BELIEVABLE BUT NOT MEMORABLE:

EXAMINING THE INTERACTION BETWEEN THE BELIEVABILITY AND MEMORABILITY OF EVIDENCE AS IT AFFECTS INFERENCES

by

Jason David Ozubko

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Arts
in
Psychology

Waterloo, Ontario, Canada, 2007

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Dual-process theories of reasoning (e.g., Gilbert, 1991; Stanovich & West, 1997) posit that decisions are mediated by two cognitive systems: a fast and automatic system which sometimes relies on past knowledge, and a conscious and effortful system which is more likely to adhere to the rules of logic. Dual-process accounts of memory (e.g., Joordens & Hockley, 2000) suggest that memory is influenced by two cognitive systems: a fast and automatic familiarity component, and a conscious and effortful recollection component. Both accounts suggest that cognition relies on two underlying systems, which are described similarly in the two literatures, suggesting some form of convergence in these two areas of research. Memory research may therefore be informed by considering decision making research, and vice versa. Combining these two theoretical perspectives, it follows that believable evidence should be less memorable than unbelievable evidence due to its shallow initial processing. Despite this fact however, when inferences are being made based on evidence retrieved from memory, believable evidence should actually have a larger impact than it does when it is provided online, whereas no change or a lesser impact should be noted for unbelievable evidence. Across 3 experiments these predictions are validated, suggesting that the impact of evidence on inferences depends not only on the believability of that evidence, but also on whether the decision is being made online or from memory. Specifically, memory-based inferences exaggerate the influence of believable but not unbelievable evidence, despite the fact that unbelievable evidence is more memorable.

Acknowledgments

I would like to thank Jonathan Fugelsang, Derek Koehler, and the University of Waterloo Department of Psychology, Cognition division for their support. I would also like to thank Jason Locklin for his assistance in designing web-based experiments and collecting data.

Table of Contents

List of Tables

Introduction

Several lines of research have begun to converge on the notion that, when evaluating information, humans are inherently sensitive to the believability of that information. For example, in the Wason Selection task (e.g., Griggs & Cox, 1982) participants must examine a set of cases to ensure that they do not violate some set rules. A typical example would be to have 4 cards presented, A, D, 3, and 7, and to inform participants that each card has a number on one side and a letter on the other. Furthermore, a rule such as "If there is an A on one side of the card then there must be a 3 on the other side" would be given. Usually, participants are unable to correctly solve this puzzle in its current, abstract form (the solution is to check the A and 7 card). If the task is given semantic content however, the task becomes much more easily solved. For example if participants are told they are police officers searching a bar for under-age drinkers, and the numbers and letters are changed to represent age and type of drink being consumed, participants are about 75% successful in the task (Griggs and Cox, 1982). Thus, it seems that prior knowledge plays some role in how participants conceptualize this task.

Fugelsang and Thompson (2000; 2001) have also demonstrated that when evaluating potential causes, participants are sensitive not just to the degree of covariation between the potential cause and the effect, but also to the causes' believability. In fact, given two potential causes (a candidate and an alternative cause), participants are more likely to discount the candidate cause and endorse the alternative cause if the alternative cause is believable, even if the alternative cause covaries with the effect to a lesser degree than the candidate cause (Fugelsang & Thompson, 2001).

Finally, Evans and colleagues (Evans, Barston, & Pollard, 1983; Evans & Curtis-Holmes, 2005) have demonstrated that when solving syllogisms, believability and validity interact such that the role of logic plays a smaller role for syllogisms with believable conclusions. That is, syllogisms with believable conclusions are more likely to be accepted than those with unbelievable conclusions. Furthermore, the validity of conclusions has a smaller effect on acceptance when syllogisms are believable, suggesting logical considerations are adhered to less when believable conclusions are presented than when unbelievable conclusions are presented. Thus, humans do not appear to evaluate information in an abstract, purely analytical fashion, but are inherently affected by the believability of information.

Dual-Process Models of Human Reasoning

Gilbert (1991) distinguishes between two possible models of human cognition, a Cartesian and Spinozan model. According to a Cartesian model, individuals learning new information first understand the information, then in a subsequent stage of processing either accept or reject that information as true[1]. A Spinozan model on the other hand, suggests that the act of learning new information also necessarily entails acceptance of that information as veridical, and that in the subsequent stage of processing individuals can possibly reject the information. Gilbert cites a wide range of psychological research, from developmental research to sleep deprivation to linguistic research, that all converges on the notion that the Spinozan model is a more accurate model of human cognition. The interesting aspect of this Spinozan account is it suggests

[1] Although the accounts presented by Gilbert focus on the truth of information, we will use the terms "truth" and "believability" interchangeably when discussing this model, under the assumption that, to the extent that individuals believe something is true, subjectively it is. Thus, the notions of believability and subjective truth are functionally equivalent in our perspective.

that human cognition is indefinitely intertwined with believability, to the point where new information cannot even be understood without first being believed.

From a Spinozan account then, information that is initially processed is automatically accepted as true. This acceptance also is likely unconscious, as it occurs as soon as information is understood, and individuals often understand concepts without much conscious effort (see Gilbert, 1991 for a discussion on this point). Obviously however, individuals do not believe everything that they hear. Hence Gilbert (1991) points out that the second phase of processing in the Spinozan account is where participants can reject ideas as unbelievable or false. Interestingly, this second phase of processing appears to be an effortful, conscious process as experimental manipulations that introduce a cognitive load have been demonstrated to reduce participants' abilities to reject false information (e.g., Gilbert, Tafarodi, & Malone, 1993). This dual-processing account of human reasoning therefore suggests that information is first processed in an automatic and potentially unconscious manner, such that information is both understood and accepted. During a subsequent processing stage, which relies on conscious attention and effort, information that is unbelievable can be rejected.

Stanovich and West (1997) have also put forth a dual-process account of human reasoning that shares many parallels with this Spinozan account and is actually theoretically compatible. Stanovich and West however focus more on the details of the processing of the two systems, rather than the order in which they occur. According to Stanovich and West, in System 1 (or the first processing stage of a Spinozan account) information is processed in an automatic, fast, and parallel manner. It is this system which operates primarily based on prior beliefs and knowledge (among other influences,

but we will focus on these). Furthermore, the processing of System 1 is not privy to conscious introspection, as only the output of this system is posted to consciousness. System 2 (or the second phase of processing in a Spinozan account) on the other hand, is a slow, deliberate, and serial processing system. It is this system that takes into account abstract notions such as logic. This system is conscious and open to introspection, but it also requires effort and attention, and is limited by working memory capacity.

According to Stanovich and West (1997), it is the interaction of these two systems that give rise to belief biases. Whereas processing by System 2 can lead to logically valid solutions to problems, System 1 processing will result in believable, but not necessarily logical, solutions being endorsed. Indeed, experimental findings, such as those discussed previously on syllogisms (e.g., Evans et al., 1983) and causal relations (e.g., Fugelsang & Thompson, 2000), support the idea that although participants are sensitive to logical considerations, they are also biased by the believability of information. Considering both Stanovich and West's dual-process account and a Spinozan account together, we can arrive at the following general account of information processing: New information is first processed in System 1, whereby the default action is to understand and accept the idea. Subsequently, in System 2, the newly acquired information is consciously scrutinized based on logical considerations. To the degree that the information is believable however, it may receive little or no processing in System 2. Thus, believable information is more likely to be processed primarily by System 1, whereas unbelievable information is more likely to be passed on from System 1 and processed more thoroughly by System 2.

Past researchers have made similar claims about the differential processing of believable and unbelievable information (e.g., Evans, 2003) and some empirical findings support the notion that believable information may be processed primarily in System 1 whereas unbelievable information is more likely to receive additional processing in System 2. For example, using a response deadline paradigm, Evans and Curtis-Holmes (2005) found that being forced to respond quickly primarily affected the accuracy of judging syllogisms with valid/unbelievable conclusions. Although the speeded responding also affected syllogisms with invalid/believable conclusions, the effect was larger on the valid/unbelievable syllogisms. If we assume that under speeded responding, the slower analytic system (i.e., System 2) has a weaker effect on responses, the larger impact on unbelievable syllogisms suggests that System 2 plays a larger role in processing unbelievable syllogisms than believable syllogisms.

Thus, the dual-process account of reasoning we wish to focus on is one that combines the considerations of both a Spinozan model and of Stanovich and West's (1997) System 1 and System 2 approach. From our consideration, these two accounts of reasoning together suggests that information is first processed in System 1, and to the degree that it is unbelievable it may face more complex, conscious scrutiny in System 2. As a result, believable information is processed primarily in System 1 whereas unbelievable information is more likely to receive additional processing in System 2.

Interestingly, in the area of memory research, dual-process accounts have also been proposed (e.g., Joordens & Hockley, 2000; Reder, Nhouyvansivong, Schunn, Ayers, Angstadt, & Hiraki, 2000). These dual-process accounts of memory suggest that memory performance may rely on two underlying systems: a fast and automatic familiarity

component, and a conscious and effortful recollection component. These two systems actually have many similarities with the two systems proposed by dual-process accounts of decision making. Thus, it may be that the two areas of research, decision making and memory, have independently begun to converge. If this is true, then both the areas of memory and decision making may be informed by one another. Thus, we now turn to a consideration of the memory literature to examine how a dual-process account of decision making may fit with a dual-process account of memory.

Dual-Process Reasoning and Memory

Hastie and Park (1986) distinguish between two types of reasoning tasks: online and memory-based. Online reasoning tasks provide participants with all necessary information to make decisions on each trial. For example, Evans et al. (1983) provided participants with sets of premises and asked participants to judge how accurate the provided conclusions were. This is an example of an online task, since all relevant information is available during the decision phase. A memory-based reasoning task however requires participants to learn some information and later to make an inference based on that information. To prevent participants from making online judgments as they learn information, participants are usually not informed what they will be using the information to decide about, or sometimes are not even told they will need to use the information to make a decision. For example, in one experiment Hastie and Park had participants listen to a 5-min conversation between two men, and then subsequently participants judged the suitability of one of the men for a computer programming job. Hence, participants had to remember the conversation, but were unaware what the

information would be used for. At test then, participants had no choice but to rely on their memory of the conversation to make their decision.

Although most studies in the area of judgment and reasoning focus on online tasks, memory-based tasks may be particularly relevant to real world situations. For example, individuals in the real world often have to make decisions based on information that is not currently being presented to them. Be it to successfully answer an exam question or to decide how much you enjoyed a movie, memory plays an important role in many decisions throughout our lives. Of particular relevance given our previous discussion of believable and unbelievable information, may be how the believability of information interacts with its memorability.

In a typical recognition memory experiment participants are shown a set of items to memorize during a study phase. During the test phase, memory is tested by asking participants to classify items as either "old" (i.e., previously presented in the study phase) or "new". Correctly calling an old item "old" is called a hit whereas mistakenly calling a new item "old" is called a false alarm, and from these two measures memorability can be estimated. Dual-process accounts of recognition memory (not to be confused with dual-process accounts of reasoning) propose that performance on recognition memory tests can be explained in terms of recollection and familiarity (e.g., Joordens & Hockley, 2000; Reder et al., 2000).

In terms of familiarity, it is assumed that items that feel familiar are simply more likely to elicit "old" responses regardless of their old/new status. Familiarity is assumed to be influenced by factors such as perceptual fluency (Jacoby & Dallas, 1981), in the sense that items that are easier to process are more likely to feel familiar to participants.

Furthermore, familiarity is an automatic and unconscious process (Yonelinas & Jacoby, 1996), such that participants believe familiar items are old, regardless of whether they actually are or not. Thus, the familiarity of a stimulus set acts to both increase the hits and false alarms. Nonetheless, items that were seen at study should feel more familiar to participants than foils, since the studied items were recently experienced (Jacoby & Whitehouse, 1989; Joordens & Merikle, 1992), therefore, based solely on familiarity based responding, hits should be higher than false alarms.

However, a second factor, recollection, also helps increase hit rates. Recollection represents the degree to which items can be consciously retrieved from memory. As only studied items can be retrieved from memory, recollection acts only on old items and therefore boosts hit rates. The degree of conscious recollection is often attributed to the degree of conscious attention or processing given during study (Joordens & Hockley, 2000; Reder et al., 2000). Thus, items that are distinctive (e.g., Geraci & Rajaram, 2002; Schmidt, 1991; Valentine, 1991) are often more consciously recollectable than other items.

Hence, both familiarity and recollection can be used to recognize studied items. However, familiarity and recollection are assumed to be quite different. Familiarity is believed to be more of an automatic, unconscious process (Jacoby & Whitehouse, 1989) whereas recollection is believed to be an effortful, conscious process (Joordens & Hockley, 2000; Reder et al., 2000). In terms of the results of a recognition memory test, one manner in which familiarity and recollection can interact is to produce a mirror effect.

The frequency-based mirror effect is the finding that low frequency words are more memorable than high frequency words, and that specifically, low frequency words have both a higher hit rate and lower false alarm rate than high frequency words (Glanzer & Adams, 1985). In terms of the general memorability of believable versus unbelievable information, we may expect to see a similar pattern. That is, the frequency-based mirror effect is usually explained by the dual-process account of memory by assuming that high frequency words have a higher false alarm rate than low frequency words because high frequency words are more familiar due to pre-experimental experience (Joordens & Hockley, 2000; Reder et al., 2000). Based solely on familiarity we may expect both higher false alarm and hit rates for high frequency words. But, low frequency words are assumed to be more distinctive than high frequency words as low frequency words are experienced less often (e.g., Geraci & Rajaram, 2002). As a result, these items are easier to recollect than high frequency words therefore hit rates for low frequency items is boosted, beyond that of high frequency words, resulting in the full mirror effect.

Our original rationale for considering memory literature was that dual-process theories of memory may be able to inform dual-process theories of decision making, and vice versa. Thus, the frequency-based mirror effect may have an analogy in the decision making literature. Specifically, turning to believable and unbelievable information we may expect to see a similar mirror effect.

That is, the greater familiarity of high frequency words is hypothesized to give rise to the higher false alarms for these items. According to our dual-process account of reasoning, when new information is perceived it is initially processed in System 1. Although both believable and unbelievable information may be initially processed in

System 1, believable information may be processed more fluently in this system. That is, given that System 1 is assumed to rely on prior knowledge and beliefs, information that is consistent with these beliefs may be more fluently processed in this system than information which is not (i.e., unbelievable information). As was mentioned previously, perceptual fluency is one factor that can help increase subjective feelings of familiarity (Jacoby & Dallas, 1981), however others have advocated conceptual fluency as having similar effects on familiarity (e.g., Gregg, Gardiner, Karayianni, & Konstantinou, 2006). Thus, if believable information is processed more fluently in System 1 than unbelievable information, we would expect believable information to generally be more familiar than unbelievable information. As a result, believable information should have both a higher hit rate and false alarm rate than unbelievable information, based solely on familiarity.

However, based on conscious recollection, we may expect the hit rates for unbelievable information to actually surpass those of believable information, giving rise to a full mirror effect. That is, unbelievable information was hypothesized previously to be more likely to receive additional processing in System 2 than believable information. Because System 2 processing was proposed to be conscious and attentive, System 2 processing should be more likely to give rise to conscious recollection. Therefore, unbelievable information should be more recollectable than believable information, which should result in a higher hit rate for these items, and a full mirror effect. In line with this suggestion, Hastie and Kumar (1979) found that when participants were rating the personalities of hypothetical individuals, sentences describing characteristics which were incongruent with the final trait rating were more memorable. Hastie and Kumar

attributed these results to extra conscious processing these incongruent sentences elicited as participants attempted to reconcile the incongruent traits with the final trait rating.

Thus, considering memory generally, it appears as if unbelievable information should be more memorable than believable information, but that specifically we may expect to see a mirror effect (i.e., more hits and fewer false alarms to unbelievable than believable information). Additionally, the mechanisms which we have suggested that would give rise to this mirror effect are consistent with both models of memory and reasoning put forth. That is, the idea that System 1 would give rise to familiarity is consistent with how both System 1 and familiarity have been characterized, as both have been suggested to be unconscious and automatic processes. Furthermore, System 2 was previously described as an effortful and conscious processing of information, precisely the type of processing which was suggested to give rise to conscious recollection. Thus, both memory and reasoning accounts seem to coincide nicely, and predict that a mirror effect should be observed between believable and unbelievable information, with superior memory for unbelievable information.

The Impact of Memory on Evidence (Premises)

Similar to some past work on logical reasoning (e.g., Evans et al., 1983), the current experiments were designed as reasoning tasks where participants would make judgments about the accuracy of some conclusion statements based on some premises. However, work on logical reasoning generally has focused on the believability of conclusions. For example, in the syllogistic reasoning task used by Evans et al., participants are presented with two premises and a conclusion and must make some decision about the validity of the conclusion. The task is setup such that the conclusions

are either believable or unbelievable, whereas the two premises together are always neutral in terms of believability. Our interest was on that of the believability of premises, not of conclusions.

Thompson (1996) points out that while a large body of work has investigated how the believability of conclusions affects decisions, significantly less work has been done investigating the believability of premises. Thompsons herself investigated the believability of premises and found that believable premises had a larger impact on decisions than did unbelievable premises. Given that this area is relatively understudied and that manipulating the believability of premises allowed us to easily include a memory aspect to our experiments (i.e., participants could memorize believable or unbelievable premises and later judge conclusions based on those premises), it seemed reasonable to focus on believability of premises rather than of conclusions.

On a more practical note however, focusing on the premises rather than the conclusions allows us to examine how the memorability of information affects decisions which are subsequently made. Had we instead chose to focus on the memorability of conclusions, the memorability information would inform us about previously made decisions, and not about how any information will affect upcoming decisions. Thus, by focusing on premises rather than conclusions we feel we are better able to combine the areas of memory and decision making.

Another deviation from typical decision making methodology is that we chose not to use logical syllogisms but rather to use a less formal reasoning task. That is, some work has demonstrated that participants may not generally perform logical tasks correctly. For example, Evans et al. (1983) point out that data from some participants

appear to indicate that these participants based their judgments about the accuracy of conclusions based solely on the conclusions themselves, ignoring the premises which they were supposed to evaluate. Dickstein (1981) examined participants' performance on a logical reasoning task and found, consistent with past work, that participants were unable to correctly reason through some types of syllogistic problems. Dickstein argues that a large part of these errors are due to participants' inability to differentiate between possible and necessary conclusions. Finally, Roberts and Sykes (2003) across two experiments demonstrated that participants were unable to use logic correctly to reason through syllogisms. Although a third experiment by Roberts and Sykes did demonstrate that participants may be able to use logic correctly in some situations, the bulk of their results simply reinforced that notion that most participants misunderstand or misuse the rules of logic.

As a result of the potential problems with using logical syllogisms, we moved to an inferences task in which there are no logically correct answers. Participants received evidence statements that could either be believable or unbelievable and then had to make an inference based on those statements. Thus, from this point on we will no longer discuss "premises" or "conclusions" but instead discuss "evidence" and "inferences", as these terms more accurately describe our paradigm. As there were no logically correct answers to any of these questions, we used normative data to judge participants' performance. Furthermore by focusing on the believability of evidence instead of the inferences, we could easily introduce a memory element to our experiments, namely by having participants memorize evidence which would later be used to make inferences.

Current Hypotheses

If believable information is more likely to be processed primarily in System 1, we hypothesized this would lead to more familiarity-based responding to believable evidence on a memory test and in the end, result in poorer memory for believable evidence compared to unbelievable evidence. However, when believable evidence is being retrieved from memory for the purposes of making an inference, this evidence may actually have a larger impact on inferences than when used online. That is, when encoding believable evidence, that evidence will be largely processed in System 1, where information is processed automatically and integrated with prior knowledge and beliefs. Thus, it seems clear why believable evidence would not be very memorable if processed in System 1, as by being integrated with past knowledge, specifics and details of the information encoded may be lost, while the general gist or message of that information may be what is retained.

When it comes to retrieving believable evidence, even though the exact verbatim item which was encoded may not be retrievable, many consistent and agreeable notions may be easily retrieved upon the attempt. That is, the prior beliefs which the believable evidence was integrated with may be accidentally retrieved as participants attempt to retrieve the specific, believable piece of information. Thus, even if participants fail to retrieve the specific believable evidence statement that was encoded during study, they may inadvertently retrieve numerous other pieces of information that all are consistent with the encoded item. As a result, this may exaggerate the impact of believable evidence, as participants will have far more believable evidence to rely on than they would in an online task. For example, imagine a participant is trying to retrieve the

believable evidence statement, “8% of men are color blind”. Even if this specific item cannot be successfully retrieved, the retrieval attempt may begin to result in prior knowledge consistent with this statement being retrieved instead. So facts like, “men don’t usually have good taste in color”, “my Dad never wears clothes that color-coordinate”, “My brother can’t tell the different between red and orange” may all be retrieved.

Prior research into item similarity has shown that when participants encounter a series of items which are easily integrated together, later memory tests demonstrate that participants may falsely recollect items that, although consistent with the set itself, were never presented. For example, Roediger and McDermott (1995) demonstrated that when participants were shown a series of words (e.g., TIRED, PILLOW, BED) that were all related to an unpresented, critical lure (e.g., SLEEP), participants later falsely recollected the critical lure at test. Roediger and McDermott suggested that as participants saw and integrated items at study, they inadvertently were also activating the critical lure in memory, and hence, that item was easily retrieved at test, even though it had not actually been presented (see also Deese, 1959 for similar findings).

Thus, in terms of believable evidence, these findings suggest that if information can be easily integrated together, it becomes difficult for participants to correctly identify exactly which specific items were presented and which were not. And critically, while attempting to retrieve certain items they may be inadvertently activating and perhaps even retrieving other items. Thus, as believable evidence has been integrated with prior knowledge, any attempt to retrieve a specific believable evidence statement may inadvertently activate or retrieve other pieces of evidence that are consistent with the

evidence statement being sought. The result is participants may actually end up with several pieces of evidence to support an inference, when they attempt to retrieve a single believable evidence statement. Thus, believable evidence may have a larger impact on inferences when retrieved from memory.

In terms of unbelievable evidence, we hypothesized this type of information should receive more processing in System 2 than believable information, thus eliciting more conscious attention and analytic processing. This may result in information being encoded in a more analytic or abstract but conscious manner. Furthermore this type of processing may be relatively free from the constraints of System 1. That is, System 2 processing would not necessarily integration information with prior knowledge or beliefs. Thus, System 2 processing should result in encoding that does not integrate information with prior knowledge or beliefs, but instead allocates more analytic and conscious effort to the information. The result is that unbelievable evidence should be more consciously recollectable later on, than believable evidence.

In terms of memory-based reasoning tasks, the fact that unbelievable evidence is consciously encoded in a recollectable manner may result in unbelievable evidence having an equivalent effect in both memory-based and online tasks. That is, when unbelievable evidence is initially processed in System 1, it may not integrate well with prior knowledge since it is, by definition, inconsistent with prior knowledge. Furthermore, subsequent processing in System 2 likely considers the information in a more abstract manner than System 1. Although this processing may lead to better conscious recollection, it likely would not result in integration with prior knowledge.

Thus, unbelievable evidence may be encoded in such a way that specific surface details are retained.

This more analytic and conscious encoding would support conscious recollection, and thus be compatible with our previous characterization of how unbelievable information is memorized. A result of this less belief-integrated encoding however, would be that when participants attempt to retrieve an unbelievable piece of evidence, they would either retrieve the exact item from study, or else fail to recollect anything. If an unbelievable piece of evidence was successfully retrieved, it should then have a similar effect on inferences as when it is provided in an online task. Thus, to the degree that unbelievable information is successfully recollected, it should have an identical effect in memory-based tasks as when it is provided in an online task.

Hence, due to the nature of how believable and unbelievable information is processed in System 1 and System 2 and how these systems interact with memory, a rather counterintuitive set of predictions can be generated. Namely, my thesis is that unbelievable information should be more memorable than believable information, but interestingly, believable information should have a larger impact in memory-based tasks than in online tasks, whereas unbelievable information should have a similar effect or else a lesser effect in memory-based tasks than online tasks, depending on how successfully unbelievable items are recollected from memory.

The current series of experiments were designed to test this theory regarding how believable and unbelievable information is processed, stored, and retrieved from memory. In Experiment 1, participants engage in an online task in which they are given either believable or unbelievable evidence and asked to make an inference based on the

evidence provided. This experiment serves as a baseline measure of how influenced participants are by different types of evidence, when that evidence is not being retrieved from memory. Experiments 2 and 3 are memory-based tasks where participants are shown the believable and unbelievable evidence before making inferences. Furthermore, in both experiments we test participants' memories to look for evidence that believable and unbelievable evidence is being encoded in a manner consistent with the dual-process account of reasoning that we have outlined.

To anticipate, Experiment 1 demonstrates that, in online tasks, believable evidence statements have a larger impact on inferences than unbelievable evidence statements. Experiment 2 and 3 however, demonstrate that this effect reverses when evidence is being retrieved from memory to make inferences. Furthermore, Experiment 2 and 3 provide some evidence that unbelievable evidence may be more memorable than believable evidence statements. The results of Experiment 3 in particular suggest that unbelievable evidence statements may be stored in a more verbatim manner, whereas believable evidence statements may be stored in a less verbatim manner.

Pre-rating Stimuli

The goal of our study was to examine how believable and unbelievable evidence affect inference judgments and how memorability interacts with this effect. Therefore, we needed a set of believable and unbelievable statements to use as stimuli throughout the series of experiments, as well as a set of statements that could be used to make inferences, based on the evidence statements.

Method

Participants. Forty-seven participants rated the believability of evidence statements and thirty-six rated the accuracy of inference statements. All participants were from the University of Waterloo and for their participation participants received 0.5 course credits towards their Introductory Psychology course.

Materials. To construct evidence statements, we searched the internet for short, interesting facts (e.g., "A blink lasts 0.3 seconds"). In total 63 facts were found to be used as evidence statements. Statements were selected such that they would be unfamiliar to most participants, would be clearly believable or unbelievable, and such that each contained a numerical value. These statements were then categorized as either believable or unbelievable by the researcher. Afterwards, for each believable evidence statement, the numerical value was changed to create an unbelievable evidence statement (e.g., "A blink lasts 2.5 seconds"), and similarly from each unbelievable statement the numerical value was changed to create a believable statement. Thus, in the end we had constructed 63 believable and 63 unbelievable evidence statements, and each believable statement had a corresponding unbelievable version. These pairings of believable and unbelievable statements will be referred to as evidence pairs.

In addition to evidence statements we needed inference statements. That is, we needed statements which participants could use the evidence statements to make inferences about (e.g., "Blinks are large and noticeable motions"). We constructed an inference statement for each evidence pair, which resulted in 63 inference statements. Furthermore, we decided to setup the inference such that the accuracy of the inference statement could either be supported or rejected based on the believability of the evidence sentence presented. Although sometimes the believable statement would support the inference statement, other times it would refute it. That is, the believable and unbelievable statements always differentially supported the inference statement. For example, one of the inference statements was, "Blinks are large and noticeable motions". The inference participants were asked to make was, "based on the evidence we provided you, is this statement accurate?" If participants had received the believable statement (i.e., "A blink lasts 0.3 seconds") then clearly this inference statement is not supported. If however participants received the unbelievable statement (i.e., "A blink lasts 2.5 seconds") then this inference is supported (at least to a greater degree). Thus, for each inference statement there was no right or wrong answer, but merely degrees of support, with some evidence statements favoring the accuracy of an inference question, and others disputing it.

Procedure. Participants completed the study on the internet. After reading a consent form and confirming their identification, they were presented with a set of instructions which told them they would be rating the believability of statements (if they were rating evidence statements), or else the accuracy of statements (if they were rating inference statements). For each evidence statement, participants rated, on a Likert scale

of 1 to 7, how believable that statement was. A rating of 1 indicated the statement was not believable and a rating of 7 indicated it was believable. For inference statements, participants rated how accurate the statements were. A rating of 1 indicated that the statement was inaccurate, whereas a rating of 7 indicated that the statement was believed to be true.

Results & Discussion

The relevant results from these two experiments are the believability ratings of evidence statements and the accuracy ratings of inference statements. No analyses were conducted on the data themselves as this data served merely as norming data to later be used as a baseline. The full results of these experiments can be found in Appendix A. However, as only a subset of stimuli would be focused on in subsequent experiments, we have presented the most relevant stimuli in Table 1.

Table 1. Pre-Rating Data and Results of Experiment 1 for Believable and Unbelievable Evidence Statements and Corresponding Inference Statements.

Believable Evidence Statements	Believability Rating	Inference Statements	Pre-Rating Accuracy	Exp 1 Accuracy	Accuracy Change
18% of a person's income is spent on transportation	4.91	For most people, transportation costs are easily afforded	2.67	3.26	0.60
In 1962, the first Wal-Mart opened up in Rogers, Arkansas	4.70	Walmart is a relatively new company	4.58	3.18	1.40**
The stomach of an adult can hold 1.5 litres of material	5.11	One jug of pop is enough to fill up an adult's stomach	4.14	4.32	0.18
Roses need 6 hours of sunlight per day to grow properly	4.80	Roses can grow even with very little sunlight	3.31	2.57	0.74*
25% of injuries by athletes involve the wrist and hand	4.62	Common injuries for athletes involve hands and wrists	4.67	4.84	0.18
38% of Americans eat breakfast everyday	4.64	No one really eats breakfast every day	2.31	2.32	0.01
women spend 55 minutes per day getting showered and dressed	5.07	Women usually wake up an hour early to get ready in the morning	4.97	4.77	0.20
40% of the states in the U.S. have severe, or extreme pollution problems	4.91	Pollution still isn't a major problem for most of the states in the US	2.67	2.87	0.20
5% of the people who use personal ads for dating are already married	4.53	If you meet a person from a personal ad, chances are they are already married	2.67	2.29	0.38
8% of men are color blind	5.02	The reason most men are bad at colour-coordinating is that they are colour blind	2.28	2.26	0.01
50% of lottery players go back to work after winning the jackpot	4.67	Pretty much no one quits their job when they win the jackpot in a lottery	3.25	2.89	0.36
In the United States, 33% of land is covered by forests	4.64	If you drive across the US most of the drive you will be driving through forests	2.19	2.58	0.38

Unbelievable Evidence Statements	Believability Rating	Inference Statements	Pre-Rating Accuracy	Exp 1 Accuracy	Accuracy Change
10% of all greeting cards are purchased by women	2.44	Dads are more likely to buy greeting cards than moms	2.22	3.58	1.36**
It costs $10 to make a $1 bill in the United States	2.31	It sometimes costs the government more to make a bill than the bill is worth	3.14	5.61	2.47**
A disposable diaper can hold up to 23 pounds of liquid	2.36	Disposable diapers can hold the weight of several children	3.19	2.66	0.54
500 Valentine's Day cards are sent each year in North America	2.47	Very few people actually send Valentine's Day cards out	3.17	4.63	1.46**
A person passes gas every 10 minutes	2.44	Someone who passes gas a few times an hour is not having a normal day	4.17	2.97	1.19**
The average person falls asleep in 77 minutes	2.51	People usually fall asleep pretty quickly when they go to bed at night	3.83	2.84	1.00**
The Snickers chocolate bar was invented in 1996	2.24	Snickers has been around since the Great Depression	3.94	2.21	1.74**
A leech has 32 brains	2.23	You can't cut off a leeches head because it has brains all over its body	2.89	3.95	1.06*
A female mouse can produce up to 200,000 babies a year	2.54	A handful of mice can produce millions of babies per year	4.11	4.66	0.55
Rats can only survive for 20 minutes without any food	2.04	Rats usually fast for several days at a time as food is usually scarce	4.67	2.63	2.04**
7% of weddings are held in a synagogue or church	2.36	People usually get married in religiously sacred buildings	4.75	3.00	1.75**
Majority of brides plan their wedding for 5 years	2.45	Most weddings are planned in under a year and therefore require the help of a wedding planner	2.36	2.46	0.10

* $p < .05$, ** $p < .01$

Using this newly pre-tested data we now turn to our primary investigation. The dual-process account we outlined in the Introduction suggests that believable information is more likely to be primarily processed in System 1, whereas unbelievable information is more likely to receive additional processing in System 2. As a result of this differential processing, unbelievable information should be more memorable than believable information. However, in terms of inferences, believable evidence should have a larger impact on inferences when it is being retrieved from memory than when it is being processed online. Unbelievable information should have a similar effect or a lesser effect when being retrieved from memory than when processed online, depending on the degree to which it is accurately recollected. To begin our investigation we first turn to an examination of how believable and unbelievable information affect inferences in an online task.

Experiment 1:

On the Believability of Evidence

Experiment 1 was designed to assess how much the accuracy ratings of inference statements change when believable or unbelievable evidence is presented to participants. So, if participants are presented with evidence either for or against an inference statement, will they change how accurate they judge that statement to be? And furthermore, which type of evidence will more strongly affect decisions? This experiment was conducted primarily to act as a baseline with which to compare later memory experiments.

Some past research has suggested that the believable information has a larger impact on decisions than does unbelievable information (e.g., Thompson, 1996). Although we hypothesized that, generally, participants may be more likely to utilize believable than unbelievable evidence when making inferences, with regard to this current experiment we did not believe a strong believability effect would emerge. That is, to pre-rate our inference statements participants were asked to judge how accurate they believed each statement to be. Likely, these judgments were based on participants' prior knowledge and beliefs. Therefore, if we were to then present participants with believable evidence (i.e., information they were previously using to rate the accuracy of inference statements), this may not result in a very large change in accuracy rating. Thus, as unbelievable evidence is most likely to be inconsistent with participants' prior beliefs, and because inference statements were likely pre-rated based on prior beliefs, unbelievable evidence statements may have a larger impact in terms of changing participants' rating of inference statements.

Yet, whether or not unbelievable evidence statements are found to have a larger impact on inferences than believable evidence statements is actually inconsequential to our primary investigation. That is, we are not as concerned with the absolute impact of believability on inference statements, as much as we were interested in how memory interacts with the impact of believability on inferences. Thus, the results of Experiment 1 serve merely as a baseline with which to judge later, memory-based experiments.

Method

Participants. Thirty-nine participants from the University of Waterloo participated in the experiment for 0.5 bonus credits towards their Introductory Psychology course.

Materials. The evidence statements and inference statements used were from the pre-rated set described earlier (see Table 1). We selected 12 believable and 12 unbelievable evidence statements, where believability was based on participants' pre-ratings, not based on prior experimental labels. As described previously, participants rated the believability of evidence statements on a 7-point Likert scale with 7 indicating high believability and 1 indicating low believability. The 12 believable evidence statements selected had a mean believability rating of 4.80 ($SD = 0.20$) and the 12 unbelievable evidence statements had a mean rating of 2.37 ($SD = 0.14$), $t(22) = 34.48$, $p < .01$. None of these statements were from the same evidence pair. This resulted in 24 unrelated evidence statements where each was either believable or unbelievable.

For each evidence statement we also selected the corresponding inference statement to accompany that evidence statement in the experiment. Inference statements for believable evidence items had a mean accuracy rating of 3.31 ($SD = 1.02$) and for

unbelievable evidence items had a mean accuracy rating of 3.54 (SD = 0.84), t(22) = 0.60, p = 0.55. Thus, inference statements were, on average, relatively neutral in terms of believability, or at least equivalent across evidence types. An additional benefit of selecting these statements to have pre-ratings close to the center of the Likert scale was to allow room for the ratings to change when evidence was presented.

Procedure. Participants completed the experiment on the internet. After reading a consent form, participants were instructed that they would be engaged in an inference task. On each trial, participants would see an evidence statement and an inference statement. Their task was to read both statements and judge how accurate the inference statement was based on the assumption that the evidence statement was true. Instructions read as follows:

> In the following experiment you will be asked to rate the believability of a series of statements. Before each statement we will first tell you a fact (e.g., "A dime has 118 ridges around the edge"). Please treat this fact as true, regardless of whether or not you believe it. Following the fact we will show you a statement (e.g., "It's pretty hard to count all the ridges on a dime"). Your task is to tell us, based only on the fact we provided, how believable you think that statement is. So for example, if you think that counting all the ridges on a dime, assuming it really does have 118 ridges would be hard, you would indicate the statement is believable.

> To indicate how accurate you believe the statement to be, use the scale below the statement. The scale will range from 1-7. A 1 indicates that you do not think the statement is accurate, and hence, do not believe it. A 7 indicates you do think the statement is accurate, and hence, you do believe it. Intermediate values indicate an intermediate level of certainty about believability; for example, a 4 indicates that you think the statement may or may not be accurate. That is, the statement seems somewhat believable, but also somewhat unbelievable. Try your best to be as accurate as you can when evaluating the statements. If you see a statement which you are unsure about, go with your best intuition about how believable it seems.

Thus, participants were instructed to treat the evidence statement as true, regardless of whether they actually believed it or not, and judge the inference statement's accuracy based on this information. Participants rated the accuracy of the inference statement on a 7-point Likert scale, identically to the pre-rating phase.

Results & Discussion

A preliminary analysis at the item level can be seen in Table 1. For each inference statement, the degree of judged accuracy from pre-rating values was compared to the degree of judged accuracy from Experiment 1, and the amount of change in judged accuracy was noted. For inference statements presented with believable evidence, only 2 of the 12 items had significantly changed compared to 9 of the 12 inference statements that had been shown with unbelievable evidence statements. Thus, this preliminary item

analysis revealed that, consistent with our predictions, believable evidence had a lesser effect than unbelievable evidence on ratings of inference statements.

However, the more relevant analysis is one conducted at the participant level. Specifically, for each inference judgment for each participant we calculated the absolute difference between the judged accuracy of the inference statement and the average pre-rating judgment for that particular inference statement. The result is, for each participant we had a measure of how much that participants' judgments of inference statements differed from the average pre-rating values. This new variable, which we term absolute inference change (as we disregarded the sign for this variable), provided a more accurate measure of how much believable and unbelievable evidence statements affected inference judgments, at a participant level. For example, if a participant rated a particular inference statement as 4 in terms of accuracy, but the pre-rating average rating for that statement was 2.4, an absolute difference of 1.6 would be calculated for that participant for that trial, indicating that seeing evidence caused an accuracy judgment that was 1.6 units deviant from the normed value. It is these absolute inference changes which we now turn to.

One sample t-tests, tested against zero, were used to test the absolute inference change for inferences read with believable and unbelievable evidence statements. These tests revealed that inferences made with both believable ($\underline{M}$ = 1.49, $\underline{SD}$ = 0.39) and unbelievable ($\underline{M}$ = 2.00, $\underline{SD}$ = 0.52) evidence statements caused a significant change in accuracy judgments relative to the normative data, $\underline{t}(38) = 24.11$, $\underline{p} < .01$,and $\underline{t}(38) = 24.18$, $\underline{p} < .01$, respectively. Thus, at the participant level, both believable and unbelievable evidence statements affected inferences. However, on average,

unbelievable evidence caused a larger change in accuracy judgments than did believable evidence, $t(38) = 4.94$, $p < .01$. Thus, it appears that although both believable and unbelievable evidence statements changed the inferences participants made, participants were nonetheless most affected by unbelievable evidence.

Earlier we suggested that this fact may have been a result of the subjective nature of our task. That is, responses to inferences had been pre-rated by participants who were likely using prior knowledge when judging the accuracy of inference statements. Thus, as believable evidence would merely support prior knowledge, providing participants with believable evidence may have done little to change these accuracy ratings. As such, it is not surprising to find that believable evidence had a smaller impact on inferences than did unbelievable evidence.

However, another interpretation is that unbelievable evidence statements in this experiment were more unbelievable than believable statements were believable. That is, unbelievable evidence statements had a mean rating of 2.37. This means that unbelievable statements, on average, were 1.37 units above the lowest point on the Likert rating scale (i.e., 1). Believable evidence statements on the other hand had a mean rating of 4.80, indicating that, on average, they were 2.20 units below the highest point on the Likert rating scale (i.e., 7). If we assume that the degree of believability or unbelievability of evidence directly predicts the change in inferences observed, then it is not surprising to find that unbelievable evidence statements had a larger impact on inferences than did believable evidence statements.

Yet, as discussed before, the exact reason why unbelievable evidence statements had a larger impact on inferences than did believable evidence statements is actually

inconsequential to our primary investigation. As we were not as concerned with the absolute impact of believability on inference statements but rather, the relative change in impact between an online and a memory-based task, the results of Experiment 1 serve merely as a baseline with which to judge later, memory-based experiments.

One final point of interest is that, although we have argued that believable evidence has a smaller impact on inferences than does unbelievable evidence in this online task, remarkably believable evidence still had a substantial effect. That is, given one piece of believable evidence, inferences changed on average 1.49 points, versus unbelievable evidence changing inferences 2.00 points. Although we have suggested that because pre-rating values were based on prior beliefs that believable evidence should have a smaller effect than unbelievable evidence, and this was borne out, clearly believable evidence still has a relatively large impact on inferences. That is, its impact is closer to that of unbelievable evidence (i.e., 2.00) than it is to no impact at all (i.e., zero). Hence, although we have claimed believable evidence has a lesser impact than unbelievable evidence in an online inference task, it should be made clear that believable evidence still has a significant effect, that is in some sense, comparable to that found for unbelievable evidence.

To summarize then, both an item-level and participant-level analysis revealed that, as was suggested may be the case, unbelievable evidence statements affected participants' inference judgments to a greater degree than did believable evidence statements. This result may be due to the fact that there are no objective correct answers in our experiments, and therefore judgments in this experiment are compared to pre-rating data, which itself is likely based on prior believable evidence participants already

possess. Alternatively it may be due to the fact that unbelievable evidence statements were more unbelievable than believable evidence statements were believable. Regardless, these results could now be used as a baseline for how believability affects inferences in our task when inferences are made online. We now turn to our investigation of how the memorability of believable and unbelievable evidence may alter this finding.

Experiment 2:

The Interaction Between Evidence Believability and Memorability

Experiment 1 demonstrates that in the online version of our inferences task, unbelievable evidence has a larger impact on inferences than does believable evidence. Based on our earlier discussion of the dual-process account of reasoning however, we may expect to see different results in Experiment 2, where a memory-based inferences task is used instead. That is, the dual-process account of reasoning we outlined in the Introduction, suggest that believable information is more likely to be primarily processed by System 1, whereas unbelievable information is more likely to receive additional processing by System 2. The result of these processing differences is that believable evidence should be encoded differently than unbelievable evidence. Whereas believable evidence may be encoded more automatically and in a manner which integrates it with existing knowledge, unbelievable information should be less integrated with existing knowledge (because it is inconsistent with that knowledge) and encoded in a more conscious and abstract manner.

As we considered the issue earlier we suggested that in a recognition memory test, a mirror effect may be observed between believable and unbelievable evidence. Namely, believable evidence may be more familiar than unbelievable evidence but unbelievable evidence may be more consciously recollectable than believable evidence. This should result in a mirror effect and specifically, superior memory for unbelievable evidence. In terms of inferences however, we predicted that an interesting interaction between memorability and believability may occur. That is, we suggest that because believable evidence may be primarily processed in System 1, this information may become

integrated with prior knowledge. Attempts to retrieve specific items to base inferences upon may inadvertently result in several pieces of evidence, all consistent with the one integrated, being accidentally retrieved. Regardless of whether or not participants are consciously aware of this fact, the fact that more evidence is now present on which to base inferences should result in more extreme changes in inferences compared to an online task. Thus, believable evidence statements should have a larger impact on inferences in Experiment 2 than in Experiment 1.

For unbelievable evidence statements, we hypothesized that these items would not integrate well with prior knowledge during System 1's initial processing. Additionally, subsequent processing in System 2 would also not integrate these items with prior knowledge, although it would make them more consciously recollectable. Thus, when attempting to retrieve an unbelievable evidence statement in order to make an inference, participants may be relatively successful in recollecting the exact item from memory. Furthermore, no additional evidence would likely be retrieved, since the unbelievable evidence was not integrated with prior knowledge, retrieval attempts should not lead to other evidence being accidentally retrieved. The result is that, to the degree unbelievable evidence statements can be consciously recollected, they should affect inferences more or less to the same degree they would if they were simply provided to participants, as in Experiment 1. Thus, unbelievable evidence statements should affect inferences to a similar degree as in Experiment 1, or if they are not perfectly recollectable, to a somewhat lesser degree.

Method

Participants. Twenty-six participants from the University of Waterloo participated in the experiment for 0.5 bonus credits towards their Introductory Psychology course.

Materials. The 24 evidence statements from Experiment 1 were used as the study items in this experiment. Twenty-four new evidence statements were also selected for this experiment to act as foils during the recognition memory test (see Table 2). These items were all unrelated to the study items; hence, 12 new believable and 12 new unbelievable statements were obtained. From the pre-rating data, the 12 new believable statements had a mean believability rating of 4.89 (SD = 0.56) and the unbelievable statements had a rating of 2.12 (SD = 0.33). Finally, the 24 inference statements from Experiment 1 were used as the inference statements in this experiment.

Table 2. Believability ratings of believable and unbelievable foils to be used during the memory test in Experiment 2.

Believable Evidence Statements	Believability Rating	Unbelievable Evidence Statements	Believability Rating
Chopsticks originated from China 4,000 years ago	5.26	A blink lasts 2.5 seconds	1.93
McDonald's restaurant has over 1.5 million employees all over the world	6.13	75% of the population is left-handed	1.96
The United States Postal Service handles 40% of the world's mail volume	4.46	It takes 2 weeks for food to be broken down in the human stomach	1.91
Each day 14 people die from asthma in North America	4.48	Men live 15 years longer than women do	1.85
People spend 33% of their life sleeping	4.87	700,000 people have been frozen after their death	2.60
American models are skinnier than 98% of American women	5.63	Heinz first started making ketchup in 1233 AD	1.82
An average American eats 60 hot dogs per year	4.49	In a year, an American kid eats 3 slices of pizza	1.64
90% of Pumpkins sold are for decoration	5.07	8% of candles that are purchased are purchased by women	2.13
Alaska has 2 times as many caribou as people	4.42	17% of children go out trick or treating for Halloween	2.50
31% of employees skip lunch entirely	4.93	70% of the human population reside in deserts	2.47
33% of accidental deaths occur in the home	4.43	Hitler was voted Time Magazine's man of the year in 1981	2.59
Only 4% of babies are born on their actual due date	4.46	The world's tallest roller coaster reaches a peak height of 6 meters	2.02

Procedure. The experiment was conducted on the internet. After reading a consent form, participants were informed they would participate in a 3-part study. In the first phase of the experiment, participants engaged in a study phase. During the study phase the 24 evidence statements from Experiment 1 were shown, one at a time. Participants read each sentence then clicked a button to continue. They were instructed to remember each statement as best as possible because later we would be testing their memory. After the study phase participants engaged in a memory test.[2] In the memory test, the 24 items from study were intermixed with 24 new, unrelated evidence statements. For each statement participants had to indicate whether the item was old (present at study) or new. Participants used a radio button on the website to indicate their decision on each trial and clicked a button to continue. During the final phase of the experiment, participants saw the 24 inference statements from Experiment 1 and were instructed to judge the accuracy of these inference statements based on the evidence read at study. As in Experiment 1, participants were instructed to treat the statements from the study phase as true when making inferences, regardless of what they really thought of them.

Results & Discussion

The results were analyzed in two parts. First we analyzed the memory data to examine what effects, if any, were present. Then we examined the inference judgments to see how these judgments differed from Experiment 1, under the assumption that any

[2] We opted to use a fixed order for the three phases of this experiment. This was because a set of early pilot data revealed that if participants engaged in the inference task before the memory task, significantly different results in the memory task were revealed. For the inference task however, it did not matter which order it occurred. That is, if it occurred before or after the memory test, the results were nearly identical, therefore, to preserve the integrity of the memory test we always had it occur before the inference phase.

differences noted were a result of participants having to remember evidence, rather than just have it provided to them.

Memory Test. The recognition memory data is presented in Table 3. Recall that the mirror effect is the finding of a greater hit rate but lower false alarm rate to one stimulus class over another. We also hypothesized that unbelievable evidence statements would be more recollectable than believable evidence statements, resulting in a greater hit rate. Additionally, believable evidence statements in general would be processed more fluently in System 1 than unbelievable evidence statements, resulting in more false alarms. Thus, a mirror effect was predicted to be found such that unbelievable evidence statements were more memorable than believable evidence statements, and specifically had more hits and fewer false alarms than believable evidence statements.

From hits and false alarms we calculated d' for both believable and unbelievable statements. The statistic d' is often considered a more comprehensive measure of memory than either hits or false alarms alone and is better able to represent memorability of items without being susceptible to issues of bias. Thus, d' was taken as our general measure of memorability, whereas hits and false alarms were examined to look for the specific mirror effect pattern we had predicted.

The data were analyzed in a 2 (old vs. new) X 2 (believable vs. unbelievable evidence statement) within-subjects ANOVA. Participants could successfully discriminate old from new items, $F(1, 25) = 762.85$, $MSe = 0.03$, $p < .01$, $\eta^2 = .97$. Additionally, although there was no difference in response to believable and unbelievable evidence statements generally, $F < 1$, there was a borderline significant interaction, $F(1, 25) = 3.39$, $MSe = 0.002$, $p = .08$, $\eta^2 = .12$. This interaction represented the fact that

unbelievable evidence statements were nearly more memorable (i.e., higher d') than believable evidence statements.

Paired-sample t-tests revealed that, although the trends in the means trended in a direction consistent with a mirror effect (i.e., more hits and fewer false alarms to unbelievable than believable evidence), neither the hit rate difference, $t(25) = 0.30$, $p = .77$, nor the false alarm rate difference was significant, $t(25) = 1.47$, $p = .15$. A main reason for this failure to find significance seemed to be that the recognition memory test was so easy that the data was contaminated by floor and ceiling effects. Indeed, 11 of the 26 participants had hits for both believable and unbelievable evidence statements of 1.00. Additionally, 23 of 26 participants had false alarms for both believable and unbelievable evidence statements of 0.

Table 3. Means (and Standard Deviations) for hits, false alarms, and d' noted for believable and unbelievable evidence statements in Experiment 2.

	Evidence Type	
	Believable	Unbelievable
Hits	.93 (.11)	.94 (.12)
False Alarms	.04 (.13)	.01 (.03)
d'	.95 (.10)	.97 (.08)

From the perspective that we had actually anticipated unbelievable statements to be more memorable than believable statements, it would make sense to conduct one-tailed tests when examining the memorability data. If we adopt this approach, then unbelievable evidence statements are more memorable than believable evidence statements (i.e., $p < .05$). Although perhaps a somewhat liberal approach, given that the hit and false alarm data appears to be hampered by ceiling and floor effects, it is still

remarkable that any noticeable pattern emerged at all. Furthermore, given our theoretical consideration of the memorability of believable and unbelievable information, we had strong a priori predictions that unbelievable statements would be more memorable than believable statements. Thus, we feel it is acceptable in this case to report that unbelievable statements were more memorable than believable statements, although clearly the difference in memorability is not large.

Inference Phase. As in Experiment 1, we were concerned with the relative difference between accuracy judgments of inference statements in this experiment as compared to the normative data. Thus, we examined the absolute inference change of inferences made based on believable and unbelievable evidence statements. As mentioned previously, absolute inference change provides a measure at the participant level as to how much change in inference judgments was noted in this experiment as compared to the pre-rating data, where no evidence was provided.

As was found in Experiment 1, a comparison of absolute inference changes against zero revealed that in this experiment both believable (M = 1.99, SD = 0.43) and unbelievable evidence (M = 1.69, SD = 0.38) significantly altered accuracy judgments compared to the normative data, $t(25) = 23.72$, $p < .01$, and $t(25) = 22.84$, $p < .01$, respectively. However, unlike Experiment 1, believable evidence statements impacted inferences more than did unbelievable evidence, $t(25) = 3.40$, $p < .01$.

A 2 (believable vs. unbelievable evidence) X 2 (Experiment 1 vs. Experiment 2) mixed ANOVA was conducted to directly compare the results of Experiment 1 and 2. No main effect for the impact of believability of evidence was found, $F(1, 63) = 1.35$, $MSe = 0.16$, $p = .25$, $\eta^2 = .02$, and no main effect for experiment was found, $F(1, 63) =$

34.22, MSe = 0.16, p = .23, η^2 = .02. However, there was a significant interaction, F(1, 63) = 34.22, MSe = 0.16, p < .01, η^2 = .36. This interaction indicated that the impact of the believability of evidence varied depending on the experiment in question. To further investigate this interaction, independent-sample t-tests were used to compare the impact of the believability of evidence between Experiment 1 and 2.

As predicted, believable evidence statements had a larger impact in this memory-based experiment, than in an online task, t(62) = 5.18, p < . 01, d = 1.32. Additionally, unbelievable evidence statements had a lesser impact on inferences in this memory-based experiment, than in the online task in Experiment 1, t(62) = 2.64, p < .01, d = 0.67. Thus, as suggested would be the case, believable evidence statements have a larger impact on inferences when retrieved from memory than when explicitly provided. Additionally, this specific experiment found that unbelievable evidence statements may have a lesser impact on inferences when retrieved from memory then when provided. According to our dual-process account of reasoning, this suggests that in Experiment 2, unbelievable evidence statements may not have been perfectly recollectable, as to the degree they are clearly recollectable we predicted they should have a similar effect on inferences as in Experiment 1.

Generally then, the findings from Experiment 2 are in line with the dual-process account of reasoning we have laid out. Although we did not find strong evidence that this difference in memorability exists, ceiling and floor effects in the hits and false alarms respectively likely contributed to this issue. Thus, we now turn to Experiment 3 with three goals in mind. First, to again demonstrate that when believable and unbelievable evidence is being retrieved from memory, believable evidence has a larger impact than it

does when it is provided (i.e., as in Experiment 1). Second, we wished to make the memory test more difficult, with the hopes of observing a clear memory benefit for unbelievable evidence. Thirdly, we wished to add another aspect to the memory test in order to investigate our hypothesis that unbelievable evidence statements are stored in a more detail-specific manner than are believable statements. Namely, by introducing the alternative member of evidence pairs of old items to act as superficially related foils at test. By replicating those results found in our inferences phase and obtaining clearer results in our memory test, we hope to provide stronger support for the specific hypotheses we have put forth about believability and memorability.

Experiment 3:

Investigating the Memory Hypothesis

Experiment 3 was an extension of Experiment 2, in an attempt to better ascertain if there are any memory differences between believable and unbelievable statements, and if these differences are consistent with our theory. The dual-process account of reasoning we put forth previously proposed that unbelievable evidence should be better recognized than believable evidence, and that this may manifest specifically as a mirror effect. Although Experiment 2 did not find strong evidence that unbelievable statements are better recognized than believable statements, this was likely hindered by floor and ceiling effects. Thus, in Experiment 3 we sought to first increase the difficultly of the memory test by introducing relatedness among the foils. That is, instead of simply using the 24 unrelated foils from Experiment 2, we used the 24 foils from Experiment 2 and the foils' corresponding statements from the evidence pairs (see Table 4). For example, "A blink lasts 2.5 seconds" was used as a foil in Experiment 2, therefore both that and "A blink lasts 0.3 seconds" were used in Experiment 3.

Table 4. Superficially related foils to previously studied evidence statements (see Table 1) and previously used foils (see Table 2).

Foils Related to Studied Believable Statements	Believability Rating	Foils Related to Studied Unbelievable Statements	Believability Rating
49% of a person's income is spent on transportation	3.41	93% of all greeting cards are purchased by women	5.11
In 1991, the first Wal-Mart opened up in Rogers, Arkansas	3.23	It costs 3 cents to make a $1 bill in the United States	4.47
The stomach of an adult can hold 20 litres of material	2.93	A disposable diaper can hold up to 7 pounds of liquid	4.50
Roses need 20 minutes of sunlight per day to grow properly	4.18	1 billion Valentine's Day cards are sent each year in North America	5.39
4% of injuries by athletes involve the wrist and hand	3.42	A person passes gas every 2 hours	4.76
6% of Americans eat breakfast everyday	3.07	The average person falls asleep in 12 minutes	4.36
women spend 3 hours per day getting showered and dressed	3.38	The Snickers chocolate bar was invented in 1930	4.91
90% of the states in the U.S. have severe, or extreme pollution problems	4.11	A leech has 1 brain	5.13
62% of the people who use personal ads for dating are already married	3.07	A female mouse can produce up to 100 babies a year	4.67
71% of men are color blind	2.80	Rats can survive up to 14 days without any food	4.93
92% of lottery players go back to work after winning the jackpot	4.28	85% of weddings are held in a synagogue or church	5.02
In the United States, 87% of land is covered by forests	2.85	Majority of brides plan their wedding for 9 months	5.43

Foils Related to Prior Believable Foils	Believability Rating	Foils Related to Prior Unbelievable Foils	Believability Rating
Chopsticks originated from China 100 years ago	2.91	A blink lasts 0.3 seconds	5.76
McDonald's restaurant has over 2,300 employees all over the world	3.56	15% of the population is left-handed	5.15
The United States Postal Service handles 99% of the world's mail volume	2.89	It takes 3 hours for food to be broken down in the human stomach	4.78
Each day 1.2 million people die from asthma in North America	2.62	Women live 7 years longer than men do	5.67
People spend 62% of their life sleeping	3.65	90 people have been frozen after their death	4.07
American models are skinnier than 12% of American women	2.91	Heinz first started making ketchup in 1876	5.20
An average American eats 7 hot dogs per year	2.87	In a year, an American kid eats 46 slices of pizza	4.76
15% of Pumpkins sold are for decoration	3.80	96% of candles that are purchased are purchased by women	5.15
Alaska has 100 times as many people as caribou	3.16	93% of children go out trick or treating for Halloween	4.84
85% of employees skip lunch entirely	3.29	30% of the human population reside in deserts	3.86
99% of accidental deaths occur in the home	2.84	Hitler was voted Time Magazine's man of the year in 1938	3.95
79% of babies are born on their actual due date	3.80	The world's tallest roller coaster reaches a peak height of 72 meters	4.73

By using related foils we hoped to increase the difficulty of the memory test. That is, past research has shown that increased perceptual fluency (e.g., Jacoby & Dallas, 1981) can increase the subjective feelings of familiarity to those items. Because foil pairs were virtually identical to one another, save for one number being altered, we believed that these items could serve to increase each other's perceptual fluency, or provide the false sense of recent exposure. That is, after seeing "A blink lasts 2.5 seconds", if participants later see "A blink lasts 0.3 seconds", this second foil should actually be processed more fluently than the first foil which should lead to an increased sense of familiarity. This false sense of familiarity should cause participants to be more likely to believe they saw this second foil on the study list, when in fact it they did not. Thus, each foil had a related foil mixed in at test. In this manner, participants should have a harder time discriminating old from new items, as both old and new items would feel familiar to participants.

However, a more relevant goal was to specifically test how believable and unbelievable information may be encoded. That is, we have supposed the unbelievable evidence statements are stored in a more detail-specific manner due to System 2 processing whereas believable statements are more likely to be integrated with prior knowledge due to System 1 processing. Although this account predicts that there should be a memory advantage for unbelievable than believable items, simpler accounts could also predict this effect. For example, if we simply assume that unbelievable evidence statements are more surprising or require more processing time, we would likely predict they should also show superior memory than believable evidence statements. This account could potentially explain the memorability effects without having to rely on a

System 1 and System 2 processing distinction. Thus, a stronger test of our dual-process account of reasoning would be to demonstrate that unbelievable evidence statements are not only more memorable than believable evidence statements, but are more memorable for the specific reasons our account suggests.

One way to examine the details of memory is to investigate what types of items participants false alarm to. That is, our account suggests that believable evidence is more likely to be encoded in such a way that details are lost because it has been integrated with prior knowledge (due to System 1 processing), whereas unbelievable evidence is less likely to be encoded in this manner (due to System 2 processing). If this is so then participants should false alarm more for superficially related but contradictory foils for unbelievable statements, as these foils would be more easily matched to unbelievable items. That is, if an unbelievable statement is stored in terms of details, and the exact details can be retrieved, then sentences that share many of those details should cause a false alarm, even if the sentences differ on some conceptual aspects. So if the unbelievable statement, "The average person falls asleep in 77 minutes" is memorized at study, and at test the superficially related foil "The average person falls asleep in 12 minutes" is shown, participants should be likely to false alarm to this foil. This is because the foil shares many surface features with the unbelievable statement, and we hypothesize that these are the details that are readily stored for unbelievable statements.

For believable statements however, false alarms to superficially related foils should be significantly lower. That is, believable evidence statements are integrated with prior knowledge, fewer specific details of the items should be encoded. As a result, the fact that the foils match the studied items on surface details should not have a large effect

on participants' judgments. Further, the foils would obviously differ from the encoded believable statements, as they would be incompatible with prior knowledge. Thus, participants should readily reject these foils.

For example, if the believable statement "Roses need 6 hours of sunlight per day to grow properly" is read at study, we hypothesize that this information will become integrated with prior knowledge and surface characteristics may be lost. As a result, participants may simply remember something like "roses need a day of sun to grow". If the superficially related foil "Roses need 20 minutes of sunlight per day to grow properly" is shown at test, participants should be less likely to false alarm to it. This is because, although this item shares lots of surface characteristics with the studied item, those surface characteristics were not encoded or stored well. Furthermore, as the believable statement from study was readily integrated with prior knowledge, it becomes obvious that the foil is inconsistent with prior knowledge, and so could not have been the item that was seen at study. Thus, participants may realize this and not false alarm to this foil.

To sum up, we will use the 24 items from the evidence pairs that correspond to the studied items as superficially related but contradictory foils during the memory test. We hypothesize that if unbelievable and believable evidence statements are being encoded and stored as we suggest, then we should see a higher false alarm rate to foils that are superficially related to unbelievable studied statements than believable studied statements. In addition, examining hits and false alarms for unrelated statements should reveal a memory advantage for unbelievable over believable evidence statements. Finally, believable evidence statements should impact inferences to a greater degree in

this experiment than in Experiment 1. Unbelievable evidence statements should impact inferences to a similar degree as in Experiment 1 or else to a lesser degree, depending on how consistently unbelievable items can be recollected from memory.

Method

Participants.

Thirty-nine participants from the University of Waterloo participated in the experiment for 0.5 bonus credits towards their Introductory Psychology course. Two participants exhibited unusual memory test results, having false alarm rates equal to or greater than hit rates. This indicated that these participants either were not completing the task correctly, or had not understood the instructions. As a result these participants were dropped from all subsequent analyses. Therefore, data from 37 participants was examined in Experiment 3.

Materials.

The stimuli from Experiment 2 were used in this experiment. Additionally, for both the study items and foils from Experiment 2 we also obtained the related items from each evidence pair (see Table 4). Thus, at test there were 24 old items, 24 foils, 24 foils which were superficially related to the first set of foils, and 24 foils that were superficially related to the old items.

Procedure.

The experiment was conducted on the internet. It was conducted identically to Experiment 2. The one difference was that in Experiment 3 we used 72 foils during the memory test, not 24 (See Table 4 for a complete list of all foils).

Results & Discussion

Similar to Experiment 2, the results of Experiment 3 were analyzed in two parts. First we analyzed the memory data, then we examined the inference judgments.

Memory Test. The memory data are presented in Table 5. False alarms were calculated as the proportion of "old" responses to any of the foils. The superficially related foils were treated separately from the typical false alarms (i.e., false alarms to the foils which were unrelated to the old items). Before turning to those data however, we first examine hits and the false alarms to foils that were unrelated to the old items, which we will term "typical" false alarms.

Table 5. Means (and Standard Deviations) for hits, false alarms, d', and false alarms to superficially related foils noted for believable and unbelievable evidence statements in Experiment 3.

	Evidence Type	
	Believable	Unbelievable
Hits	.86 (.12)	.90 (.13)
False Alarms	.03 (.04)	.02 (.03)
d'	.91 (.07)	.94 (.07)
Related Foils	.03 (.04)	.09 (.11)

For the hits and typical false alarms we had identical predictions in this experiment as in Experiment 2. Namely, unbelievable evidence statements should be more memorable than believable evidence statements, and specifically that a mirror effect should be present with more hits and fewer false alarms to unbelievable than believable evidence statements. As before, d' was taken as our general measure of memorability while hits and false alarms were also examined to look for the specific mirror effect we had predicted.

The recognition memory data were analyzed in a 2 (old vs. new) X 2 (believable vs. unbelievable evidence statement) within-subjects ANOVA. Participants could

discriminate old from new items, $F(1, 36) = 1894.33$, $MSe = 0.01$, $p < .01$, $\eta^2 = .98$, and although there was no overall main effect of statement believability, $F(1, 36) = 2.44$, $MSe = 0.004$, $p = .13$, $\eta^2 = .06$, there was an interaction between old/new status and believability, $F(1, 36) = 6.99$, $MSe = 0.004$, $p < .05$, $\eta^2 = .16$, which indicated that unbelievable evidence statements were more memorable (i.e., higher d') than believable evidence statements.

Paired sample t-tests revealed that this interaction indicated a mirror effect (i.e., more hits and fewer false alarms to one unbelievable than believable evidence statements) may be present in the data. That is, hit rates for unbelievable evidence statements were higher than for believable evidence statements, $t(36) = 2.25$, $p < .05$, and false alarm rates were marginally lower for unbelievable than for believable evidence statements, $t(36) = 1.82$, $p = .08$. Thus, although unbelievable evidence statements were more memorable than believable evidence statements, and some evidence for a mirror effect was found (i.e., more hits to unbelievable than to believable evidence statements), the false alarm rate difference was not significant, although in the predicted direction.

Although the false alarm rate different between believable and unbelievable evidence was not technically significant, it is worthwhile to note that in both Experiment 2 and 3 the false alarm rate difference was in the direction we had predicted. Furthermore, if we take our mirror effect predictions as strong a priori predictions, we could argue that a one-tailed test would be more appropriate in these cases, resulting in the false alarm rate difference being significant at $p < .05$. However, as superficially related foils had been introduced specifically to examine whether believable and unbelievable evidence statements were being memorized in a manner consistent with our

dual-process account of reasoning, an examination of those foils is actually more critical with respect to our theory than is the failure to find a significant false alarm rate difference.

In terms of superficially related foils, recall that we predicted that false alarms for superficially related foils for studied unbelievable statements should be significantly higher than for studied believable statements. This effect was found as predicted, $\underline{t}(36) = 3.38$, $\underline{p} < .01$. This finding confirms our earlier suggestions, that believable evidence statements may be more integrated with prior knowledge, with fewer surface characteristics having been encoded. Unbelievable evidence statements on the other hand, may be stored in a less integrated, more detail-specific manner than believable statements. Thus, believable evidence statements may have been encoded in a manner consistent with more primary processing having occurred in System 1, whereas unbelievable evidence statements may have been encoded in a manner consistent with more subsequent processing having occurred in System 2.

Given that the results of the memory test confirm our earlier predictions about how evidence statements are memorized, we now turn to an examination of the inference data to investigate if similar effects to those seen in Experiment 2 are evident.

Inference Phase. As in both Experiment 1 and 2, we were interested in examining the relative difference between accuracy judgments of inference statements in this experiment as compared to the normative data. Thus, we examined the absolute inference change of inferences made based on believable and unbelievable evidence statements.

As in previous experiments, single-sample t-tests used to compare absolute inference changes against zero revealed that both believable (M = 2.10, SD = 0.51) and unbelievable evidence (M = 2.20, SD = 0.49) significantly altered accuracy judgments, $t(36) = 24.91$, $p < .01$, and $t(36) = 28.90$, $p < .01$, respectively. Thus, believable and unbelievable evidence had an effect on inferences made. However, of more interest was how the inferences made in this memory experiment compared to those of an online task, such as Experiment 1.

A 2 (believable vs. unbelievable evidence) X 2 (Experiment 1 vs. Experiment 3) mixed ANOVA was conducted to directly compare the results of Experiment 1 and 3. Believable evidence generally had a larger impact than unbelievable evidence, $F(1, 74) = 16.95$, $MSe = 0.17$, $p < .01$, $\eta^2 = .19$, and larger changes in inferences were noted in Experiment 3 versus Experiment 1, $F(1, 74) = 15.47$, $MSe = 0.26$, $p < .01$, $\eta^2 = .17$. However, both of these main effects seemed driven by a significant interaction, $F(1, 74) = 11.49$, $MSe = 0.17$, $p < .01$, $\eta^2 = .13$. Specifically, believable evidence statements had a larger impact in this experiment than in Experiment 1, $t(74) = 5.40$, $p < .01$, $d = 1.26$, however unbelievable evidence statements had a similar impact in this experiment as in Experiment 1, $t(74) = 0.88$, $p = .38$, $d = 0.20$. Additionally, in Experiment 2, believable evidence statements were found to have a larger impact than unbelievable statements on inferences, however, this effect was not evident in this experiment, $t(36) = 0.57$, $p = .58$. Thus, unlike Experiment 2 where unbelievable evidence statements had a lesser impact on inferences than in Experiment 1, in Experiment 3 their impact was equivalent to that of Experiment 1, an issue which we will return to shortly in the General Discussion.

According to our dual-process theory of reasoning, believable evidence statements should affect inferences to a much greater degree when being retrieved from memory than when provided. This is because System 1 processing has integrated believable evidence with prior knowledge, and this prior knowledge could be inadvertently retrieved to support believable inferences, when participants attempt to retrieve a believable evidence statement. In line with this suggestion, compared to Experiment 1, believable evidence statements were found to affect inferences more in this experiment. Thus, compared to when believable and unbelievable evidence is given to participants, when retrieved from memory believable evidence has a larger effect on inferences, whereas, in this experiment, unbelievable information has a similar effect as when given.

General Discussion

As researchers continue to investigate how individuals understand and use information that is either believable or unbelievable, the interaction between this factor and memory will become increasingly important. That is, in everyday life, individuals often make judgments based on knowledge they posses, which sometimes can be either believable or unbelievable. Very rarely are individuals being provided with all of the relevant information when making decisions, and so, they must rely on memory.

Our experiments demonstrate that participants make stronger inferences based on believable evidence when it was being retrieved from memory (Experiments 2 and 3) than when it was simply provided (Experiment 1). For unbelievable evidence however, participants appeared to make similar inferences regardless of whether the unbelievable evidence was provided (Experiment 1) or retrieved from memory (Experiment 3),

although there was some evidence participants may make less extreme inferences when unbelievable evidence is retrieved from memory (Experiment 2).

When initially considering the impact of unbelievable evidence statements retrieved from memory on inferences, we hypothesized that, because unbelievable evidence statements were more likely to receive additional processing by System 2, these items should be encoded in such a way that conscious recollection is later supported. As these items had also not been integrated with prior knowledge, when participants attempt to retrieve a specific unbelievable evidence statement, they should be fairly good at recollecting the exact item they had studied. Thus, to the degree that unbelievable evidence statements are successfully recollected, they should impact inferences in a similar manner as they did in an online task (i.e., Experiment 1). However, in Experiment 2 unbelievable evidence statements had a lesser impact on inferences compared to Experiment 1, whereas in Experiment 3 they had a similar impact on inferences compared to Experiment 1.

One explanation for this result may have been that the introduction of superficially related foils to the memory test inadvertently improved the recollectability of unbelievable evidence statements. That is, we hypothesized that superficially related foils for unbelievable evidence statements, by being so similar to the studied evidence statements, should be likely to cause participants to false alarm to these items. However, a side effect of introducing these superficially related foils may have been that when a superficially related foil was seen, participants recollected the studied item which it corresponded to. Thus, during the memory test, participants may have recollected each

unbelievable evidence statement twice: once when the studied item was presented, and once when the superficially related foil was presented.

Past research has demonstrated that if participants are given practice in retrieving information, memory for that information actually improves, a phenomenon call the testing effect (e.g., Chan & McDermott, 2007; Chan, McDermott, & Roediger, 2006). That is, although the initial intuition may be that, for the best final memory performance, participants should be given extra study time, the typical testing effect result is that practice test sessions actually improve final memory better than do extra study sessions. This finding has also been proposed as a possible explanation of the generation effect (Slamecka & Graf, 1978). The generation effect being the finding that items which are generated from a cue during study are better recognized than those simply read at study. One interpretation of this effect suggests that generated items are essentially retrieved from memory (i.e., past knowledge) based on the cue, whereas read items are simply perceived. Through retrieval practice, generated items would therefore be predicted to be more memorable than read items, which is exactly what is found. Of particular relevance for us, Chan and McDermott (2007) found that practice test sessions actually improve recollection on a recognition memory test, even if there is no detectable difference in hit rates between the extra study and the extra test conditions. The implication for our work being that extra practice recollecting unbelievable items during the memory test may make those items more recollectable during the inferences phase.

If we assume that in Experiment 3 participants recollect unbelievable evidence statements twice (i.e., once when the actual studied item is tested and once when the superficially foil is presented) whereas in Experiment 2 participants recollect

unbelievable evidence statements only once (i.e., only when the actual studied item is tested), we would predict that recollection for unbelievable evidence statements during the final, inference phase should be better in Experiment 3 than in Experiment 2. Further, our dual-process account of reasoning earlier proposed that, to the degree unbelievable evidence statements could be recollected, they should impact inferences similarly to an online task. Thus, we would expect in Experiment 3, where participants have had more practice recollecting unbelievable evidence statements, that these unbelievable evidence statements would have, at most, an equal impact on inferences as in Experiment 1 (which is precisely what we found). Furthermore, in Experiment 2, where participants received less practice recollecting unbelievable evidence statements during the memory test, these evidence statements may have been less recollectable during the inferences phase, and as a result have had a lesser impact on inference (which again, is precisely what we found).

Thus, the differential impact of unbelievable evidence in Experiment 2 as compared to Experiment 3 may simply have been due to the fact that by introducing superficially related foils in Experiment 3, unbelievable evidence statements now had two opportunities to be retrieved. This extra retrieval practice may have made unbelievable evidence statements particularly easy to recollect during the inferences phase of Experiment 3, resulting in these statements having a relatively larger impact compared to Experiment 2. Finally, as believable evidence statements were suggested to be less likely to be processed by System 2, these items were also suggested to be less recollectable. Thus, these testing effects may be less relevant for believable evidence statements and indeed, no noticeably large difference was observed for believable evidence statements in Experiment 3 compared to Experiment 2.

In terms of Stanovich and West's (1997) dual-process account of reasoning, these experiments provide some general support for that theory, especially if it takes into account the Spinozan model described by Gilbert (1991). The combined account we outlined in the Introduction was able to make detailed predictions about how information should be processed and subsequently reveal itself in both a memory test and in an inferences task. These predictions largely bore out in our results, supporting the account.

Furthermore, these results suggest a possible connection between the areas of memory and reasoning research. Namely, both the dual-process account of reasoning and of memory suggest that there are two fundamental processes or systems, one of which is automatic and unconscious (System 1 or familiarity) and the other which is attentive and consciously effortful (System 2 or recollection). It is of particular interest to note that these two theories converge in the sense that, believable materials are predicted by the dual-process account of reasoning to rely more heavily on System 1 and predicted by the dual-process account of memory to be more influenced by familiarity. Similarly, unbelievable materials are predicted by the dual-process account of reasoning to rely more heavily on System 2 and predicted by the dual-process account of memory to be more influenced by recollection. Thus, these two theories appear to not only be compatible but perhaps to be cataloguing the same areas of mind as one another, albeit from slightly different perspectives. This fact suggests that the dual-process accounts of reasoning may be informed from future work on the dual-process account of memory, and vice versa.

As a case in point, our particular results highlight the fact that the effect of evidence on inferences inherently depends on the nature of the task. That is, is the task

online or memory-based? Memory-based tasks appear to exaggerate the influence of believable evidence while either shrinking that of unbelievable evidence or having no effect. Future research would do well to take into account the type of task being used. Particularly, as most decisions in the real world are not online tasks, the effects of evidence on inferences and of premises on conclusions being examined in research may be misrepresenting the actual real world impact of evidence or premises.

References

Chan, J. C. K. & McDermott, K. B. (2007). The testing effect in recognition memory: A dual process account. Journal of Experimental Psychology: Learning, Memory, and Cognition, 33(2), 431-437.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. III (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related material. Journal of Experimental Psychology: General, 135(4), 553-571.

Dickstein, L. S. (1981). Conversion and possibility in syllogistic reasoning. Bulletin of the Psychonomic Society, 18(5), 229-232.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. Journal of Experimental Psychology, 58(1), 17-22.

Evans, J. St. B. T. (2003). In two minds: Dual-process accounts of reasoning. Trends in Cognitive Sciences, 7(10), 454-459.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. Memory & Cognition, 11(3), 295-306.

Evans, J. St. B. T. & Curtis-Holmes, J. (2005). Rapid responding increases belief bias: Evidence for the dual-process theory of reasoning. Thinking & Reasoning, 11(4), 382-389.

Fugelsang, J. A. & Thompson, V. A. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. Canadian Journal of Experimental Psychology, 54(1), 15-32.

Fugelsang, J. A. & Thompson, V. A. (2001). Belief-Based and covariation-based cues affect causal discounting. Canadian Journal of Experimental Psychology, 55(1), 70-76.

Geraci, L., & Rajaram, S. (2002). The orthographic distinctiveness effect on direct and indirect tests of memory: Delineating the awareness and processing requirements. Journal of Memory and Language, 47(2), 273-291.

Gilbert, D. T. (1991). How mental systems believe. American Psychologist, 46(2), 107-119.

Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. Journal of Personality and Social Psychology, 65(2), 221-233.

Glanzer, M. & Adams, J. K. (1985). The mirror effect in recognition memory. Memory & Cognition, 13, 8-20.

Gregg, V. H., Gardiner, J. M., Karayianni, I., & Konstantinou, I. (2006). Recognition memory and awareness: A high-frequency advantage in the accuracy of knowing. Memory, 14(3), 265-275.

Griggs, R. A. & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. British Journal of Psychology, 73(3), 407-420.

Hastie, R. & Kumar, P. A. (1979). Person memory: Personality traits as organizing principles in memory for behaviors. Journal of Personality and Social Psychology, 37(1), 25-38.

Hastie, R. & Park, B. (1986). The relationship between memory and judgment depends on whether the judgment task is memory-based or on-line. Psychological Review, 93(3), 258-268.

Jacoby, L. L. & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. Journal of Experimental Psychology: General, 110, 306-340.

Jacoby, L. L. & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. Journal of Experimental Psychology: General, 118, 126-135.

Joordens, S. & Hockley, W. E. (2000). Recollection and familiarity through the looking glass: When old does not mirror new. Journal of Experimental Psychology: Learning, Memory and Cognition, 26, 1534-1555.

Joordens, S., & Merikle, P. M. (1992). False recognition and perception without awareness. Memory & Cognition, 20, 151-159.

Reder, L. M., Nhouyvansivong, A., Schunn, C. D., Ayers, M. S., Angstadt, P. & Hiraki, K. (2000). A mechanistic account of the mirror effect for word frequency: A computational model of remember/know judgments in a continuous recognition paradigm. Journal of Experimental Psychology: Learning, Memory, and Cognition, 26, 294-320.

Roberts, M. J. & Sykes, E. D. A. (2003). Belief bias and relational reasoning. The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 56A(1), 131-154.

Roediger, H. L. III, & McDermott, K. B. (1995). Creating false memories: Remembering words not presented on lists. Journal of Experimental Psychology: Learning, Memory and Cognition, 21, 803-814.

Schmidt, S. R. (1991). Can we have a distinctive theory of memory? Memory & Cognition, 19(6), 523-542.

Slamecka, N J. & Graf, P. (1978). The generation effect: Delineation of a phenomenon. Journal of Experimental Psychology: Human Learning & Memory, 4(6), 592-604.

Stanovich, K. E. & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. Journal of Educational Psychology, 89(2), 342-357.

Thompson, V. A. (1996). Reasoning from false premises: The role of soundness in making logical deductions. Canadian Journal of Experimental Psychology, 50(3), 315-319.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion, and race in face recognition. The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 43A(2), 161-204.

Yonelinas, A. P. & Jacoby, L. L. (1996). Noncriterial recollection: Familiarity as automatic, irrelevant recollection. Consciousness and Cognition: An International Journal, 5(1-2), 131-141.

Appendix A:

Mean believability ratings and accuracy ratings for evidence statements and inference statements based on pre-testing measures

Believable Sentence	Mean	Unbelievable Sentence	Mean	Inference Statement	Mean
93% of all greeting cards are purchased by women	5.11	10% of all greeting cards are purchased by women	2.44	Dads are more likely to buy greeting cards than moms	2.22
A dime has 118 ridges around the edge	4.02	A dime has 15 ridges around the edge	2.89	It's pretty hard to count all the ridges on a dime	4.50
A person uses 57 sheets of toilet paper each day	3.70	A person uses 8 sheets of toilet paper each day	3.26	People only use one or two pieces of toilet paper each time they go to the washroom	2.14
A toilet has 100 times more bacteria than an office desk	3.82	An office desk has 400 times more bacteria than an a toilet	4.20	Toilets are infested with more bacteria than other areas in the house or workplace	3.36
Chopsticks originated from China 4,000 years ago	5.26	Chopsticks originated from China 100 years ago	2.91	Chopsticks are a relatively modern invention	1.64
It costs 3 cents to make a $1 bill in the United States	4.47	It costs $10 to make a $1 bill in the United States	2.31	It sometimes costs the government more to make a bill than the bill is worth	3.14
A disposable diaper can hold up to 7 pounds of liquid	4.50	A disposable diaper can hold up to 23 pounds of liquid	2.36	Disposable diapers can hold the weight of several children	3.19
The life span of a dollar bill is 1 and 1/2 years	2.84	The life span of a dollar bill is 16 years	4.00	Paper money is usually replaced every couple of years	3.75
1 billion Valentine's Day cards are sent each year in North America	5.39	500 Valentine's Day cards are sent each year in North America	2.47	Very few people actually send Valentine's Day cards out	3.17
The average North American car contains 300 pounds of plastics	4.28	The average North American car contains 2,000 pounds of plastics	3.09	Most of the weight in cars nowadays is from plastics	2.83
There are 200 parts in a typical telephone	4.09	There are only 6 parts in a typical telephone	3.74	Phones actually have a lot of complicated parts inside	5.22
18% of a person's income is spent on transportation	4.91	49% of a person's income is spent on transportation	3.41	For most people, transportation costs are easily afforded	2.67

Appendix A Continued

Believable Sentence	Mean	Unbelievable Sentence	Mean	Inference Statement	Mean
In 1962, the first Wal-Mart opened up in Rogers, Arkansas	4.70	In 1991, the first Wal-Mart opened up in Rogers, Arkansas	3.23	Walmart is a relatively new company	4.58
McDonald's restaurant has over 1.5 million employees all over the world	6.13	McDonald's restaurant has over 2,300 employees all over the world	3.56	McDonald's is a relatively large employer of people around the world	5.64
The United States Postal Service handles 40% of the world's mail volume	4.46	The United States Postal Service handles 99% of the world's mail volume	2.89	A letter mailed from anywhere in the world eventually passes through the US	1.42
25% of kids in the USA are overweight	4.27	85% of kids in the USA are overweight	4.16	Being overweight is still more uncommon for children than being a healthy weight	3.75
A blink lasts 0.3 seconds	5.76	A blink lasts 2.5 seconds	1.93	Blinks are large and noticeable motions	6.17
15% of the population is left-handed	5.15	75% of the population is left-handed	1.96	Most things are not left-handed because left-handed people are a minority	6.17
Each day 14 people die from asthma in North America	4.48	Each day 1.2 million people die from asthma in North America	2.62	Asthma kills hundreds of millions of people annually in North America	3.03
From all the oxygen that a human breathes, 20% goes to the brain	4.33	From all the oxygen that a human breathes, 95% goes to the brain	3.98	The brain needs oxygen to survive, the rest of the body doesn't really need oxygen much	1.69
It takes 3 hours for food to be broken down in the human stomach	4.78	It takes 2 weeks for food to be broken down in the human stomach	1.91	Your body is inefficient and takes many days to break down food	2.00
A person passes gas every 2 hours	4.76	A person passes gas every 10 minutes	2.44	Someone who passes gas a few times an hour is not having a normal day	4.17
People spend 33% of their life sleeping	4.87	People spend 62% of their life sleeping	3.65	People spend more time sleeping than awake	1.94
The average person falls asleep in 12 minutes	4.36	The average person falls asleep in 77 minutes	2.51	People usually fall asleep pretty quickly when they go to bed at night	3.83
American models are skinnier than 98% of American women	5.63	American models are skinnier than 12% of American women	2.91	American models have realistic body sizes	1.78

Appendix A Continued

Believable Sentence	Mean	Unbelievable Sentence	Mean	Inference Statement	Mean
The stomach of an adult can hold 1.5 litres of material	5.11	The stomach of an adult can hold 20 litres of material	2.93	One jug of pop is enough to fill up an adult's stomach	4.14
Women live 7 years longer than men do	5.67	Men live 15 years longer than women do	1.85	Grandma's usually outlive grandpa's	5.75
It takes 5 seconds for light to get from the sun to earth	3.27	It takes 8 minutes for light to get from the sun to earth	4.27	Light can travels vast distances in space almost instantaneously	5.31
Roses need 6 hours of sunlight per day to grow properly	4.80	Roses need 20 minutes of sunlight per day to grow properly	4.18	Roses can grow even with very little sunlight	3.31
90 people that have been frozen after their death	4.07	700,000 people that have been frozen after their death	2.60	Asking to be frozen after your death is a fairly rare request	5.06
25% of injuries by athletes involve the wrist and hand	4.62	4% of injuries by athletes involve the wrist and hand	3.42	Common injuries for athletes involve hands and wrists	4.67
38% of Americans eat breakfast everyday	4.64	6% of Americans eat breakfast everyday	3.07	No one really eats breakfast every day	2.31
An average American eats 60 hot dogs per year	4.49	An average American eats 7 hot dogs per year	2.87	Hot dogs are actually rarely eaten by Americans	1.83
Heinz first started making ketchup in 1876	5.20	Heinz first started making ketchup in 1233 AD	1.82	Heinz has been a merchant since the medieval times	2.03
In a year, an American kid eats 46 slices of pizza	4.76	In a year, an American kid eats 3 slices of pizza	1.64	Kids like pizza and eat a good amount in a year	5.53
80% of households have oatmeal in their kitchen	4.13	99% of households have oatmeal in their kitchen	2.77	It is incredibly rare to find a household that doesn't have oatmeal in it	3.17
90% of Pumpkins sold are for decoration	5.07	15% of Pumpkins sold are for decoration	3.80	Most pumpkins are sold around Halloween to make jack-o-laterns	5.61
The Snickers chocolate bar was invented in 1930	4.91	The Snickers chocolate bar was invented in 1996	2.24	Snickers has been around since the Great Depression	3.94
A crocodile can run up to a speed of 16 kilometres per hour	4.39	A crocodile can run up to a speed of 100 kilometres per hour	2.63	Crocodiles can run at highway speeds because they're so low to the ground	2.08

Appendix A Continued

Believable Sentence	Mean	Unbelievable Sentence	Mean	Inference Statement	Mean
A leech has 1 brain	5.13	A leech has 32 brains	2.23	You can't cut off a leeches head because it has brains all over its body	2.89
A female mouse can produce up to 100 babies a year	4.67	A female mouse can produce up to 200,000 babies a year	2.54	A handful of mice can produce millions of babies per year	4.11
Rats can survive up to 14 days without any food	4.93	Rats can only survive for 20 minutes without any food	2.04	Rats usually fast for several days at a time as food is usually scarce	4.67
Alaska has 2 times as many caribou as people	4.42	Alaska has 100 times as many people as caribou	3.16	Alaska has more wildlife than people	5.14
31% of employees skip lunch entirely	4.93	85% of employees skip lunch entirely	3.29	Eating lunch at work is not the norm	5.86
25% of Americans don't know that the sun is a star	4.40	95% of Americans don't know that the sun is a star	3.65	Only scientists tend to know that the sun is a star	4.08
85% of weddings are held in a synagogue or church	5.02	7% of weddings are held in a synagogue or church	2.36	People usually get married in religiously sacred buildings	4.75
96% of candles that are purchased are purchased by women	5.15	8% of candles that are purchased are purchased by women	2.13	Boyfriends are more likely to buy candles than girlfriends	1.78
women spend 55 minutes per day getting showered and dressed	5.07	women spend 3 hours per day getting showered and dressed	3.38	Women usually wake up an hour early to get ready in the morning	4.97
40% of the states in the U.S. have severe, or extreme pollution problems	4.91	90% of the states in the U.S. have severe, or extreme pollution problems	4.11	Pollution still isn't a major problem for most of the states in the US	2.67
5% of the people who use personal ads for dating are already married	4.53	62% of the people who use personal ads for dating are already married	3.07	If you meet a person from a personal ad, chances are they are already married	2.67
33% of accidental deaths occur in the home	4.43	99% of accidental deaths occur in the home	2.84	You're more likely to die during work, school, or in transit than at home	3.81
93% of children go out trick or treating for Halloween	4.84	17% of children go out trick or treating for Halloween	2.50	Trick-or-treating is still a very common yearly tradition	5.08

Appendix A Continued

Believable Sentence	Mean	Unbelievable Sentence	Mean	Inference Statement	Mean
8% of men are color blind	5.02	71% of men are color blind	2.80	The reason most men are bad at colour-coordinating is that they are colour blind	2.28
Only 4% of babies are born on their actual due date	4.46	79% of babies are born on their actual due date	3.80	Doctor's are remarkably accurate in predicting the exact day a baby will be born	3.17
50% of lottery players go back to work after winning the jackpot	4.67	92% of lottery players go back to work after winning the jackpot	4.28	Pretty much no one quits their job when they win the jackpot in a lottery	3.25
30% of the human population reside in deserts	3.86	70% of the human population reside in deserts	2.47	Most cultures around the world are desert-dwelling cultures	5.19
40% of people end up marrying their first love	3.84	90% of people end up marrying their first love	2.65	The reason most marriages end if divorce is that most people marry their first love	2.03
In the United States, 33% of land is covered by forests	4.64	In the United States, 87% of land is covered by forests	2.85	If you drive across the US most of the drive you will be driving through forests	2.19
12 men have landed on and explored the moon	3.98	216 men have landed on and explored the moon	2.78	Because moon missions are so rare, only a handful of men have ever landed on the moon	5.06
Hitler was voted Time Magazine's man of the year in 1938	3.95	Hitler was voted Time Magazine's man of the year in 1981	2.59	Before World War II Hitler was actually a popular leader due to his economic reforms	5.36
The world's tallest roller coaster reaches a peak height of 72 meters	4.73	The world's tallest roller coaster reaches a peak height of 6 meters	2.02	The world's tallest roller coaster is about as tall as a tall man	1.14
James Bond made his debut in the 1952 novel "Casino Royale"	4.38	James Bond made his debut in the 1765 novel "Casino Royale"	3.11	The first James Bond tale was written hundreds of years ago	2.50
Majority of brides plan their wedding for 9 months	5.43	Majority of brides plan their wedding for 5 years	2.45	Most weddings are planned in under a year and therefore require the help of a wedding planner	2.36