B})] = [p(\text{A}) + p(\text{B}) + p(\text{C})] \div [p(\text{A-or-B}) + p(\text{C-or-A}) - p(\text{A})] = [p(\text{A}) + p(\text{B}) + p(\text{C})] \div [p(\text{C-or-A}) + p(\text{B-or-C}) - p(\text{C})] = 1$. Hence, whether memory for these Jeopardy! facts preserves the countable additivity property was evaluated by simply testing the null hypothesis that none of these ratios differed from 1.

The universal event property specifies that the sum of the occurrence probabilities of a series of incompatible events cannot exceed 1 and that the sum must be <1 if the series is not exhaustive. Because the DEM candidates A, B, and C are not an exhaustive series for any of the incompatibility relations, whether memory for these Jeopardy! facts preserves the universal event property was evaluated by simply testing the null hypothesis that $p(A) + p(B) + p(C) < 1$. The three candidates are not an exhaustive series because there are additional incompatible candidates (i.e., D, E, …) for each DEM item. For instance, consider the last example in Table 2. Although remembering that Jupiter is the largest planet in the solar system is incompatible with simultaneously remembering that Earth or Saturn is the largest planet, there are at least five other planets that are sources of such incompatible memories. Consequently, $p(\text{Jupiter}) + p(\text{Earth}) + p(\text{Saturn}) < 1$ if memory obeys the universal event property.

Last, in all our experiments, the statistical analyses that were used to detect violations of the additivity, countable additivity, and universal event rules could be conducted in two ways: (a) at the level of individual DEM facts, where the acceptance probabilities for a fact's six probes are estimated by aggregating data across the responses of individual participants to each probe and (b) at the level of individual participants, where these acceptance probabilities are estimated by aggregating data across the responses to each probe type for the 100 DEM facts for each participant. When we evaluated deep distortions at the fact and participant levels, the respective patterns were the same. Consequently, to avoid repetition, the detailed results that we report for the three logical rules in all experiments, whether descriptive statistics or significance tests, are for the fact level. Violations of the additivity, countable additivity, and universal event properties are tested for statistical significance three times apiece in each of our experiments. To avoid alpha slippage, the .01 level of confidence, rather than the traditional .05 level, was used for all those tests.

Readers may think that it would also be desirable to conduct these analyses in a third way, too—namely, at the Participant × Fact Level; that is, at the level of each participants' responses to all six probes for each individual DEM fact. However, as discussed in prior psychometric analyses of levels of measurement in memory research (e.g., Jacoby & Shrout, 1997), it is impossible to measure effects at this third level. That is because the only way to estimate relevant quantities at that level (i.e., the probabilities of each individual participant accepting each of the six probes for each fact) is to administer all the probes for each fact repeatedly to each participant. However, that procedure contaminates the data with powerful repeated-testing artifacts.

Transparency and Openness

Consistent with this journal's editorial policy, this article was prepared in accordance with Level 2: Requirement of the Transparency and Openness Promotion Guidelines. The DEM instrument and the raw data of the present experiments are contained in the online supplemental materials. The software and instructions for analyzing data with the conjoint-recognition model may be found at https://osf.io/p8uh6/.

Results

Deep distortion findings are presented in two waves. First, we report findings for the additivity, countable additivity, and universal event rules of incompatibility relations. Second, we report results for the previously untested compensation rule.

Additivity, Countable Additivity, and Universal Event

As an advance organizer, there were two major patterns: (a) pronounced deep distortions for all three logical rules, and (b) a common psychological theme to these distortions—namely, subadditivity (memory probabilities always exceeded the values that logic allows). These patterns can be seen by inspecting the data in the first column of Table 3, which displays the three deep distortion measures for the DEM that were described above.

Before analyzing the results in Table 3, a preliminary question that must be answered is whether the correct facts in the trios of incompatible candidates for DEM items are part of the long-term memories of this participant population. Note, first, that $p(A)$, the probability of remembering that the correct candidate was indeed correct was .69. When we surveyed a sample of Jeopardy! programs broadcast in 2022, this value was in the difficulty range of $400 facts, which means that the average difficulty of DEM items is in the lower range of difficulty for Jeopardy! facts. Second, if these facts are part of the long-term memories of the majority of participants, $p(A)$ will be greater than either $p(B)$ or $p(C)$, and $p(\text{B-or-C})$ will be smaller than either $p(\text{A-or-B})$ or $p(\text{C-or-A})$. The descriptive statistics for these six measures appear in the first column of Table 4, and the observed mean judgment probabilities obviously conform to these two patterns. To test them for statistical reliability, we computed a 2 (probe type: conventional vs. disjunctive) × 3 (candidate: A, B, C) analyses of variance (ANOVAs). The results confirmed both patterns. More explicitly, there was a large Probe Type × Candidate interaction, $F(2, 198) = 45.35$, $MSE = .023$, $p < .001$, $\eta_p^2 = .31$, in addition to main effects for probe type and candidate. When this interaction was decomposed with post hoc tests, the order of the judgment probabilities for the three conventional probes was $p(A) > p(B) > p(C)$, where C is the false fact that was less often accepted. Consistent with that ordering, the order of the judgment probabilities for the three disjunctive probes was $p(\text{B-or-C}) < p(\text{A-or-B}) = p(\text{C-or-A})$.

Returning to the deep distortion measures in Table 3, The additivity, countable additivity, and universal event rules were all violated by wide margins. With respect to additivity, remember that if the sum of the probabilities of remembering each of two incompatible

Table 3

Deep Distortion Measures for Experiments 1 and 2

	Experiment	
Distortion measures	1	2
Additivity		
$p(A) + p(B)]/p(\text{A-or-B})$	1.45	1.68
$[p(B) + p(C)]/p(\text{B-or-C})$	1.52	1.69
$[p(A) + p(C)]/p(\text{C-or-A})$	1.48	1.59
Countable additivity		
$[p(A) + p(B) + p(C)]/[p(\text{A-or-B}) + p(\text{B-or-C}) - p(B)]$	1.83	2.21
$[p(A) + p(B) + p(C)]/[p(\text{A-or-B}) + p(\text{C-or-A}) - p(A)]$	1.96	2.65
$[p(A) + p(B) + p(C)]/[p(\text{C-or-A}) + p(\text{B-or-C}) - p(C)]$	1.82	1.96
Universal set		
$p(A) + p(B) + p(C)$	1.54	1.42
$p(A) + p(B)$	1.13	1.08
$p(A) + p(C)$	1.10	1.00

Note. A = remembering the correct candidate fact for an item; B = remembering the first incorrect candidate fact for an item; C = remembering the second incorrect candidate fact for an item.

Table 4

Mean Probabilities and Standard Deviations for the Six Memory Judgments in the Distortion in Everyday Memory Instrument in Experiments 1 and 2

	Memory judgment					
	Conventional probes			Disjunctive probes		
Experiment	A	B	C	A or B	B or C	C or A
Experiment 1						
M	.69	.44	.41	.78	.56	.76
SD	.19	.21	.20	.16	.21	.16
Experiment 2						
M	.66	.43	.34	.70	.50	.67
SD	.20	.21	.19	.20	.20	.19

Note. A = remembering the correct candidate fact for an item; B = remembering the first incorrect candidate fact for an item; C = remembering the second incorrect candidate fact for an item.

candidate facts equals the probability of remembering their disjunction, none of the ratios in the left column of Table 3 will differ reliably from 1. Obviously, however, they do, with all values being far larger than 1. Their grand average is 1.48. One-sample *t* tests established that all these ratios were significantly different from 1, $t(99) = 12.93$, 9.53, and 11.47 (all $ps < .001$).

In all instances, memory judgments violated additivity by exceeding the required value; that is, the judgments were subadditive. Subadditivity has a straightforward psychological meaning: Because the facts are incompatible, people endorse more facts than can possibly be true (Abadie & Camos, 2019; Greene & Naveh-Benjamin, 2020; Lampinen et al., 2006; Nieznański, 2020).

Continuing to countable additivity, note that the ratios in Table 3 that bear on this rule are also >1. Also as before, one-sample *t* tests showed that all of them were significantly different from 1, $t(99) = 11.53$, 10.75, and 11.04 (all $ps < .001$). When these results are combined with those that were just reported, participants' memory for Jeopardy! facts was wholly nonadditive. Because the values of all the additivity and countable additivity measures were >1, memory for incompatible candidate facts was consistently subadditive. As before, the psychological meaning of this pattern is that people endorsed more facts than can possibly be true. Note, too, that the level of deep distortion was even greater for countable additivity than for additivity: The mean value of the ratios increased by 27%, from 1.48 for additivity to 1.87 for countable additivity, which is a significant increase ($p < .005$).

Next, consider the universal event measure, $p(A) + p(B) + p(C)$. Unlike additivity and countable additivity, the predicted value is <1 rather than 1 because, as we saw, A, B, and C are not an exhaustive series of candidates for Jeopardy! facts. No statistical test is required to reject that prediction because the observed value of $p(A) + p(B) + p(C)$ in the first column of Table 3 is 1.54. A one-sample *t* test showed that this value was significantly >1, $t(99) = 14.80$, $p < .001$.

An especially remarkable finding that illustrates the strength of this third deep distortion is that memory overflowed the unit interval to such a degree that violations of the universal event rule even occurred for subquantities of the $p(A) + p(B) + p(C)$ measure—specifically, for $p(A) + p(B)$ and $p(A) + p(C)$. It can be seen in the last two rows of Table 3 that neither of these quantities was <1, when they should be far below considering that only two incompatible candidates were involved. Moreover, both subquantities were significantly >1, $t(99) = 4.98$ and 3.64 (both $ps < .001$), respectively. These results echo the subadditivity theme that emerged in the additivity and countable additivity analyses. Psychologically, by every measure, participants endorsed more facts as true than is logically possible.

Compensation Among Incompatible Memories

The three candidates for each of our everyday facts are mutually incompatible in the real world—each is an actual fact only to the extent that the others are not. If memory is to preserve that reality structure, it must behave isomorphically—the remembering frequencies of the three candidate facts must exhibit strong negative correlations. As mentioned in connection with tests of deep distortions (see the earlier Analyses section), such correlations can be computed in either of two ways: (a) at the level of candidate facts, averaged over participants, or (b) at the level of participants, averaged over candidate facts. The resulting bivariate partial correlations are reported in the first column of Table 5.

Three patterns stand out in Table 5. First and most important, the relations among these memory probabilities are not isomorphic with how incompatible facts tradeoff in the real world. There are six correlations, and their grand average (.12) is not even negative, let alone strongly so. Second, although two of the six correlations were negative, neither was reliable, and they were confined to the fact-level analyses. Third, the only reliable correlations were positive rather than negative: Remembering candidate fact B significantly increased the probability of remembering the incompatible candidate fact C, and conversely, in both the fact and participant analyses.

In sum, the fact-level and participant-level correlations provide clear evidence that memories of everyday facts do not preserve the compensation property of incompatibility relations among such facts. Incompatible facts are perfectly complementary, but the corresponding memories are only weakly related, with the dominant relation being positive rather than complementary.

Table 5

Partial Correlations Among the Memory Judgment Probabilities for the Incompatible Candidate Facts A, B, and C in Experiments 1 and 2

	Experiment		
Measurement level	1	2	*M*
Fact level			
r_{AB}	−.20	−.16	−.18
r_{AC}	−.03	−.14	−.09
r_{BC}	.35**	.22*	.29*
M	.04	−.03	.01
Participant level			
r_{AB}	.11	.32**	.28*
r_{AC}	.07	.26*	.22*
r_{BC}	.39**	.29*	.34**
M	.19	.32**	.26*

Note. A = remembering the correct candidate fact for an item; B = remembering the first incorrect candidate fact for an item; C = remembering the second incorrect candidate fact for an item.
* $p < .05$. ** $p < .001$.

Experiment 2

We began with the possibility that memories of everyday facts might not exhibit the deep distortions that have been observed in word-list experiments. Now that deep distortions have been detected for everyday memories, we consider the possibility that, ironically, they may be more robust than in word-list experiments—in particular, that they may be less susceptible to manipulations that are designed to suppress them. We saw that, theoretically, these phenomena result from reliance on noncompensatory gist memories, and thus, it should be possible to suppress them by shifting retrieval in the direction of verbatim traces. Here, a principle known as task calibration has previously been applied to reduce various gist-driven illusions in reasoning and memory (Corbin et al., 2015). This principle grows out of a line of research showing that on memory tests, there are various manipulations that cause retrieval to shift away from general familiarity/plausibility toward literal verbatim memories (e.g., Reder, 1982, 1987). According to this principle, the mix of gist and verbatim retrieval is affected by how fine-grained the response demands of memory and reasoning tasks are, with verbatim reliance increasing as response demands become more fine-grained (Wolfe & Reyna, 2010). A standard example is making simple categorical judgments on memory or reasoning tasks (course-grained) versus making confidence judgments on a numerical scale (fine-grained). Consistent with task calibration, illusions such as semantic false memory, decision framing, and preference reversals decline with confidence judgments.

Also consistent with task calibration, Brainerd et al. (2017) found that deep distortions in word-list experiments virtually disappeared when confidence judgments were added to categorical judgments like those in Experiment 1. This procedure was implemented in Experiment 2 to test the hypothesis that everyday memory distortions are less susceptible to verbatim suppression than distortions in word-list experiments. FTT expects that levels of deep distortion will decline when verbatim memory is cued, but the magnitude of this effect necessarily depends on accessibility: The less accessible verbatim traces are, the smaller the suppression effect from any form of verbatim cuing.

In that connection, a case can be made that verbatim traces should be generally less accessible for everyday memories than they are in word-list experiments. A key consideration is the relative forgetting rates of verbatim and gist traces. Over the interval between encoding and retrieval, a well-replicated finding from both classic memory research (Kintsch & van Dijk, 1978) and recent research (Abadie & Camos, 2019; Greene & Naveh-Benjamin, 2020; Lampinen et al., 2006; Nieznański, 2020) is that verbatim traces become inaccessible more rapidly than gist. This interval is usually far longer for everyday memories than in word-list experiments—a matter of months or years as compared to a few minutes or a few days. The denouement is that deep distortions in memory for everyday facts may be more resistant to suppression from fine-grained confidence judgments than the memories in word-list experiments have been found to be.

To evaluate these possibilities, we added a confidence judgment procedure to the methodology of Experiment 1—specifically, after responding to each memory probe, participants rated how confident they were in their response on a three-point scale. In this design, the confidence tasks are not treated as direct tests of episodic memory; that is the role of DEM items. Rather, the confidence tasks are treated as metacognitive evaluations of the quality of the memories that were retrieved on the probes that preceded them.

Method

Participants

The participants in Experiment 2 were 96 undergraduates who enrolled to fulfill a course requirement, and who received extra credit for participating. Their mean age was 20.00 years (range = 18–23), and the sample was 79% female, 18% male, and 3% unreported. A sensitivity analysis of the data of prior experiments (Brainerd, 2021) indicated that a sample size of 90 provided sufficient power ($1 - \beta = .80$) to detect medium-size effects for the variables of interest to us. However, all individuals who enrolled were allowed to participate—hence, the eventual sample size of 96.

Materials and Procedure

The materials and procedure were the same as in Experiment 1, except for a single change in the response format during conjoint-recognition tests. The instructions were modified to inform participants that following each memory judgment, they would rate their confidence in its accuracy. All 100 probes were presented individually, centered on the computer screen, and printed in black font. The two radio buttons for the memory response appeared below the probe. After answering, the participant was presented with the question "How confident are you in your answer?" Three radio buttons labeled "very sure," "moderately sure," and "slightly sure" appeared below this question. For each probe, the participant first selected a memory response and then selected a confidence rating. The confidence scale was applied in the same manner regardless of whether the participants response was to accept or reject the probe; that is, the participant made the same very sure/moderately sure/slightly sure rating following both types of judgments. The participant was given 10 s to respond to each memory probe, but had no time limit for the confidence rating.

Results

Additivity, Countable Additivity, and Universal Event

As in Experiment 1, there were two main patterns: (a) pronounced deep distortions, and (b) a common theme to the distortions—namely, subadditivity (memory probabilities always exceeded the values that logic allows). These patterns can be seen by inspecting the data in the second column of Table 3, which displays the three deep distortion measures for the DEM.

Also as in Experiment 1, we must determine whether the correct facts in the trios of incompatible candidates for DEM items are part of the long-term memories of this participant population. If they are, p(A) will be greater than either p(B) or p(C), and p(B-or-C) will be smaller than either p(A-or-B) or p(C-or-A). The descriptive statistics in the second column of Table 4 conform to both predicted patterns. To test the patterns for statistical reliability, we computed a 2 (probe type: conventional vs. disjunctive) × 3 (candidate: A, B, C) ANOVA of the judgment probabilities. As in Experiment 1, there was a large Probe Type × Candidate interaction, $F(2, 198) = 53.19$, $MSE = .025$, $p < .001$, $\eta_p^2 = .35$. When it was decomposed with post hoc tests, the order of the memory probabilities for the three conventional probes was $p(\mathrm{A}) > p(\mathrm{B}) > p(\mathrm{C})$, and the order of the memory probabilities for the three disjunctive probes was $p(\text{B-or-C}) < p(\text{A-or-B}) = p(\text{C-or-A})$.

The additivity, countable additivity, and universal event rules were again violated by wide margins. With respect to additivity, all three ratios in the second column of Table 3 are >1, with a grand average of 1.65. One-sample *t* tests established that all these ratios were significantly different from 1, $t(99) = 9.21$, 8.47, and 9.10 (all $ps < .001$). Concerning countable additivity, all three ratios in the second column of Table 3 were also >1, with a grand average of 2.26. One-sample *t* tests again showed that all these ratios were significantly different from 1, $t(95) = 8.77$, 5.81, and 4.57 (all $ps < .001$). Finally, consider the universal event measure $p(\mathrm{A}) + p(\mathrm{B}) + p(\mathrm{C})$. We know that the value for this measure that logic requires is <1 because A, B, and C are not an exhaustive series for these facts. No statistical tests are required to reject that prediction because the observed sum in the second column of Table 3 is >1.

All the results in the second column of Table 3 echo the subadditivity theme that emerged in Experiment 1 for the additivity, countable additivity, and universal event measures. Psychologically, the participants in both experiments over-remembered in ways that violate the constraints of logic.

Finally, although the earlier analysis showed that $p(\mathrm{A})$ was greater than either $p(\mathrm{B})$ or $p(\mathrm{C})$, an examination of the individual DEM items showed this was not true for all 100 facts. When the data were pooled for Experiments 1 and 2, there were 20 facts for which $p(\mathrm{A})$ was equal to or smaller than either $p(\mathrm{B})$ or $p(\mathrm{C})$. This does not affect our findings for deep distortion in Table 3, however. When those analyses were repeated using only the data of DEM items for which $p(\mathrm{A})$ was greater than either $p(\mathrm{B})$ or $p(\mathrm{C})$, all the distortions were still observed.

Suppression of Deep Distortions?

This experiment implemented a manipulation (confidence judgments) that has produced two key effects in word-list experiments: (a) It has reduced the probabilities of remembering old, similar, and new items as old, and (b) it has erased deep distortions like those in Table 3 (see Brainerd et al., 2017). Inspection of the results in the lower halves of Tables 3 and 4 reveals that this manipulation had the first effect but not the second. On the one hand, similar to word-list experiments, each of the six memory probabilities in Table 4 is higher in Experiment 1 than in Experiment 2—roughly 11% higher. The grand means of these memory probabilities are 0.61 for Experiment 1 and 0.55 for Experiment 2, which is a reliable difference, $t(99) = 8.47$, $p < .001$.

On the other hand, unlike word-list experiments, a comparison of the paired values of the three deep distortion measures for Experiments 1 and 2 reveals no suppression effect at all. If everyday memories were susceptible to such suppression, all the values in the first column of Table 3 would be notably larger than the corresponding values in the second column. It is apparent at a glance that this is not the case, and instead, these measures had slightly larger values in the second experiment than in the first (grand $M\text{s} = 1.66$ and 1.89).

A likely explanation of the absence of a suppression effect, which will be evaluated later with the conjoint-recognition model (see General Discussion section), revolves around the accessibility of verbatim traces. In word-list experiments, confidence judgments suppressed deep distortions because they shifted retrieval in the direction of verbatim traces. Notice, however, that confidence judgments could not produce such a shift if verbatim traces are no longer accessible or are very difficult to access. As mentioned earlier, this may be the case in memory for everyday facts, owing to the long delay between when they were stored and when memory tests are administered. A report of the confidence data themselves—the frequencies and percentages of slightly sure, moderately sure, and very sure ratings for the various conditions of this experiment—may be found in the online supplemental materials.

Compensation Among Incompatible Memories

We know that if memory is to obey the logical rules that govern incompatibility relations, the memory probabilities of the three candidate facts for DEM items must be mutually compensatory; they must tradeoff the way incompatible events tradeoff in the real world. The relevant correlations can be computed at either the fact level, averaged over participants, or at the participant level, averaged over facts. Both sets of bivariate partial correlations for Experiment 2 are reported in the second column of Table 5.

The same three patterns stand out as in Experiment 1. First and foremost, there was no compelling evidence that these memory probabilities tradeoff in the manner that incompatible events tradeoff. There are six correlations, and their grand average (.15) is not even negative. Second, although two of the six correlations were negative, neither was reliable, and they were confined to the fact-level analyses. Third, the only reliable correlations were positive rather than negative. In fact, the other four correlations, though modest, were all positive and reliable. For the participant level correlations, in particular, remembering any one of these incompatible candidates slightly but significantly increased the probability of remembering each of the other two.

As in Experiment 1, then, neither the fact-level correlations nor the participant-level correlations provided any evidence that everyday memories preserve the compensation property of incompatibility relations among everyday facts. Although incompatible facts are perfectly complementary, the corresponding memories are only weakly related, with the dominant relation being positive rather than negative.

Experiments 3 and 4

Experiments 1 and 2 are the first to measure deep distortions in memory for everyday facts, and they are the first to measure the memory counterpart of the compensation property of incompatibility relations. Therefore, it seemed advisable to attempt to replicate all aspects of their results with larger, older, and more diverse participant samples. Experiment 3 was a replication of Experiment 1, with a large sample of adults drawn from a survey platform. Experiment 4 was a replication of Experiment 2, also with a large sample of adults drawn from the same survey platform. Beyond replication, the more specific objectives of Experiments 3 and 4 were the same as those of Experiments 1 and 2, respectively.

Method

Participants

The participants in Experiment 3 were 200 adults who were drawn from the Prolific survey platform. They enrolled for compensation, and all were residents of the United States. Their mean age was 32.39 years (range = 20–51), and the sample was 59% female, 38%

male, and 3% nonbinary. The participants in Experiment 4 were 201 adults from the same survey platform who enrolled for compensation and were residents of the United States. Their mean age was 33.97 years (range = 18–56), and the sample was 54% female, 45% male, and 1% nonbinary. Because Experiment 3 was a replication of Experiment 1 and Experiment 4 was a replication of Experiment 2, the earlier power calculations applied here. However, considering that those power calculations were based on data from undergraduate populations like those in the first two experiments, we decided to double the sample size to ensure reliability.

Materials and Procedure

Experiment 3 was a replication of Experiment 1 with a much larger participant sample, in which the participants were 13 years older on average and the gender composition was more balanced. Except for those features, the materials and procedures of Experiments 1 and 3 were the same. Similarly, Experiment 4 was a replication of Experiment 2 with a much larger participant sample, in which the participants were 14 years older on average and the gender composition was more balanced. Except for those features, the materials and procedures of Experiments 2 and 4 were the same.

Results

The values of the three deep distortion measures for each experiment are reported in Table 6. Descriptive statistics for the six memory judgments for the DEM items are reported in Table 7. As before, we present the results for additivity, countable additivity, and universal event first, followed by the results for compensation among incompatible memories.

Additivity, Countable Additivity, and Universal Event

As in Experiments1 and 2, there were two key patterns: (a) pronounced deep distortions and (b) a subadditivity theme of these deep distortions (i.e., memory probabilities always exceeded the values required by logic). These patterns can be seen at a glance in the upper half of Table 6, which displays the three deep distortion measures for the DEM. Each of these values was derived from the descriptive statistics for the six memory probes that appear in Table 7.

Table 6
Deep Distortion Measures for Experiments 3 and 4

	Experiment	
Distortion measures	3	4
Additivity		
$p(\text{A}) + p(\text{B})]/p(\text{A-or-B})$	1.42	1.43
$[p(\text{B}) + p(\text{C})]/p(\text{B-or-C})$	1.38	1.34
$[p(\text{A}) + p(\text{C})]/p(\text{C-or-AC})$	1.45	1.43
Countable additivity		
$[p(\text{A}) + p(\text{B}) + p(\text{C})]/[p(\text{A-or-B}) + p(\text{B-or-C}) - p(\text{B})]$	1.95	1.64
$[p(\text{A}) + p(\text{B}) + p(\text{C})]/[p(\text{A-or-B}) + p(\text{C-or-A}) - p(\text{A})]$	1.59	1.71
$[p(\text{A}) + p(\text{B}) + p(\text{C})]/[p(\text{C-or-A}) + p(\text{B-or-C}) - p(\text{C})]$	1.42	1.52
Universal set		
$p(\text{A}) + p(\text{B}) + p(\text{C})$	1.81	1.41
$p(\text{A}) + p(\text{B})$	1.12	1.07
$p(\text{A}) + p(\text{C})$	1.09	1.02

Note. A = remembering the correct candidate fact for an item; B = remembering the first incorrect candidate fact for an item; C = remembering the second incorrect candidate fact for an item.

Table 7
Mean Memory Probabilities and Standard Deviations for the Distortion in Everyday Memory Instrument in Experiments 3 and 4

	Measures					
	Conventional probes			Disjunctive probes		
Experiment	A	B	C	A or B	B or C	C or A
Experiment 3						
M	.70	.42	.39	.79	.56	.79
SD	.19	.19	.20	.14	.19	.13
Experiment 4						
M	.68	.39	.34	.75	.51	.76
SD	.20	.19	.21	.15	.21	.15

Note. A = remembering the correct candidate fact for an item; B = remembering the first incorrect candidate fact for an item; C = remembering the second incorrect candidate fact for an item.

As before, we begin with the question of whether the correct fact candidates of DEM items are part of the long-term memories of this participant population. First, it can be seen in Table 7 that $p(\text{A})$, the probability of remembering the correct candidate, was .70 in Experiment 3 and .69 in Experiment 4. Similar to the first two experiments, those probabilities were in the range of $400 Jeopardy! facts, according to our survey of 2022 broadcasts. Second, if the correct candidates for DEM items are in the long-term memories of this participant population, $p(\text{A})$ will be greater than either $p(\text{B})$ or $p(\text{C})$, and $p(\text{B-or-C})$ will be smaller than either $p(\text{A-or-B})$ or $p(\text{C-or-A})$. For Experiment 3, a 2 (probe type: conventional vs. disjunctive) × 3 (candidate: A, B, C) ANOVAs confirmed this pattern. It produced a large Probe Type × Candidate interaction, $F(2, 198) = 66.94$, $MSE = .021$, $p < .001$, $\eta_p^2 = .40$. Planned comparisons revealed that this interaction showed that $p(\text{A})$ was significantly larger than $p(\text{B})$ or $p(\text{C})$ for conventional probes and $p(\text{B-or-C})$ was significantly smaller than $p(\text{A-or-B})$ and $p(\text{C-or-A})$ for disjunctive ones. Experiment 4 also produced a large Probe Type × Candidate interaction, $F(2, 198) = 77.23$, $MSE = .022$, $p < .001$, $\eta_p^2 = .44$, and post hoc tests revealed the same two patterns among the individual memory probabilities.

Turning to the three deep distortions, the additivity, countable additivity, and universal event rules of incompatibility relations were once again violated by wide margins in both experiments. With respect to additivity, the ratios in Table 6 must not differ reliably from 1. They are not close to 1, however. The average value was 1.42 in Experiment 3 and 1.40 in Experiment 4. One-sample t tests established that all these ratios were significantly different than 1 in both Experiment 3, $t(99) = 15.56$, 10.64, and 13.24 (all $ps < .001$), and Experiment 4, $t(99) = 12.43$, 9.23, and 11.69 (all $ps < .001$).

As in the first two experiments, all violations of the additivity rule are in the subadditive direction. Psychologically, these participants also remembered too much, endorsing more facts than could logically be true, rather than too few, which is a known characteristic of reliance on gist memory (Brainerd, 2022; Brainerd & Reyna, 2018).

Continuing to countable additivity, for this rule to be obeyed, the ratios in Table 6 must not differ significantly from 1. However, they are all well above 1, averaging 1.65 in Experiment 3 and 1.63 in

Experiment 4. One-sample t tests showed that all these ratios were significantly different from 1 in both Experiment 3, $t(99) = 15.47$, 15.37, 14.00 (all ps $< .001$), and Experiment 4, $t(99) = 12.06$, 10.74, and 11.06 (all ps $< .001$). Thus, both the countable additivity and additivity results show that these participants' memory for logically incompatible candidate facts were completely nonadditive. Also, the participants' memories were completely subadditive because all the ratios were >1. Similar to Experiments 1 and 2, the level of deep distortion was somewhat higher (16% higher) for countable additivity than for additivity.

Last, consider the universal event measure in Table 6, $p(\mathrm{A}) + p(\mathrm{B}) + p(\mathrm{C})$, and remember that this rule requires that the sum must be <1. As in the first two experiments, no statistical tests are required to reject this prediction because the observed values of $p(\mathrm{A}) + p(\mathrm{B}) + p(\mathrm{C})$ are far above 1, averaging 1.61. Further, one-sample t tests showed that these values were significantly >1 in both Experiment 3, $t(99) = 14.85$, $p < .001$, and Experiment 4, $t(99) = 11.08$, $p < .001$. In short, the probabilities of remembering incompatible candidates for DEM facts overflowed the unit interval by wide margins.

An especially surprising outcome in the first two experiments was that this third deep distortion was so strong that memory also overflowed the unit interval for two of the subquantities of the universal event measure—specifically, for $p(\mathrm{A}) + p(\mathrm{B})$ and for $p(\mathrm{A}) + p(\mathrm{C})$. It can be seen in Table 6 that this happened again in Experiments 3 and 4. Specifically, $p(\mathrm{A}) + p(\mathrm{B})$ and $p(\mathrm{A}) + p(\mathrm{C})$, which should both be far below 1 under the universal event rule, were slightly above 1 in both experiments. Moreover, both subquantities were significantly >1 in Experiment 3, $t(99) = 4.71$ and 3.86 (both ps $< .001$), and $p(\mathrm{A}) + p(\mathrm{B})$ was significantly >1 in Experiment 4, $t(99) = 2.64$, $p < .01$.

Summing up, the results of Experiments 3 and 4 agree with those of Experiments 1 and 2 in three critical respects. First, there was strong evidence of deep distortions in memory for everyday facts. Second, all three logical rules were violated because memory probabilities were subadditive rather than additive. Third, regardless of which of the three indexes is considered, participants endorsed more of the candidate facts than could possibly be true.

To determine whether there were any significant differences in performance as a function of age in Experiments 3 and 4, we pooled the data of the two experiments and conducted a one-way analysis of covariance (ANCOVA). Specifically, we pooled the data and computed a one-way ANCOVA of acceptance proportions using a six-level probe factor (A, B, C, AB, AC, BC) and using age as a covariate. The age covariate was not reliable in this ANCOVA. Of course, this only means age was not a reliable predictor of probe performance with the age range of the participant samples of these experiments. It might be reliable with more extreme age separation (e.g., young vs. elderly adults).

Suppression of Deep Distortions?

Experiment 4 added a verbatim cuing manipulation (confidence judgments) to the basic design of Experiment 3. In word-list experiments, we know that confidence judgments (a) reduce the probabilities of remembering old, similar, and new items as old and (b) erase deep distortions like those in Table 6. In the first two experiments, however, the manipulation produced the first effect, but it did not affect the strengths of the three deep distortions. In the present experiments, it had both effects, though the effects were small in absolute terms.

First, inspection of the mean memory probabilities in Table 7 reveals that each of the six probabilities was lower in Experiment 4, where participants also made confidence judgments, than in Experiment 3, where they only made categorical judgments. The magnitude of the effect was small (grand Ms $= 0.61$ and 0.57), although it was reliable, $t(98) = 7.41$, $p < .001$. Second, an inspection of the deep distortion measures in Table 6 reveals that there was no suppression effect for additivity or countable additivity, their average differences being very small (M_{Δ}s $= 0.02$ and 0.03) and unreliable. However, there was a suppression effect for the universal event index. The values of this index were 1.81 when participants only made categorical judgments versus 1.41 when they also made confidence judgments—a reliable difference, $t(98) = 4.77$, $p < .001$.

As in the first two experiments, a likely explanation of these results, which will be evaluated later with the conjoint recognition model (see General Discussion), is concerned with the accessibility of verbatim traces. According to that explanation, verbatim cuing manipulations are less effective with everyday memories than they are in word-list experiments, owing to decreased accessibility of verbatim traces. According to that explanation, such manipulations cannot dramatically shift retrieval toward verbatim traces with everyday memories because such traces become largely inaccessible over the long delays between storage and test. As in Experiment 2, a report of the confidence data of Experiment 4—the frequencies and percentages of slightly sure, moderately sure, and very sure ratings for the various conditions of this experiment—may be found in the online supplemental materials.

Compensation Among Incompatible Memories

Both the fact-level and participant-level partial correlations among the memory probabilities for incompatible fact candidates in these experiments are displayed in Table 8. The patterns are remarkably similar to those in Experiments 1 and 2. The most instructive similarity is

Table 8
Partial Correlations Among the Memory Judgment Probabilities for the Incompatible Candidate Facts A, B, and C in Experiments 3 and 4

	Experiment		
Measurement level	3	4	*M*
Fact level			
r_{AB}	.03	−.06	−.02
r_{AC}	−.31**	−.06	−.19
r_{BC}	.47**	.51**	.49**
M	.06	.13	
Participant level			
r_{AB}	.14	−.20*	−.02
r_{AC}	−.01	.22*	.11
r_{BC}	.33**	.34**	.34**
M	.15	.12	

Note. A = remembering the correct candidate fact for an item; B = remembering the first incorrect candidate fact for an item; C = remembering the second incorrect candidate fact for an item.
* $p < .05$. ** $p < .005$.

that there was no general support for the hypothesis that the probabilities of remembering these incompatible candidates trade off in the manner that incompatible facts trade off in the real world. There are 12 correlations in all, and their grand average (.12) is neither reliable nor negative. A second similarity is that although there were two reliable negative correlations, they were weak, accounting for an average of only 7% of the variance. A third similarity is that the only substantial and consistently reliable correlations were positive rather than negative, for r_{BC}. All four r_{BC} correlations were reliable, positive, and accounted for an average of 18% of the variance. In short, although incompatible facts are perfectly complementary, the corresponding memories are only weakly related, and the overall relation is positive rather than strongly negative.

General Discussion

In this section, we consider two broad topics. The first consists of the major new things that have been learned about deep distortions in these experiments. The other is how best to explain those patterns theoretically. To evaluate theoretical explanations, we will apply the conjoint-recognition model to the data of the four experiments.

New Findings About Deep Distortions

Three patterns are of principal significance. The first is the extensive support that these experiments supply for the conclusion that deep distortions are not mere curiosities of laboratory experiments with word lists. In all the present experiments, memory for everyday facts was infected with the same deep distortions that have been detected with word lists. It violated three logical rules that govern incompatibility relations among events in the world. This is an instructive result because heretofore, such violations have only been observed in word-list experiments, which opens the door to ecological validity arguments about their theoretical significance and generality. According to such arguments, differences in the content and conditions of laboratory versus everyday memories are so substantial that the latter might adhere to logical constraints even though the former do not. Although that is a real possibility, we noted earlier that a contrasting argument can be made that everyday and laboratory memories are both susceptible to deep distortions because people rely on gist traces in both realms. Our findings are congruent with that line of reasoning.

The second pattern of prime significance is that memory for everyday facts always overshot the bounds of logic. Explicitly, the sum of the probabilities of remembering each of two incompatible candidates always exceeded the probability of remembering their disjunction, the sum of the probabilities of remembering each of three incompatible candidates always exceeded the probability of remembering their disjunction by an even wider margin, and the sum of the probabilities of remembering each of three incompatible candidates always overflowed the unit interval. In fact, rather astonishingly, the sums of the probabilities of remembering just two of these incompatible candidates—$p(\text{A}) + p(\text{B})$ and $p(\text{A}) + p(\text{C})$—also overflowed the unit interval.

The psychological significance of the second pattern runs as follows. When real-world facts are incompatible—as in not being able to stand on your head, stand on your feet, and lie in a prone position at the same time—logic constrains their permissible occurrence probabilities so that the additivity, countable additivity, and universal event rules are all satisfied. However, when incompatible facts are remembered, as in the DEM instrument, the probabilities of remembering them behave illogically. Those probabilities are always subadditive, which means that people consistently remember more about incompatible facts than is logically possible—facts that are physically incompatible do not seem so to memory.

The third pattern of prime significance is concerned with the logical rule that incompatible events are mutually compensatory; that the probability of one can only increase as the probability of others decreases. Empirically, this means strong negative correlations. Although incompatible events always correlate in this manner (e.g., correlations among the observed instances of vehicles turning left, turning right, and driving straight ahead at an intersection), we now know that the corresponding memories do not. For each DEM item, candidate A is correct, whereas candidates B and C are both incorrect. Thus, there are two types of correlations that evaluate compensation, both of which should be strongly negative: (a) correlations between correct and incorrect memories (r_{AB} and r_{AC}) and correlations between incorrect memories (r_{BC}). Over the four experiments, the first type of correlation was essentially zero (grand $M_{r\text{AB}} = 0.02$ and grand $M_{r\text{AC}} = 0.01$), and remarkably, the second type of correlation was always positive (grand $M_{r\text{BC}} = 0.37$, $p < .001$). In short, although people are aware, metacognitively, that it is impossible for more than one of these candidates to be true, memory does not see it that way.

Explaining Deep Distortions

We now consider the second topic, in which the conjoint-recognition model was applied to our data to test theoretical explanations. We already considered the fact that deep distortion research was originally motivated by the hypothesis that people rely on gist memories at least some of the time in most forms of episodic memory. Further, the patterns that we have just described are consistent with gist reliance, as noted earlier. However, that is not a direct test of such an explanation, which would require that rates of gist and verbatim retrieval be measured to determine whether the values fall out in accordance with the explanation.

Such tests are described in the two subsections that follow. In the first, we sketch the specific version of the conjoint-recognition model that is used with DEM data. In the second, we fit that model to the data of Experiments 1–4, estimate its parameters, and determine whether the values of its gist, verbatim, and bias parameters are consistent with the process explanation of the findings.

Conjoint-Recognition Model for the DEM

To see whether noncompensatory gist can explain deep distortions, a conjoint-recognition model was developed for the DEM, where such models provide estimates of the contributions of verbatim retrieval, gist retrieval, and response bias to performance on memory and reasoning tasks (Brainerd et al., 1999). Although versions of the model have been applied to certain reasoning tasks (Abadie et al., 2013, 2017; Singer & Remillard, 2008; Singer & Spear, 2015), the most extensive applications have been in memory experiments with word and picture lists (e.g., Abadie & Camos, 2019; Greene & Naveh-Benjamin, 2020; Lampinen et al., 2006; Nieznański, 2020; Obidziński, 2021; Obidziński & Nieznański, 2017). In such experiments, the model measures the contributions of verbatim retrieval, gist retrieval, and response bias to true and false memory for items and sources (for a review, see Brainerd et al., 2022).

All versions of the model have two features in common. First, they contain three types of parameters: (a) verbatim parameters that support acceptance of true items/sources (e.g., candidate A in DEM facts) and rejection of false items/sources that are related to true ones (e.g., candidates B and C in DEM facts); (b) gist parameters that support acceptance of both true items/sources and related false items/sources; and (c) bias parameters that support acceptance of both true items/sources and related items/sources when verbatim and gist retrieval fail. Second, these models are defined over experimental paradigms that contain at least three types of memory probes: (d) judgments about true items/sources (e.g., judgments about A candidates in DEM facts); (e) judgments about related false items/sources (e.g., judgments about B and C candidates in DEM facts); and (f) judgments about disjunctions of true and related false items (e.g., the A-or-B, C-or-A, and B-or-C judgments in DEM facts). We mentioned earlier that these features have been implemented in between-participants designs, within-participants designs, and multiple-choice designs, using slightly different versions of the conjoint-recognition model (see Abadie & Camos, 2019; Abadie et al., 2013; Gong et al., 2016; Greene & Naveh-Benjamin, 2020, 2022; Nieznański, 2020; Obidziński, 2021).

The model version that is appropriate for our experiments is displayed in Equations A1–A6 of the Appendix. Note that its parameters measures three distinct processes. First, V_A is the probability that participants retrieve a verbatim trace of learning candidate A, which leads to acceptance of the A, A-or-B, and C-or-A probes as true coupled with rejection of the other three probes as false. Next, G_A, G_B, and G_C are the probabilities that probes containing the candidates in the subscripts produce acceptance of the probes as true. Hence, G_A is the probability that the A, A-or-B, and C-or-A probes produce retrieval of gist traces that support acceptance, G_B is the probability that the B, A-or-B, and B-or-C probes produce retrieval of gist traces that produce acceptance, and G_C is the probability that the C, C-or-A, and B-or-C probes produce retrieval of gist traces that support acceptance. Last, b is a bias parameter. It is simply the probability that when verbatim and gist retrieval both fail for any of the probes, response bias causes participants to accept that candidate as true.

Analysis of the model's equations shows how it predicts violations of the additivity, countable additivity, and universal event rules of incompatibility relations. For the additivity rule, analysis reveals that it will be violated, and violations will be in the subadditive direction when these parameter relations hold: $V_A < 1$, $G_B > 0$, and $G_C >$ 0. In that connection, algebraic manipulation indicates that under those constraints, the sum of A1 and A2 must be greater than A4, the sum of A1 and A3 must be greater than A5, and the sum of A2 and A3 must be greater than A6. Note, further, that the additivity rule will be satisfied if the last two conditions do not hold. Turning to the countable additivity rule, analysis reveals that it will also be violated in the subadditive direction when these parameter relations hold: $V_A < 1$, $G_A > 0$, $G_B > 0$, and $G_C > 0$. The analysis also shows that countable additivity will be satisfied if the last three conditions do not hold. Concerning the universal event rule, for this rule to be satisfied, the upper bound of $p(\text{A}) + p(\text{B}) + p(\text{C})$ must always be 1. However, analysis reveals that this sum will be $>$1 when $V_A < 1$ and $G_A > 0$, $G_B > 0$, and $G_C > 0$. Finally, with respect to all three rules, analysis reveals that the strengths of observed violations will be inversely proportional to the value of V_A and directly proportional to the values of G_A, G_B, and G_C.

Testing Theoretical Explanations With the Model

These explanations of deep distortions can be tested by estimating the model's retrieval parameters and determining whether their values fall out in the manner that was just described. First, however, the model must be fit to the data of each experiment to determine whether it delivers statistically tolerable accounts of those data. As mentioned in the Appendix section, the fit test is a $G^2(1)$ statistic with a critical value of 3.84 to reject the null hypothesis that the model fits and, therefore, a more complex model is not required. The results of the fit tests for the present experiments appear in the first column of Table 9. None rejected the null hypothesis, and indeed, the mean value of the four $G^2(1)$ statistics (0.31) is only a small fraction of the critical value.

As fits were acceptable, we estimated the model's verbatim, gist, and bias parameters for each experiment. The *M*s and *SD*s of these five parameters are reported in rows 1, 2, 3, and 4 of Table 9. Inspection of the parameter estimates, within and between experiments, reveals four patterns of theoretical significance. First and foremost, the estimates confirm the above explanations of deep distortions inasmuch as their values satisfy the conditions under which violations of additivity, countable additivity, and universal event in a subadditive direction should be observed. Specifically, with respect to violations of additivity, countable additivity, and universal event, notice that $V_A < 1$, $G_A > 0$, $G_B > 0$, and $G_C > 0$ in all experiments. Further, when the grand mean values of these parameters were used to compute an average predicted value of $p(\text{A}) + p(\text{B}) + p(\text{C})$, the predicted value was 1.48, which was only slightly smaller than the average observed value (1.55). Recall that even the subquantity $p(\text{A}) + p(\text{B})$ always overflowed the unit interval to a statistically reliable extent. That pattern is also explained by the parameter values in Table 9: The average predicted value of $p(\text{A}) + p(\text{B})$ is $>$ 1 and only slightly larger than the average observed value (1.16 vs. 1.09).

The second salient pattern in Table 9 pertains to the earlier discussion of the retrieval differences between memory tasks such as the DEM and laboratory experiments with mean value of V_A in Table 9 being only 0.02, and the values in Experiments 2 and 3 not being reliably $>$0. Over the four experiments, the grand mean of this parameter indicates that verbatim retrieval failed 98% of the time. Thus, with memory for everyday facts, the type of retrieval that suppresses deep distortions was largely absent, which is consistent with theoretical expectations. Obviously, this also explains why a manipulation that has been effective at reducing deep distortions in laboratory experiments (confidence judgments) was ineffective with memory for everyday facts. This pattern can be contrasted with the much higher levels of verbatim retrieval that have been reported in laboratory conjoint-recognition experiments, where values of the verbatim retrieval parameter are far larger (for reviews, see Brainerd et al., 2022; Greene & Naveh-Benjamin, 2023). For instance, in the corpus of conjoint-recognition data that was reviewed by Brainerd et al., the mean value of the verbatim retrieval parameter for old test items was 0.37 for experiments that implemented between-participants designs and 0.30 for experiments that implemented within-participants designs.

We conducted two follow-up analyses of the finding that verbatim retrieval was close to zero in these experiments. First, we sought to determine whether verbatim retrieval might be substantially higher in a theoretically specified subset of the data. A

Table 9

Goodness-of-Fit Tests and Means and (Standard Deviations) of the Conjoint-Recognition Models' Parameters for Experiments 1–4

	Statistic					
Experiment	G^2	V_A	G_A	G_B	G_C	b
	Parameter estimates for the unrestricted model					
1	0.01	0.06 (.36)	0.55 (.29)	0.28 (.16)	0.24 (.15)	0.26 (.17)
2						
All data	0.42	0	0.50 (.35)	0.21 (.12)	0.09 (.10)	0.28 (.18)
High confidence	0.50	0	0.86 (.04)	0.24 (.07)	0.17 (.07)	0
3						
All data	0.40	0.01 (.24)	0.66 (.12)	0.34 (.11)	0.26 (.10)	0.12 (.09)
High confidence	0.17	0.01 (.20)	0.89 (.15)	0.18 (.25)	0.14 (.20)	0
4	0.40	0.02 (.31)	0.64 (.19)	0.26 (.11)	0.21 (.10)	0.18 (.10)
	Parameter estimates for a restricted model with $V_A = 0$					
1	0.03	0	0.60 (.05)	0.26 (.09)	0.22 (.10)	0.24 (.10)
2	0.42	0	0.50 (.06)	0.21 (.08)	0.09 (.10)	0.28 (.09)
3	0.39	0	0.66 (.03)	0.33 (.06)	0.25 (.06)	0.12 (.09)
4	0.39	0	0.61 (.03)	0.25 (.06)	0.21 (.06)	0.17 (.08)

Note. The fit statistic G^2 has one degree of freedom for the unrestricted model and two degrees of freedom for the restricted model. The high confidence parameter estimates were computed for a subset of the data of Experiments 2 and 4, for which correct memory judgments were accompanied by the highest confidence rating (very sure).

familiar finding in the false memory literature is that people assign higher confidence ratings to correct memory judgments when levels of verbatim retrieval are higher (e.g., Brainerd & Reyna, 2005; Gallo, 2006). As confidence ratings were obtained in Experiments 2 and 4, we used those data to test the hypothesis that V_A would be much higher when confidence was high. Specifically, we refit the model to only those judgments to which participants assigned the highest rating on the confidence scale. The results appear in rows 3 and 5 of Table 9, where it can be seen that the hypothesis was rejected: $V_A = 0$ for high confidence judgments in both experiments.

In the second reanalysis, we tested the hypothesis that verbatim retrieval levels for everyday facts were so low that the V_A parameter can be discarded in these experiments; that V_A is so small that it is not required to achieve model fit. To test that hypothesis, we refit the model to the data of all the experiments, subject to the constraint that $V_A = 0$. The results are reported in the lower half of Table 9, where the hypothesis proved to be correct because all four fit statistics were acceptable, $M^2_{G(2)} = 0.31$.

The last two patterns in Table 9 are concerned with the relative magnitudes of the three gist-retrieval parameters, on the one hand, and between-experiment variability in the bias parameter, on the other. Concerning the gist-retrieval parameters, theory expects that true alternatives should be better retrieval cues for their semantic gist than false alternatives. That prediction has been confirmed in laboratory experiments, where estimates of gist-retrieval parameters for old items are larger on average than estimates of gist-retrieval parameters for similar items (Brainerd et al., 2022). The prediction also holds for memory for everyday facts, as estimates of the G_A parameter (grand $M = 0.58$) were more than twice the estimates of the G_B and G_C parameters (grand Ms $= 0.27$ and 0.21, respectively). These were highly reliable differences: $G^2(1)$ tests of the null hypothesis that $G_A = G_B$ were significant in all four experiments (all ps $< .001$), and so were $G^2(1)$ tests of the null hypothesis that $G_A = G_C$ (all ps $< .001$).

With respect to response bias, inspection of the last column of Table 9 shows that bias to accept plausible alternatives when verbatim and gist retrieval both failed was roughly twice as strong in the first two experiments ($M_b = 0.27$) as in the last two ($M_b = 0.15$). This may seem surprising to some readers, who might expect that bias levels would be higher in the more diverse participant samples of Experiments 3 and 4. Actually, however, this result can be viewed as a natural continuation of previously reported developmental trends in bias. A common developmental finding with recognition is that measures of response bias decrease steadily from childhood to adolescence to the early 20 s (e.g., Holliday et al., 2011). It would be expected that this trend should continue for at least a few more years, but the decade immediately following the early 20 s is rarely studied. In our case, however, it was: The participants in the last two experiments were 33 years old on average, whereas the participants in the first two experiments were 20 years old on average.

Summing up, the model analyses provided direct tests of theoretical explanations of the high levels of deep distortion that were observed in memory for everyday facts. On the one hand, the verbatim retrieval process that suppresses deep distortions was near-floor in all experiments. On the other hand, the noncompensatory gist processes that support deep distortions were present at substantial levels in all experiments.

Bias Correction With the Model

A tactical advantage of the model is that it can address the familiar signal detection hypothesis that responses to recognition memory probes are influenced by bias as well as the contents of memory. The standard approach to this hypothesis is to measure performance with the memory parameters of some theoretical model that provides separate memory and bias parameters (e.g., see Macmillan & Creelman, 1991; Snodgrass & Corwin, 1988). That can be done with the present model; that is, the statistics in Tables 3 and 6 that evaluate the three deep distortions can be recalculated with the memory parameters V_A, G_A, G_B, and G_C, using Equations A1–A6. When that is done, all the distortions are still observed, and the relevant statistical tests are still highly reliable. To illustrate, recall that in Experiment 1, the additivity rule was violated for the three ratios $[p(\mathrm{A})+p(\mathrm{B})]/p(\text{A-or-B})$, $[p(\mathrm{B})+p(\mathrm{C})]/p(\text{B-or-C})$, and $[p(\mathrm{A})+p(\mathrm{C})]/p(\text{A-or-C})$. Their observed values were 1.45, 1.52, and 1.48, respectively. When these ratios are recalculated with the memory parameter estimates in Table 9, the additivity property is still violated by a wide margin, the ratios' observed values being 1.29, 1.25, and 1.28. Similarly, recall that in Experiments 3 and 4, the universal event rule $p(\mathrm{A})+p(\mathrm{B})+p(\mathrm{C})<1$ was violated. Their observed values were 1.81 and 1.41, respectively. When their values are recalculated with the memory parameter estimates in Table 9, the universal event rule is still violated, the observed values being 1.24 and 1.10, respectively.

Although those analyses showed that all deep distortions were still present when a common type of bias correction was applied to the data, we consider two further interpretive questions about response bias: (a) the potential inclusion of implausible distractors in the DEM to provide model-free measures of bias and (b) the use of different bias parameters for nondisjunctive DEM judgments (i.e., A, B, and C) versus disjunctive judgments (A-or-B, A-or-C, and B-or-C). Concerning a, in laboratory word list studies of deep distortion, recognition tests have included unrelated distractors (Brainerd, 2021). Unrelated distractors do not preserve the surface form of old items or their gist. This means that they are relatively pure measures of response bias because their acceptance or rejection is not supported by either verbatim or gist retrieval. For that reason, unrelated distractor data can be used to correct true and false memory responses for bias by subtracting the false alarm rate for unrelated distractors from, respectively, the hit rate and from the false alarm rate for related distractors (e.g., Holliday et al., 2011; Schacter et al., 1999). It might be thought that a similar approach could be implemented with the DEM by adding a fourth candidate for each fact that unlike A, B, and C, is implausible. (For instance, if the country where Einstein was born is the target fact and the plausible candidates are A = Germany, B = Austria, C = Switzerland, an implausible candidate would D = Japan.) This approach does not work, however, because improbable candidates are false-gist distractors rather than unrelated distractors. Specifically, improbable candidates violate the core gist of target facts. Thus, they can be confidently rejected if that gist is retrieved, and they will not be accepted if it is retrieved (Brainerd et al., 2006). Rather than being pure measures of bias, then, gist retrieval figures in judgments about implausible distractors.

Turning to question b, the conjoint recognition model that has been used in laboratory studies of deep distortions contains more bias parameters than the present model. In particular, it contains separate bias parameters for nondisjunctive probes (analogous to A, B, and C in the DEM) versus disjunctive probes (analogous to A-or-B, A-or-C, and B-or-C in the DEM). Across hundreds of data sets that have been analyzed with that model, the bias level of the nondisjunctive parameter (call it b) is lower than the bias level of the disjunctive parameter (call it b_{XY}). For instance, their mean values in the data sets reviewed by Brainerd et al., 2022) were $b = .16$ and $b_{\mathrm{XY}} = .22$. It is not possible to include separate b and b_{XY} parameters in the present model because that would saturate the parameter space, making it impossible to compute fit tests (see Appendix section). However, that leaves us with the question of how the deep distortions that were measured with the DEM would be affected if $b < b_{\mathrm{XY}}$, and the earlier bias corrections could have been made using both of these parameters. Might the deep distortions now disappear or shrink to negligible levels?

It turns out that this cannot happen, for mathematical reasons. To begin, recall that the expression for the universal event property is $p(\mathrm{A})+p(\mathrm{B})+p(\mathrm{C}) \leq 1$. As no disjunctive probabilities appear in this expression, only the nondisjunctive bias parameter is used to correct the acceptance probabilities, and hence, the result cannot depend on the relative magnitude of b versus b_{XY}. Next, consider the expression $[p(\mathrm{A})+p(\mathrm{B})]/p(\text{A-or-B})$ for the additivity property and the expression $[p(\mathrm{A})+p(\mathrm{B})+p(\mathrm{C})]/[p(\text{A-or-B})+p(\text{B-or-C})-p(\mathrm{B})]$ for the countable additivity property. Note that in both instances, only nondisjunctive probabilities appear in the numerator, and disjunctive probabilities appear only in the denominator. Therefore, only b would be used to correct numerators for bias, and b_{XY} would only be used to correct denominators for bias. It follows that if $b < b_{\mathrm{XY}}$, violations of additivity and countable additivity will necessarily be larger if both parameters are used to make bias corrections than if only b is used (as in our experiments). In sum, although it is possible that $b < b_{\mathrm{XY}}$ in the DEM, as in word list experiments, using both parameters to make bias corrections would not cause any of the three deep distortions to shrink or disappear, relative to the results of using only b.

Episodic or Semantic Memory?

A general theoretical question that should be considered in closing is, What type of memory system does the DEM tap? This instrument consists of probes for correct and incorrect candidates for everyday facts, such as the largest planet in the solar system or the first U.S. president to face impeachment. Throughout this article, we have treated such probes as everyday counterparts of the laboratory episodic memory tasks that were originally used to study deep distortions. That interpretation is grounded in the reality that people learn DEM facts by encoding everyday objects and events in specific places at specific times with accompanying contextual details (e.g., a map of the solar system in a middle-school science class, a textbook chapter in a high school history class). According to FTT, people store literal verbatim traces of those experiences and gist traces of their meaning, either or both of which can be retrieved when responding to DEM items. Thus, both types of traces are episodic in the traditional sense that they have autobiographical indexes. Under that interpretation, there are two main points of difference between the DEM and laboratory tasks: (a) The delay between encoding and test is far greater for the DEM, which favors gist retrieval, and (b) everyday facts are encoded in multiple contexts, so that the corresponding traces carry many more contextual details than their laboratory counterparts. Owing to the latter feature, traces that are retrieved by everyday facts seem more abstract than those that are retrieved on laboratory tasks because they do not specify unique episodic sources (Nelson & Shiffrin, 2013).

In contrast, it could also be argued that the key point of difference between the DEM and laboratory tasks is that they tap fundamentally different memory systems. According to this interpretation, the laboratory tasks that have been used to study deep distortions tap episodic memory, but the DEM taps semantic memory. Consistent with that hypothesis, semantic memory is usually defined as general knowledge that people acquire during the course of their lives, such as concepts, word meanings, and ideas (e.g., McRae & Jones, 2013; Tulving, 2002). Further, everyday facts of the type that figure in the DEM are often mentioned as examples of information that is stored in a semantic memory system that is distinct from episodic memory because the information does not derive from experiences of specific events. According to this interpretation, the results of our experiments show that semantic memory is illogical, too; that it also violates the logical rules that govern incompatibility relations.

In short, the broad implications of our results can be interpreted in two ways. On the one hand, they can be viewed as demonstrating that deep distortions extend from laboratory episodic memory to real-world episodic memory, and on the other hand, they can be viewed as demonstrating that deep distortions extend from episodic to semantic memory. It is also conceivable that for individual people, some DEM facts tap episodic memories of real-world experience, and others tap semantic memories of information to which people have not been directly exposed. If so, our results demonstrate that deep distortions extend to both real-world episodic memory and to semantic memory.

The task of deciding among these possibilities is far beyond the scope of this paper, but a final point should be mentioned that bears on this matter. A hypothesis that has been proposed in dual-trace models of episodic memory is that the episodic–semantic distinction may be a distinction without a difference because a separate semantic memory system is not required to explain most of the knowledge effects that are customarily attributed to it (e.g., schema abstraction, prototype retrieval, differential forgetting of surface and semantic content, semantic priming). Models such as FTT (Brainerd et al., 1999), retrieving effectively from memory (Shiffrin & Steyvers, 1997), storing and retrieving knowledge and events (Nelson & Shiffrin, 2013), and the dual-trace version of Minerva2 all assume that people store two types of episodic traces: traces of the literal surface form of experience (variously called verbatim, item, or exemplar traces) and traces of events' meanings, relations, and patterns (variously called gist, knowledge, or implicit traces). Simulation studies of such models (e.g., Johns et al., 2012) have shown that they can account for a broad range of classic effects that have traditionally been attributed to a semantic memory system that is distinct from episodic memory.

Constraints on Generality

The target population of Experiments 1 and 2 our experiments was university undergraduates. That population was chosen to preserve comparability with prior research on deep distortions with laboratory memory tasks, as the distortion effects in Table 3 were originally established with that populations, and subsequent experimentation has relied overwhelmingly on that population. As our objective was to determine those deep distortions are also present in memory for everyday facts, we began with that population. In Experiments 3 and 4, however, the target population was older, more diverse, nonuniversity individuals who had been identified by an online survey platform. The results for participants from this population were virtually identical to those for the university participants.

The research reported in this article was supported by United States Department of Agriculture Grant (NYC-321444). The deep distortion in everyday memory instrument and the raw data of the experiments are provided in the online supplemental materials. The research that we report has not been previously disseminated.

References

Abadie, M., & Camos, V. (2019). False memory at short and long term. *Journal of Experimental Psychology: General*, *148*(8), 1312–1334. https://doi.org/10.1037/xge0000526

Abadie, M., Waroquier, L., & Terrier, P. (2013). Gist memory in the unconscious-thought effect. *Psychological Science*, *24*(7), 1253–1259. https://doi.org/10.1177/0956797612470958

Abadie, M., Waroquier, L., & Terrier, P. (2017). The role of gist and verbatim memory in complex decision making: Explaining the unconscious-thought effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(5), 694–705. https://doi.org/10.1037/xlm0000336

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist*, *44*(9), 1185–1193. https://doi.org/10.1037/0003-066X.44.9.1185

Bookbinder, S. H., & Brainerd, C. J. (2017). Emotionally negative pictures enhance gist memory. *Emotion*, *17*(1), 102–119. https://doi.org/10.1037/emo0000171

Brainerd, C. J. (2021). Deep memory distortions. *Cognitive Psychology*, *126*, Article 101386. https://doi.org/10.1016/j.cogpsych.2021.101386

Brainerd, C. J. (2022). Deep distortions. *Memory*, *30*(1), 5–9. https://doi.org/10.1080/09658211.2020.1844756

Brainerd, C. J., Bialer, D. M., & Chang, M. (2022). Fuzzy-trace theory and false memory: Meta-analysis of conjoint recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(11), 1680–1697. https://doi.org/10.1037/xlm0001040

Brainerd, C. J., Nakamura, K., Chang, M., & Bialer, D. M. (2019). Super-overdistribution. *Journal of Memory and Language*, *108*, Article 104027. https://doi.org/10.1016/j.jml.2019.104027

Brainerd, C. J., Nakamura, K., & Murtaza, Y. (2020). Explaining complementarity in false memory. *Journal of Memory and Language*, *112*, Article 104105. https://doi.org/10.1016/j.jml.2020.104105

Brainerd, C. J., Nakamura, K., Reyna, V. F., & Holliday, R. E. (2017). Overdistribution illusions: Categorical judgments produce them, confidence ratings reduce them. *Journal of Experimental Psychology: General*, *146*(1), 20–40. https://doi.org/10.1037/xge0000242

Brainerd, C. J., & Reyna, V. F. (2005). *The science of false memory*. Cambridge University Press.

Brainerd, C. J., & Reyna, V. F. (2018). Complementarity in false memory illusions. *Journal of Experimental Psychology: General*, *147*(3), 305–327. https://doi.org/10.1037/xge0000381

Brainerd, C. J., Reyna, V. F., & Aydin, C. (2010). Remembering in contradictory minds: Disjunction fallacies in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 711–735. https://doi.org/10.1037/a0018995

Brainerd, C. J., Reyna, V. F., & Estrada, S. (2006). Recollection rejection of false narrative statements. *Memory*, *14*(6), 672–691. https://doi.org/10.1080/09658210600648449

Brainerd, C. J., Reyna, V. F., Gomes, C. F. A., Kenney, A. E., Gross, C. J., Taub, E. S., & Spreng, R. N. (2014). Dual-retrieval models and neurocognitive impairment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(1), 41–65. https://doi.org/10.1037/a0034057

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*(2), 413–439. https://doi.org/10.1037/a0025645

Brainerd, C. J., Reyna, V. F., & Mojardin, C. (1999). Conjoint recognition. *Psychological Review*, *106*(1), 160–179. https://doi.org/10.1037/0033-295X.106.1.160

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. https://doi.org/10.1016/j.jml.2015.06.006

Corbin, J. C., Reyna, V. F., Weldon, R. B., & Brainerd, C. J. (2015). How reasoning, judgment, and decision making are colored by gist-based: A fuzzy-trace theory approach. *Journal of Applied Research in Memory and Cognition*, *4*(4), 344–355. https://doi.org/10.1016/j.jarmac.2015.09.001

Diamond, N. B., Abdi, H., & Levine, B. (2020). Different patterns of recollection for matched real-world and laboratory-based episodes in younger and older adults. *Cognition*, *202*, Article 104309. https://doi.org/10.1016/j.cognition.2020.104309

Didyk, P., & Nieznański, M. (in press). Choice-induced preference change and episodic memory: Conscious recollection of choice restraints contrary to choice changes in ratings in the free-choice paradigm. *Consciousness and Cognition*.

Gallo, D. A. (2006). *Associative illusions of memory*. Psychology Press.

Gong, X., Xiao, H., & Wang, D. (2016). Emotional valence of stimuli modulates false recognition: Using a modified version of the simplified conjoint recognition paradigm. *Cognition*, *156*, 95–105. https://doi.org/10.1016/j.cognition.2016.08.002

Greene, N. R., Chism, S., & Naveh-Benjamin, M. (2022). Levels of specificity in episodic memory: Insights from response accuracy and subjective confidence ratings in older adults and in younger adults under full or divided attention. *Journal of Experimental Psychology: General*, *151*(4), 804–819. https://doi.org/10.1037/xge0001113

Greene, N. R., & Naveh-Benjamin, M. (2020). A specificity principle of memory: Evidence from aging and divided attention. *Psychological Science*, *31*(3), 316–331. https://doi.org/10.1177/0956797620901760

Greene, N. R., & Naveh-Benjamin, M. (2022). Effects of divided attention at encoding on specific and gist representations in working and long-term memory. *Journal of Memory and Language*, *126*, Article 104340. https://doi.org/10.1016/j.jml.2022.104340

Greene, N. R., & Naveh-Benjamin, M. (2023). Adult age-related changes in the specificity of episodic memory representations: A review and theoretical framework. *Psychology and Aging*, *38*(2), 67–86. https://doi.org/10.1037/pag0000724

Holliday, R. E., Brainerd, C. J., & Reyna, V. F. (2011). Developmental reversals in false memory: Now you see them, now you don't!. *Developmental Psychology*, *47*(2), 442–449. https://doi.org/10.1037/a0021058

Jacoby, L. L., & Shrout, P. E. (1997). Toward a psychometric analysis of violations of the independence assumption in process dissociation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(2), 505–510. https://doi.org/10.1037/0278-7393.23.2.505

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2012). A synchronization account of false recognition. *Cognitive Psychology*, *65*(4), 486–518. https://doi.org/10.1016/j.cogpsych.2012.07.002

Kassin, S. M. (2005). On the psychology of confessions: Does innocence put innocents at risk? *American Psychologist*, *60*(3), 215–228. https://doi.org/10.1037/0003-066X.60.3.215

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

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). NOVA Science Publishers.

Lampinen, J. M., Smith, A. M., & Wells, G. L. (2019). Four utilities in eyewitness identification practice: Dissociations between receiver operating characteristic (ROC) analysis and expected utility analysis. *Law and Human Behavior*, *43*(1), 26–44. https://doi.org/10.1037/lhb0000309

Lampinen, J. M., Watkins, K. N., & Odegard, T. N. (2006). Phantom ROC: Recollection rejection in a hybrid conjoint recognition signal detection model. *Memory*, *14*(6), 655–671. https://doi.org/10.1080/09658210600648431

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*(4), 361–381. https://doi.org/10.1002/bdm.752

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge University Press.

McNally, R. J., Clancy, S. A., & Schacter, D. L. (2001). Directed forgetting of trauma cues in adults reporting repressed, recovered, or continuous memories of childhood sexual abuse. *Journal of Abnormal Psychology*, *110*(1), 151–156. https://doi.org/10.1037/0021-843X.110.1.151

McRae, K., & Jones, M. (2013). Semantic memory. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology* (pp. 206–216). Oxford University Press.

Nachson, I., & Slavutskay-Tsukerman, I. (2010). Effect of personal involvement in traumatic events on memory: The case of the Dolphinarium explosion. *Memory*, *18*(3), 241–251. https://doi.org/10.1080/09658210903476530

Nelson, A. B., & Shiffrin, R. M. (2013). The co-evolution of knowledge and event memory. *Psychological Review*, *120*(2), 356–394. https://doi.org/10.1037/a0032020

Nieznański, M. (2020). Levels-of-processing effects on context and target recollection for words and pictures. *Acta Psychologica*, *209*, Article 103127. https://doi.org/10.1016/j.actpsy.2020.103127

Nieznański, M., Obidziński, M., Niedzialkowska, D., & Zyskowska, E. (2019). False memory for orthographically related words: Research in the simplified conjoint recognition paradigm. *The American Journal of Psychology*, *132*(1), 57–69. https://doi.org/10.5406/amerjpsyc.132.1.0057

Obidziński, M. (2021). Response frequencies in the conjoint recognition memory task as predictors of developmental dyslexia diagnosis: A decision-trees approach. *Dyslexia*, *27*(1), 50–61. https://doi.org/10.1002/dys.1655

Obidziński, M., & Nieznański, M. (2017). False memory for orthographically versus semantically similar words in adolescents with dyslexia: A fuzzy-trace theory perspective. *Annals of Dyslexia*, *67*(3), 318–332. https://doi.org/10.1007/s11881-017-0146-6

Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, *89*(3), 250–280. https://doi.org/10.1037/0033-295X.89.3.250

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

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*(3), 318–339. https://doi.org/10.1037/0033-295X.95.3.318

Schacter, D. L., Israel, L., & Racine, C. (1999). Suppressing false recognition in younger and older adults: The distinctiveness heuristic. *Journal of Memory and Language*, *40*(1), 1–24. https://doi.org/10.1006/jmla.1998.2611

Shiffrin, R. M., & Steyvers, M. (1997). Model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Singer, M., & Remillard, G. (2008). Veridical and false memory for text: A multiprocess analysis. *Journal of Memory and Language*, *59*(1), 18–35. https://doi.org/10.1016/j.jml.2008.01.005

Singer, M., & Spear, J. (2015). Phantom recollection of bridging and elaborative inferences. *Discourse Processes*, *52*(5–6), 356–375. https://doi.org/10.1080/0163853X.2015.1029858

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*(1), 1–25. https://doi.org/10.1146/annurev.psych.53.100901.135114

Uhl, J., Schaffrath, J., Schwartz, B., Poster, K., & Lutz, W. (2022). Within and between associations of clinical microskills and correct application of techniques/strategies: A longitudinal multilevel approach. *Journal of Consulting and Clinical Psychology*, *90*(6), 478–490. https://doi.org/10.1037/ccp0000738

Unsworth, N., McMillan, B. D., Brewer, G. A., & Spillers, G. J. (2013). Individual differences in everyday retrospective memory failures. *Journal of Applied Research in Memory and Cognition*, *2*(1), 7–13. https://doi.org/10.1016/j.jarmac.2012.11.003

White, R. (2021). Naturalistic studies of long-term autobiographical memory. *Applied Cognitive Psychology*, *35*(6), 1641–1643. https://doi.org/10.1002/acp.3891

Wolfe, C. R., & Reyna, V. F. (2010). Semantic coherence and fallacies in estimating joint probabilities. *Journal of Behavioral Decision Making*, *23*(2), 203–223. https://doi.org/10.1002/bdm.650

(*Appendix follows*)

Appendix

Basic Model

Let $p(\mathrm{A})$, $p(\mathrm{B})$, $p(\mathrm{C})$, $p(\text{A-or-B})$, $p(\text{C-or-A})$, and $p(\text{B-or-C})$ be the respective probabilities of accepting each of the six memory probes for any DEM item. Let V_{A} be the probability of successful retrieval of a verbatim memory of learning that candidate A is a true fact, which produces acceptance of the A, A-or-B, and C-or-A probes but rejection of the other three probes. Let G_{A} be the probability of successful retrieval of a gist memory of the target fact with any probe that contains candidate A, which produces acceptance. Let G_{B} be the probability of successful retrieval of a gist memory of the target fact with any probe that contains candidate B, which produces acceptance. Let G_{C} be the probability of successful retrieval of a gist memory of the target fact with any probe that contains candidate C, which produces acceptance. Let b be a bias parameter that produces acceptance of any plausible probe when verbatim and gist retrieval both fail. (Remember that in the DEM, the two false candidates are not unrelated but are plausible alternatives.) "Successful retrieval" refers to the fact that verbatim and gist retrieval are both threshold processes that proceed in parallel (see Brainerd et al., 2022): Both verbatim and gist information accumulate during the course of retrieval, but the strength of one of them must cross a threshold for a response (acceptance or rejection) to occur. If neither does so, the response is indeterminate, and participants must resort to bias processes (measured by parameter b in Equations A1–A6).

The six judgment probabilities are expressed as a function of verbatim, gist, and bias parameters as follows:

$$p(\mathrm{A}) = V_{\mathrm{A}} + (1 - V_{\mathrm{A}})G_{\mathrm{A}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{A}})b, \tag{A1}$$

$$p(\mathrm{B}) = (1 - V_{\mathrm{A}})G_{\mathrm{B}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{B}})b, \tag{A2}$$

$$p(\mathrm{C}) = (1 - V_{\mathrm{A}})G_{\mathrm{C}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{C}})b, \tag{A3}$$

$$p(\mathrm{A\text{ - or - }B}) = V_{\mathrm{A}} + (1 - V_{\mathrm{A}})G_{\mathrm{A}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{A}})G_{\mathrm{B}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{A}})(1 - G_{\mathrm{B}})b, \tag{A4}$$

$$p(\mathrm{A\text{ - or - }C}) = V_{\mathrm{A}} + (1 - V_{\mathrm{A}})G_{\mathrm{A}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{A}})G_{\mathrm{C}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{A}})(1 - G_{\mathrm{C}})b, \text{ and} \tag{A5}$$

$$p(\mathrm{B\text{ - or - }C}) = (1 - V_{\mathrm{A}})G_{\mathrm{B}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{B}})G_{\mathrm{C}} + (1 - V_{\mathrm{A}})(1 - G_{\mathrm{B}})(1 - G_{\mathrm{C}})b. \tag{A6}$$

The data space contains six degrees of freedom (the six judgment probabilities), whereas the model contains five theoretical quantities (the four retrieval parameters plus one bias parameter). Therefore, the model can be fit to the data of Experiments 1–4 with the likelihood ratio L_5/L_6, where L_6 is the a posteriori probability of a set of data when all six empirical probabilities are free to vary and L_5 is the a posterior probability of the same data when only the five theoretical quantities are free to vary. Here, $-2\ln(L_5/L_6)$ is a $G^2(1)$ statistic with a critical value of 3.84, which tests the null hypothesis that the model fits the target data against the hypothesis that a more complex model is needed.

When the model fits a set of target data, hypotheses about the relative contributions of verbatim and gist retrieval to memory judgments can also be tested with likelihood ratios of the form L_4/L_5. Here, L_5 is the a posteriori probability of the target data when all five theoretical quantities are free to vary, and L_4 is the a posteriori probability of the same data when a single constrain has been imposed on them that follows from some hypothesis about the relative contributions of verbatim and gist retrieval. Theoretically important examples of such hypotheses are (a) that verbatim retrieval is less probable than any form of gist retrieval, so that V_{A} must be smaller than G_{A} or G_{B} or G_{C}, and (b) gist retrieval is more probable with a correct fact than an incorrect one, so that G_{A} must be larger than G_{B} or G_{C}. The test statistic for such hypotheses is $-2\ln(L_4/L_5)$, which is a $G^2(1)$ statistic with a critical value of 3.84. The relations between likelihood ratios and this test statistic are spelled out in some theorems that appear in Riefer and Batchelder (1988). Briefly, when a ratio of two likelihoods is constructed, -2 times the natural log of the likelihood ratio is asymptotically χ^2, so that the statistic tests the null hypothesis that the two likelihoods do not differ reliably.

In the literature on conjoint recognition, all such models contain parameters that measure retrieval of verbatim traces of true items (V_{A} in this model), parameters that measure retrieval of gist traces (the parameters G_{A}, G_{B}, and G_{C} in this model), and parameters that measure nonmemory bias acceptances (b in this model). Similar to the DEM, many of those models are defined over experimental procedures that provide six degrees of freedom (e.g., Abadie & Camos, 2019; Greene & Naveh-Benjamin, 2020; Nieznański, 2020), so that the maximum number of parameters that can be estimated and still test fit is five. However, there is an alternative paradigm that provides nine degrees of freedom. There, it is traditional to measure two additional retrieval parameters (for a review, see Brainerd et al., in press)—namely, phantom recollection (illusory vivid memories of false items) and erroneous recollection rejection (illusory vivid rejection of true items).

Some further comments are in order about the reliability with which different parameters are estimated in multinomial models such as the present one. The reliability with which different parameter are estimated depends on how much of the target data are used in the estimation process, with reliability increasing as more data are used. For any parameter P, its reliability depends on how many terms of a model's equations contain the parameter or its inverse $1 - P$. In Equations A1–A6, it can be seen that the present model has a total of 19 terms. V_{A} or $1 - V_{\mathrm{A}}$ appear in all of them, each gist parameter or its inverse appear in eight of them, and the bias parameter appears in six of them. Thus, estimates of the verbatim parameter are more reliable than estimates of the other parameters.

It should be noted that in the conjoint-recognition literature, a multiple-choice testing paradigm has also been used to measure the model's verbatim, gist, and bias parameters (see Abadie & Camos, 2019; Abadie et al., 2017; Greene & Naveh-Benjamin, 2020, 2022, 2023; Nieznański, 2020; Nieznański et al., 2019). In that paradigm, participants are presented with all the candidates for a target fact and asked to choose among them (Is Jupiter, Saturn, or Earth the largest planet?) This alternative testing procedure cannot be used with our

(*Appendix continues*)

experiments because it provides only two degrees of freedom, whereas at least six degrees of freedom are required to estimate five parameters (a verbatim parameter, three gist parameters, and a bias parameter) and test fit.

Predicting Violations of Additivity

For two incompatible candidate facts X and Y, let $p(\mathrm{X})$, $p(\mathrm{Y})$, and $p(\text{X-or-Y})$ be the probabilities of remembering that X is true, that Y is true, and that their disjunction is true, where X is the true fact. For the additivity property to hold, the model must predict that $p(\mathrm{X}) + p(\mathrm{Y}) = p(\text{X-or-Y})$. Model expressions for these quantities appeared above. Substituting those expressions for $p(\mathrm{X})$, $p(\mathrm{Y})$, and $p(\text{X-or-Y})$ yields

$$V_X + (1 - V_X)(G_X + G_Y) + (1 - V_X)(2 - G_X - G_Y)b > V_X + (1 - V_X) G_X + (1 - V_X)(G_Y - G_X) + (1 - V_X)(1 - G_X)(1 - G_Y)b. \tag{A7}$$

where X is the correct fact and Y is incorrect, so that additivity is violated in the subadditive direction. To see that this inequality holds, we subtract the quantity on the right side from the quantity on the left side, which must equal zero for additivity to be preserved. When the subtraction is performed, however, the resulting quantity is $(1 - V_X)G_XG_Y + b(1 - V_X)(1 - G_XG_Y)$, which must be greater than 0 as long as $V_X < 1$, $G_X > 0$, and $G_Y > 0$. Note that if participants do not rely on noncompensatory gist (i.e., the gist parameters are zero) and only rely on compensatory verbatim retrieval, the additivity property holds because Equation A7 becomes

$$V_X(1 - V_X)b = V_X(1 - V_X)b. \tag{A8}$$

When the gist parameters in Equation A7 are restored, the value of $p(\mathrm{X}) + p(\mathrm{Y})$ becomes larger than the value of $p(\text{X-or-Y})$.

If X and Y are both incorrect, Equation A7 becomes

$$(1 - V_X)(G_X + G_Y) + (1 - V_X)(2 - G_X - G_Y)b > (1 - V_X)G_X + (1 - V_X)(G_Y - G_X) + (1 - V_X)(1 - G_X)(1 - G_Y)b. \tag{A9}$$

This inequality holds because subtraction of the quantity on the left side from the quantity on the right side leaves the residual $(1 - V_X)G_Y + (1 - V_A)(1 - G_XG_Y)b$. Thus, the model predicts that additivity will be violated for any pair of mutually incompatible candidates and that violations will be in the subadditive direction rather than the superadditive direction.

Predicting Violations of Countable Additivity

For three incompatible facts X, Y, and Z, where X is correct and Y and Z are both incorrect, let $p(\mathrm{X})$, $p(\mathrm{Y})$, $p(\mathrm{Z})$, and $p(\text{X-or-Y-or-Z})$ be the respective probabilities of remembering that X is true, that Y is true, that Z is true, and that their disjunction is true. For the countable additivity property to hold, the model must predict that $p(\mathrm{X}) + p(\mathrm{Y}) + p(\mathrm{Z}) = p(\text{X-or-Y-or-Z})$. Model expressions for the terms on the left side of the equation appear above. The model's expression for $p(\text{X-or-Y-or-Z})$ is

$$p(\text{X-or-Y-or-Z}) = V_X + (1 - V_X)(G_X + (1 - G_X)G_Y + (1 - G_X)(1 - G_Y)G_Z) + (1 - V_X)(1 - G_X)(1 - G_Y)(1 - G_Z)b. \tag{A10}$$

Substituting the model's expressions for $p(\mathrm{X})$, $p(\mathrm{Y})$, $p(\mathrm{Z})$ and $p(\text{X-or-Y-or-Z})$ yields

$$V_X + (1 - V_X)(G_X + G_Y + G_Z) + (1 - V_A)(3 - G_X - G_Y - G_Z) b > V_X + (1 - V_X)[G_X + (1 - G_X)G_Y + (1 - G_X)(1 - G_Y)G_Z)] + (1 - V_X)(1 - G_X)(1 - G_Y)(1 - G_Z)b. \tag{A11}$$

This inequality holds because subtraction of the expression on the right side of the inequality from the expression on the right side does not equal 0 but, rather, leaves a residual quantity. As with additivity, then, the model predicts that countable additivity will be violated and that violations will be in the subadditive direction. This will be true as long as $V_X < 1$, $G_X > 0$, $G_Y > 0$, and $G_Z > 0$. Note that Equation A11 can be expanded to include any arbitrary number of incompatible candidate facts by simply writing the model's expression for the additional facts.

Predicting Violations of Universal Event

For three incompatible facts X, Y, and Z, such that X is correct and both Y and Z are incorrect, let $p(\mathrm{X})$, $p(\mathrm{Y})$, and $p(\mathrm{Z})$ be the probabilities of remembering that X is true, that Y is true, and that Z is true. For the universal event property to hold, the model must predict that $p(\mathrm{X}) + p(\mathrm{Y}) + p(\mathrm{Z}) \leq 1$. However, the model's expression for the sum of the three remembering probabilities is

$$V_X + (1 - V_X)(G_X + G_Y + G_Z) + (1 - V_X)(3 - G_X - G_Y - G_Z)b. \tag{A12}$$

For the universal event property to be preserved, the sum of the terms in this expression must be ≤1 for all values of the five parameters. However, algebraic manipulation reveals that the upper bound of this expression is not 1. Rather, holding the value of the bias parameter constant, the value of the expression overflows 1 as the value of the verbatim parameter decreases and the values of the gist parameters increase.

Received December 4, 2022
Revision received October 17, 2023
Accepted November 11, 2023 ■