THE UNIVERSITY OF CHICAGO

PARADOXES AND PROBABILITIES: THE CONJUNCTION PROBLEM AND LAY STRATEGIES FOR COMBINING ELEMENTS IN LEGAL CLAIMS

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE UNIVERSITY OF CHICAGO

BOOTH SCHOOL OF BUSINESS

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

KRIN IRVINE

CHICAGO, ILLINOIS

MARCH 2024

For my grandmother, Lottie. Thank you for being there. I hope you're proud.

For Bill Smith, who looked me straight in the eyes in high school while telling me to "get a PhD."

Thank you for your belief in me, and the invitation to believe in myself.

Contents

List of Figures

List of Tables

Acknowledgments

During my time at Booth I have had the opportunity to grow as a researcher and learn from so many people. Thank you to the many professors and students who were a part of that journey.

I also want to thank Jane, as head of the Behavioral Science PhD program, along with the entire PhD office team, for the above and beyond support I received which made it possible for me to get to the finish line.

Thank you to Oleg and Tess for being a part of my dissertation committee, helping make this current project better, and for everything I've learned from you over the years.

To George and Reid, the co-chairs of my dissertation committee, you have done all of the above and more. Your unwavering support and tenacity to see this through with me is the reason I am able to have a finished dissertation and to earn my PhD. I am extremely grateful for your guidance on all of the projects we've worked on together, including this dissertation. Thank you for the ideas, challenges, mentorship, care, patience, and sense of humor.

Abstract

In a civil legal case, the plaintiff usually has to prove their case at a “more likely than not” standard (>50%) in order to win their claim. Legal claims typically involve multiple individual elements that also need to be proven to the same standard. In a civil claim with two elements relying on independent evidence, if the likelihood of each element was 60%, it is debatable whether the claim should win or lose. Each element has met the threshold and intuition suggests that the claim should win. However, probability theory would suggest multiplying these probabilities to reach a 36% likelihood of the overall case, which would dictate a loss for the plaintiff. Some state legal instructions contain a “conjunction problem” by stating both that a decision-maker should decide the claim based on “all of the evidence” as well as a statement that the claim is required to win if “all of the elements” are found to be true.

Across four studies we examine how lay participants choose to combine elements into overall case-level decisions. We find very few people naturally follow a multiplicative combination rule across various contexts. In Studies 1 and 2 participants read detailed case descriptions and estimate their personal element probabilities. In Study 1 we find participants are sensitive to case strength details, but provide equivalent answers to two-element versus four-element cases, which sharply violates the multiplication rule. In Study 2 we do not observe differences based on whether or not the conjunction problem is included in the legal instructions. In Studies 3 and 4 participants read abstracted case descriptions and we provide stipulated element probabilities across multiple cases, allowing categorization of participants into combination strategies. In Study 3 we stipulate numeric probabilities and find that an introductory tutorial about probabilities reduces randomness and increases the use of

conjunctive multiplication. In Study 4 we stipulate verbal probabilities and find that translating these verbal probabilities into numeric ones on each case page reduces randomness. Across these studies we find that most people are averaging the strength of evidence for the elements to reach an overall conclusion about the case, and that the conjunction-multiplying rule is followed by less than 10% of the respondents.

Introduction

When faced with a choice between two or more alternatives, people often accumulate evidence until they feel they have enough information to make a decision or to take action. People do this every day, whether deciding which experts to believe, choosing whether to wear a mask, or determining how many times to eat burnt toast before replacing the toaster. Scientific and applied research often relies on statistical criteria to determine if a result is supported by enough data or is statistically significant before treating the finding as reliable and well-established enough to have an impact on theoretical or practical conclusions (Benjamin, et al., 2017; Wasserstein & Lazar, 2016; Wilkinson & APA Task Force, 1999).

In the legal system, decisions are defined by a burden of proof (which party is required to prove their case) and by a standard of proof that defines the level of certainty or amount of evidentiary support necessary to reach a decision on the verdict. Legal doctrine and rules of practice prescribe how these two concepts combine the formality of statements of the law with the informality of human decision processes.

Civil legal cases usually place the burden of proof on the plaintiff and require a “preponderance of the evidence” standard of proof. This means the plaintiff in a case has to convince the decision-maker that their version of events is more likely to be true than not. If we express this standard of proof numerically, we might say the threshold to prevail is about 51%.

The standard of proof becomes more complex when we consider that legal claims almost always involve multiple elements that also have the same standard of proof threshold. For example, we might consider a simplified civil dispute in which the plaintiff needs to prove two elements to the judge or jury. First, they need to prove that the defendant did something

wrong. Second, they need to prove that they were actually harmed. If the judge or jury believes the evidence demonstrates that each element is 60% likely to be true, what does that imply about the likelihood that the overall claim should succeed? Intuitively, it seems as if the plaintiff has satisfied the standard of proof, as each element has been proven beyond the preponderance standard, and therefore the plaintiff should win the overall claim. We will make one more assumption: that the evidence provided for each issue was completely independent. This does little to change the intuitive answer, but in this situation probability theory tells us to multiply these probabilities to get an outcome of 36% likely that the plaintiff's case has been proven (.60 X .60), which would then compel a loss of the overall claim.

In both the real-world and academic legal discussions, there is disagreement about what the outcome should be for multi-element decision questions like these. States have differing jury instructions about how to handle this situation, and careful reading and application of these instructions could result in conflicting answers.[1] Thus, under some actual jury instructions, a "conjunction problem" (or "paradox") can arise when combining the element findings into the overall case finding.

Legal evidence scholars debate the normative answer to the posed question and also whether there is any conjunction problem with how the law currently handles situations like this. One view suggests that jurors should, and already do, assess the overall "relative plausibility" of

[1] Schwartz & Sober (2017) provide this example of the conjunction problem from DC: "[t]he party who makes a claim ..has the burden of proving it. This burden of proof means that the plaintiff must prove every element of [his/her] claim by a preponderance of the evidence." combined with "If [Plaintiff] proves each element, your verdict must be for [Plaintiff]. If [Plaintiff] does not prove each element, your verdict must be for [Defendant]." The states without the conjunction problem can leave room to interpret that you could still find for the Defendant even if every element is proven. A stronger version is Schwartz & Sober's "aggregate elements," where there is language that indicates considering the elements together as a whole, like their example from the Eighth Circuit: "Your verdict must be for plaintiff ... and against defendant ..on plaintiff's claim ... if all the following elements have been proved."

competing explanations and then choose the one that seems most likely (Allen & Pardo, 2019a, 2019b; Pardo & Allen, 2008). In this way, if the decision-maker compared more than one less likely than not explanation, they would still choose the highest among their available options, ignoring any formal burden of proof requirements.[2]

A second view is that in most jurisdictions, even if all the elements of a claim are found to be proven, this is not conclusive, and the decision-makers must also believe the evidence threshold is being met for the overall claim in order for the plaintiff to win (Schwartz & Sober, 2019, 2017). These authors extensively detailed the law and instructions across states and determined that most do not give rise to the conjunction problem. They also argue that much of the time, elements are probabilistically dependent, so the simple multiplication of probabilities is not even the correct mathematical way to combine elements.[3]

A third theory proposes changing the probability requirements that need to be met across different kinds of cases to balance the costs and benefits of verdicts with different social consequences (Nance, 2019, 2016). The motivation for this theory is that the requirements of proof should be stricter (the threshold to convict or to find for the plaintiff should be higher) when the stakes in the case are higher (severity of the punishment, amount of the award being contested).

---

[2] For example, if the decision-maker was breaking down their entire set of beliefs, they might think they believe the plaintiff's explanation is 35% likely, the defendant's explanation is 25% likely, and there's a 40% likelihood that there is some other explanation entirely. Since they have to find for one of the parties, they take the best of the options available and find in favor of the plaintiff.

[3] In our discussion of probability theory, we will usually ignore the complexity of interdependent probabilities. But we address the dependency complexity with our empirical methods and demonstrate violations of the probability-theory calculation, whether or not the component probabilities are independent or dependent. Briefly, if two component event probabilities (A and B) are independent, the conjunction probability (that both events occur) is p(A and B) = p(A) X p(B). If they are dependent, the calculation is more complex: p(A and B) = p(A) X p(B|A), or "probability of event A multiplied by the probability of event B given that event A has occurred."

A fourth theory replaces conventional probability theory with a novel interpretation in terms of belief functions (Clermont, 2019, 2015).[4] This theory compares the minimum probability that the plaintiff proved their case with the maximum probability that the defendant proved their case. This process is related to the element-by-element approach the law already requires.

Most of the discussion in this area is theoretical instead of empirical. The goal of the present research is to provide some descriptive information about how typical jurors reason when confronted with conjunction reasoning problems. Although the legal discussion and disputes are focused on normative questions of the proper manner of reasoning about conjunctions, some scholars have made arguments based on claims about jurors' behavior. In any case, information about jurors' behavior can tell us whether they are likely to violate normative principles when they decide based on their common sense and intuition.

In the psychological research on conjunction reasoning in legal contexts we found only two empirical studies that address the question of how mock jurors respond to the conjunction problem. Goldsmith (1978) presented Norwegian law students with three examples of criminal cases in which evidence items could be described in terms of component and conjoint probabilities. The students were asked to assign probabilities to the evidence items. For example, in a car-theft case, the students were asked to estimate: "What is the probability a cigarette lighter (found in the stolen car) belongs to the defendant?" "What is the probability that the lighter was left in the car by the defendant?" "What is the probability the lighter belongs to

[4] The belief functions approach does not lead to a single standard of proof value; rather, it calculates a distribution of strength of evidence across multiple hypotheses (e.g., strength of evidence supporting the plaintiff, supporting the defendant, supporting both parties, supporting neither party) (Shafer, 1990, 1976). For purposes of the present research, we ignore the belief functions model, as there are no current proposals to implement such a standard of proof in practice, and it is unclear how jurors might be instructed to apply such a standard.

the defendant AND that the defendant left the lighter in the car?" Then, the experimenter calculated the conjoint probability from the two-component events and compared it to the same student's direct assessment of the conjunction event. The general finding was that participants' conjunction estimates were higher than the conjunctions calculated from the estimates of the component probabilities.

Goldsmith's experiment does not address our primary interest in jurors' reasoning about elements of a verdict. However, his conclusions suggest that there is heterogeneity in lay reasoning about events and their conjunctions. He concluded that his students could be sorted into three "combination rule" strategies: (A) students who estimated the conjunction as the average of the two-component probabilities, (B) those who overweighted the higher component probability, and (C) those who overweighted the lower component probability. We will also observe considerable heterogeneity in the present studies.

Another relevant empirical study compared special verdict forms versus general verdict forms in a civil mock trial (Wiggins & Breckler, 1990). Special verdict forms typically ask jurors to make decisions on a number of relevant elements before providing an overall case decision. Participants in the Wiggins & Breckler six-hour study watched videos of different stages of a trial and filled out two verdict forms. The authors did not observe differences in trial outcome based on the type of form used, though they did find that special verdict forms resulted in higher damages awards. The special verdict form used in the study did dictate a finding for the plaintiff if all of the plaintiff's elements were proven by a preponderance of the evidence, so theoretically, it could have contained some examples of the conjunction problem, though that was not a factor of consideration for the study.

Turning to research by psychologists outside the legal context, the most common results reflect Goldsmith's findings. Even with explicit numerical probabilities, people tend to overestimate the conjunctive probability of multiple events (Bar-Hillel, 1973). For example, when making choices between a simple gamble (drawing one marble from an urn) versus a compound gamble (drawing multiple marbles, with replacement, from an urn), they overestimate the success rate of the compound gamble.

There are multiple explanations for how this kind of error occurs. Think-aloud tests suggest that people might be engaging in an anchor-and-subtract method of getting to a conjunction probability estimate (Doyle, 1997). Recent research comparing various models suggests that people might anchor on the lowest value and adjust from there (Fan et al., 2019).

Perhaps the most famous example of overestimation of conjunctions is the "Linda Problem" (Tversky & Kahneman, 1983), where lay people believe that a fictional Linda is more likely to be "a bank teller and a feminist" than simply "a bank teller." Although the "Linda Problem" is not stated numerically, and it is likely to involve different reasoning processes than its numerical analogs, it provides a strong suggestion that lay people do not generally respect the multiplicative conjunction probability rule.

In the present research, study participants were asked to make decisions about civil dispute scenarios that pose various conjunction reasoning problems. Study 1 tests experimental participants' judgments of cases in which the overall verdict depends on either two or four elements. We obtain the surprising result that, on average, participants are completely insensitive to the number of elements. This provides a strong challenge to any interpretation that assumes the decisions are rational or consistent with any interpretation of normative legal

principles. Study 2 tests participants' sensitivity to variations in the instructions provided on how to decide the overarching verdict in a multi-element case. Again, the basic result is that participants do not respond any differently when they are instructed to apply two different legal principles to the multi-element decision problem. These results imply that jurors will not decide multi-element verdicts in a manner that can be interpreted as rational or as consistent with the expectations of how verdict instructions should affect decisions.

In Studies 3 and 4, we shift to the question of what psychological models describe participant reasoning processes. In Study 3, we provide participants with numerical probabilities expressing the strength of evidence supporting each element. In Study 4, we present the strength of evidence with verbal probability terms and ask the participants to provide us with their numerical estimates of the strength of proof. Then, we identify the combination rules that describe the reasoning processes of typical participants. We follow up by categorizing participants according to each individual's dominant reasoning strategy. Overall, we find only a few participants reason consistently with the multiplicative probability theory rule for inferring conjunction probabilities, regardless of the instructions they are given. Furthermore, there is considerable heterogeneity across individuals, such that individuals can be sorted into a few common reasoning strategies.

Following the implications of the behavioral research we just reviewed, we represent alternate reasoning strategies as algebraic combination rules applied to subjective degrees of belief, usually called subjective probabilities or decision weights. One proposal is that nonexperts reason approximately according to the rules of elementary probability theory when making uncertain judgments and decisions, and that the probability of the conjunction is a

multiplicative function of the probabilities of the components (Benjamin, 2019; Peterson & Beach, 1967; Phillips & Edwards, 1966).

However, as noted above, behavioral research has identified a variety of combination rules, most often expressed in the form of simple or weighted averages (Bar-Hillel, 1973; Doyle, 1997; Fan et al., 2019). The averaging judgment process is usually described as serial adjustment or anchor-and-adjustment with attention initially focused on early, recent, or extreme values (e.g., “anchoring” on the first event’s subjective probability and “adjusting” for the values associated with later events, cf. Anderson 1981; Tversky & Kahneman, 1974). In the present analyses, we classify participants into categories according to the combination rule that best describes their conjunction reasoning strategy. Our identified strategies include conjunction (multiplying the component probabilities), simple averaging, overweighting the highest component, and overweighting the lowest component. In addition to these algebraic strategies, we will also consider strategies that favor one of the two elements in a multi-element decision.

Another contribution we make in this paper is providing empirical data about the question of how participants resolve cases when the conjunction problem could exist. The dilemma exists if the combined probability dictates a loss while considering the elements one at a time dictates a win. We name these instances of “Conjunction Divergence” and find that most of the time, participants in this situation choose an overall win for the case, even though the multiplicative conjunctive combination rule implies the plaintiff loses.

Study 1: Number of Elements and Strength of Evidence

This first study aimed to determine whether participants behave differently based on whether they have seen two- or four-element cases, given either strong or weak case evidence. Based on pilot testing, we predicted we would detect no difference between two and four elements but a significant difference between case evidence strengths.

Methods

The hypotheses were tested using a 2 (plaintiff's case strength: *weak* vs. *strong*) x 2 (number of elements: *two* vs. *four*) between-subjects experimental design. The sample size was determined prior to data collection via a power analysis, which suggested a per-cell sample size of 195. (See Appendix for more details on the power analysis.) Given four cells, 800 participants were recruited on Amazon Mechanical Turk (MTurk) and paid $1 to complete an online survey. Our final data set includes 766 participants.[5]

After completing generic consent and introductory information, participants were given a basic description of probability and then asked four probability questions. There was a simple probability (one die roll), a simple conjunction probability (two coin flips), a complex conjunction probability (two answers with different probabilities), and a filler (weather). (See Appendix for

[5] Participants were recruited via an MTurk HIT with a target size of 800 participants. MTurk IDs were collected at the beginning of the survey, and IP addresses were tracked. Before downloading data, partial survey responses were closed and recorded for comparison against the completed surveys. Removed from the data set are any survey results that had a duplicate ID or duplicate IP address, or did not have a matching MTurk HIT, and the partially completed surveys. All of these data removals were performed before any data analysis. Furthermore, an additional 6 participants were removed because of incomplete data (survey edits removed a forced response requirement, and this small number of people left one of the element probability fields blank).

complete instructions and questions.) Participants entered a number for each question into a text box and then were allowed to continue with the survey regardless of their answers.

Next, participants viewed a page of details about a hypothetical court case between a plaintiff, Bill, and a defendant, Steve. They read that Bill had bought a house from Steve and that there was now a dispute about repair costs for a plumbing problem. Steve had been aware of the problem and hired a plumber to fix it, and on a seller disclosure form, he had listed the plumbing as in “excellent” condition. We randomly assigned participants to one of two plaintiff case strength conditions: weak or strong.

In the strong condition, participants saw that Steve thought he had a one-year fix for the plumbing problem, that Bill had hired an excellent inspector and the problem would have been difficult to discover, that the form was reviewed carefully and believed to be accurate, and that the plumbing issue would have significantly decreased the sales price of the house.

In the weak condition, participants saw that Steve thought he had a permanent fix for the plumbing problem, that Bill had hired a cheap inspector and was warned about a potential problem, that the disclosure form usually included inaccurate statements, and that the plumbing issue would not have affected the sales price of the house. (See Appendix for full text.)

After the details of the case, participants were told that the “attorneys agree that the above facts are accurate” and to “imagine you a jury member in this case” before moving to the next page.

The next page included legal instructions. This page informed participants that the lawyers “stipulated that listing the plumbing as ‘excellent’ on the Seller's Disclosure form was a

false statement."[6] Participants were given both a whole claim requirement ("To win his case, Bill needs to prove that it is more probably true than not true that Steve committed fraud by making this false statement.") as well as a requirement for each element ("The plaintiff needs to prove that each of the following propositions is more probably true than not true...") before seeing case elements.

The potential case elements that participants could see were as follows.

Element 1 (e1): The false statement was of a material fact.

Element 2 (e2): Steve knew the statement was false.

Element 3 (e3): Bill reasonably believed the statement.

Element 4 (e4): Bill's damages resulted from his reliance on the statement.

Participants were randomly assigned to one of two conditions varying in the number of case elements: *two* or *four*. In the *four*-element condition, participants saw all four of the above elements. In the *two*-element condition, participants saw only two elements: first, either e1 or e2 and then either e3 or e4. Finally, depending on which elements they had seen, participants were also provided with definitions of "material," "knowing," "reasonably believed," and/or "reliance on a statement." (See Appendix for full text.)

Next, participants saw four pages on which to answer questions about their thoughts on the case. They were asked about the overall case by both a binary question ("Do you find for the Defendant or for the Plaintiff?" with answer choices of "I find in favor of Bill and against Steve" and "I find in favor of Steve and against Bill") and a probability question ("What is the

[6] Earlier pilot studies did not include this stipulation. It was added to reduce potential dependency between the elements. Without it, most of the elements refer to the "false statement," creating interdependency between answers.

probability that the Plaintiff (Bill) has proven his overall case against the Defendant (Steve)?") They were also asked about each of the case elements by both a binary question ("For each proposition below, indicate whether you find that it is more probably true than not." "Do you find that…" with the answer choices of "Yes" or "No" for each element) and a probability question ("For each proposition below, indicate what the probability is that it occurred." "What is the probability that...").

The probability questions contained an open-ended text box that would accept any numerical answer. The order in which questions were asked was also randomized for each participant. Participants either saw the overall case-question pages first or the case-elements pages first. Within each category, they also either saw the binary question(s) or probability question(s) first.

After these four pages of questions, participants were asked to provide importance ratings on a scale of 1 ("Not at all important") to 7 ("Extremely important") for each element they had seen. All of the case pages (case information, legal instructions, four pages of binary/probability questions, importance questions page) had a common header with "Plaintiff: Bill (buyer) vs. Defendant: Steve (seller)." Additionally, the five pages with questions provided links participants could click on in order to display either the case information or legal instructions again. The last page of the survey had a variety of demographic questions.

Results

The analyses for this study will start by looking at the effects of conditions on overall case decisions, including a discussion of an equivalence testing method. The next section will

examine the internal consistency between participants' case outcome decisions. The final section will be an analysis of conjunction divergence frequency and those case outcomes.

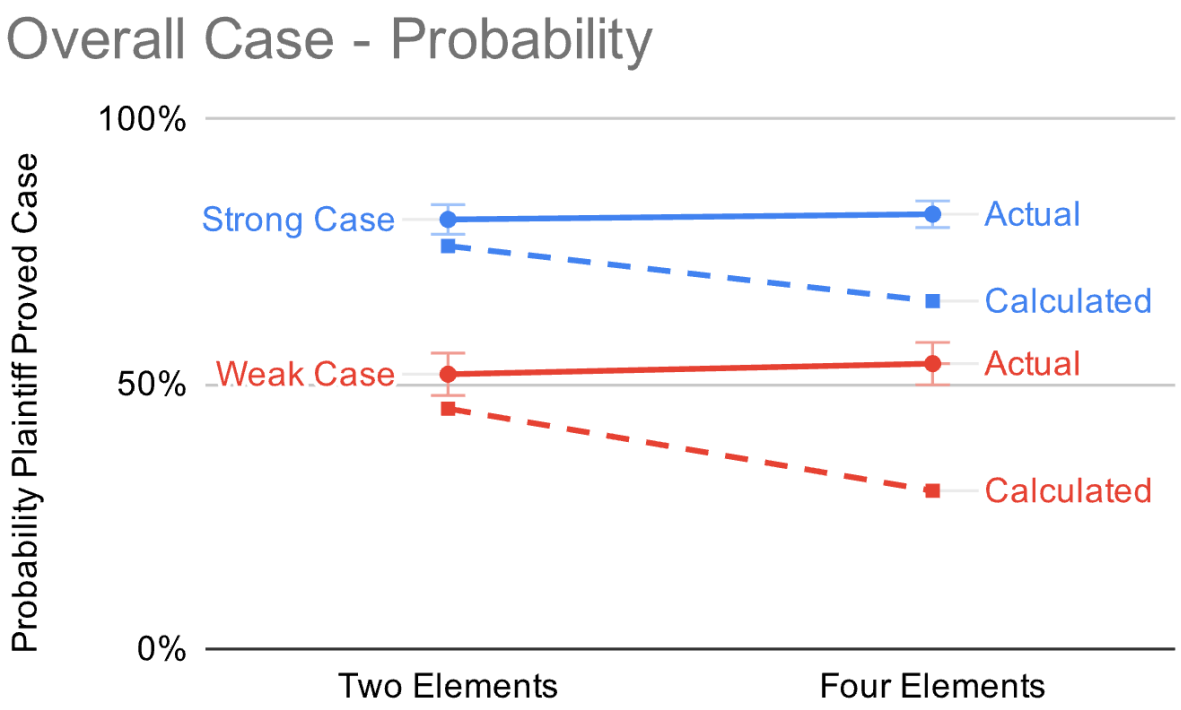
Case Outcomes by Condition

As predicted, participant answers were more in favor of the plaintiff when the *strong* case evidence was provided, compared to the *weak* case evidence. This was the case for overall case probabilities (*strong*: M=82, SD=18; *weak*: M=53, SD=28), t(645.9)=16.56, p<.001, and overall case binary decisions (*strong*: 94%; *weak*: 55%), χ2(1, N=766)=151.77, p<.001.[7] The patterns from these tests were similar for most comparisons between element questions and when only considering subsets of the data by condition. (See Appendix for full results.) This main effect of strong versus weak case evidence demonstrates that participants were paying attention to the scenario details and responding to differences.

Also, as predicted, significant differences were not observed when comparing *two* versus *four* case elements. This was the case for overall case probabilities (*two*: M=66, SD=28; *four*: M=68, SD=27), t(764)=0.77, p=.441, and overall case binary decisions (*two*: 75%; *four*: 74%), χ2(1, N=766)=0.17, p=.338 (see Figure 1).

[7] The t-tests reported here and elsewhere in the paper are Welch two sample t-tests.

Figure 1. Study 1 Case Probability Outcomes



In addition to not rejecting the null hypothesis that these means and proportions were equal using standard tests, the two one-sided tests (TOST) procedure was used to perform equivalence tests (Lakens, 2017). For equivalence test power calculations (in Appendix) and the equivalence bounds in TOST testing, we use effect sizes of Cohen's *d* (for means) and Cohen's *h* (for proportions) equal to −0.3 and 0.3, which correspond to the ability to detect a "small" effect size (Cohen, 1998). A significant equivalence test (combined with a nonsignificant null hypothesis test) indicates that differences are statistically equivalent to zero.

When comparing two versus four case elements, the equivalence test showed results statistically equivalent to zero for overall case probabilities (ΔL=-8.4, d=-0.3; ΔU=8.4, d=0.3),

t(764)=3.38, p<.001, and overall case binary decisions (ΔL=-0.12, h=-0.3; ΔU=0.12, h=0.3), z=3.28, p=.001.

For the overall case probability, an additional, more stringent equivalence test was performed based on finding a difference at least as large as the smallest difference if participants were following a conjunction multiplication rule. (See Appendix for details on this calculation.) Since the known products had a directional difference, an inferiority equivalence test (testing only in one direction) was used. This equivalence test using these more restrictive bounds was significant (ΔU=3.8, d=0.13), t(764)=2.64, p=.004.

These equivalence tests demonstrate that participants are insensitive to the change between two versus four case elements. This supports the idea that participants are not engaging with overall probabilities as conjunctions. This is explored further in the sections below.

Case Outcome Consistency

Subjects were, in general, consistent in their overall case decisions. When they chose the binary option of the plaintiff winning, their probability ratings were higher (plaintiff wins: M=78, SD=19; plaintiff loses: M=35, SD=25), t(274.8)=22.07, p<.001. This was true for both the *strong* evidence (plaintiff wins: M=83, SD=17; plaintiff loses: M=61, SD=25), t(24.4)=4.19, p<.001, and *weak* evidence cases (plaintiff wins: M=70, SD=20; plaintiff loses: M=32, SD=23), t(337.6)=17.57, p<.001.

The majority of participants (~73%) were internally consistent with matches between their own probability rating and binary option, but there were consistency errors. Of the participants who said that the plaintiff should win, 12% (68/571) did not observe the 51% cutoff

(for more probably true than not true) and chose the case in favor of the plaintiff despite having probability ratings that were not at least 51% (this number drops to 5%, 27/571, when also including probability ratings that were exactly at 50%). This technical error was more frequent in the participants who saw the *weak* evidence cases versus the *strong* evidence cases (probability<51: *strong*: 6%; *weak*: 22%), $\chi 2(1, N=571)=27.67$, p<.001, (probability<50: *strong*: 3%; *weak*: 8%), $\chi 2(1, N=571)=5.29$, p=.021.

Conjunction Divergence

In each study, in addition to that study's main analysis question, we pay special attention to cases that we name "conjunction divergence." Conjunction divergence occurs when there is a difference between a decision based on whether each element wins versus a decision based on whether the conjunctive product of element probabilities wins. (See Table 1 for examples.)

Table 1. Conjunction Divergence Illustrative Examples

Probability of 1st Element	Probability of 2nd Element	Product of Probabilities	Elements Outcome	Product Outcome	Conjunction Divergence?
20%	40%	8%	Lose	Lose	No
20%	60%	12%	Lose	Lose	No
60%	70%	42%	Win	Lose	Yes
70%	90%	63%	Win	Win	No

For each study, we will note how frequently conjunction divergence occurred. (In Study 3, we stipulated element probabilities, so we were in control of this.) Then, within these instances of conjunction divergence, we will examine the rate at which participants choose that

the case should win (aligning with the outcome based on elements) versus that the case should lose (aligning with the outcome based on conjunction/mathematical product).

In this study, 5.4% of the 766 cases had conjunction divergence. Of these 41 cases, 80.5% had winning case outcomes (matching elements), while the remaining 19.5% were losses (matching conjunction probability).

There were no statistically significant differences observed in the rate of conjunction divergence across evidence strength (*strong*: 4.1%; *weak*: 6.6%), $\chi 2(1, N=766)=2.24$, p=.135. Within conjunction divergence cases, all participants with *strong* cases (100.0%) and fewer participants with *weak* cases (68.0%) chose a case outcome matching the elements winning, $\chi 2(1, N=41)=6.36$, p=.012. This is not surprising since, even within the small number of conjunction divergence cases, overall probability answers were significantly higher for *strong* evidence cases than *weak* evidence cases (*strong*: M=74.7, SD=6.7; *weak*: M=63.4, SD=18.5), t(32.7)=2.79, p=.009, similar to the pattern observed for the whole set of cases.

The rate of conjunction divergence varied across the number of elements (*two*: 3.6%; *four*: 7.1%), $\chi 2(1, N=766)=4.65$, p=.031, which is to be expected since, by definition, conjunction divergence requires a multiplied probability product less than 50%, and multiplying four elements together will achieve that more readily than multiplying two elements. Within conjunction divergence cases, no differences were observed in the number of participants choosing a case outcome matching the elements winning (*two*: 71.4%; *four*: 85.2%), $\chi 2(1, N=41)=1.11$, p=.292. Within these conjunction divergence cases, overall probability answers were significantly higher for cases with more elements (*two*: M=58.2, SD=19.5; *four*: M=72.7, SD=11.2), t(17.6)=2.58, p=.019.

Study 1 Summary

This study used detailed scenario text to communicate the strength of the evidence in support of the plaintiff's case. The cases consisted of either two or four elements. This study showed that participants were sensitive to evidence strength, with their ratings and case findings increasing with stronger evidence. This study also showed that participants were almost perfectly insensitive to the number of elements required to prove a plaintiff's case. This second finding provides a dramatic contradiction of the Conjunction Rule for combining per-element evidence strength to yield an assessment of the global strength of the plaintiff's case.

Study 2: Varying Legal Instructions

The next study aimed to determine whether participants respond to changes in legal instructions that specify information directly relevant to the conjunction problem.

Methods

The study used a two-condition (legal instructions: *non-conjunction problem* or *conjunction problem*) between-subjects experimental design. We recruited 100 participants on MTurk who were paid $1 to complete an online survey. Our final data set includes 93 participants.[8]

The overall structure of this study was similar to the first study, including the probability quiz, case information, legal instructions, case and element decisions (with randomized order), and then demographic information. For this study, only the weak version of the case information (see Appendix) were used to allow for a more even split between plaintiff wins and losses. Additionally, all participants saw all four case elements from the four-element condition of Study 1.

The experimental manipulation occurred on the legal information page. Some of the participants saw an instruction that could be considered a conjunction problem instruction, while others saw a non-conjunction problem version. According to Schwartz & Sobers's classifications (2017), these would be considered, relatively, an "Elements Only, Mandatory" instruction and an "Entailment Check" instruction.

[8] We used the same data removal process detailed in footnote 5.

In the *conjunction problem* condition, participants read, “To win his case, Bill needs to prove that each of the following propositions is more probably true than not true:” and “If you find from your consideration of all the evidence that all of these propositions are more probably true than not true, then your verdict should be for Bill.” (Emphasis added for this paper. These phrases were not highlighted for the research participants.)

In the *non-conjunction problem* condition, participants read, “To win his case, Bill needs to prove that it is more probably true than not true that Steve committed fraud by making this false statement. In a fraud case, the plaintiff needs to prove that each of the following propositions is more probably true than not true:” (See the Appendix for full text of both conditions.)

These instructions were modeled on actual jury instructions. Both versions included the requirement that “each of” the propositions needs to be proven true. The *conjunction problem* instruction further demanded that if the propositions are “all” true, then finding in favor of the plaintiff is necessary. The *non-conjunction problem* instruction left room for the participant to find against the plaintiff, even if the propositions are all true.

Results

The analyses for this study will start by looking at the effects of condition assignment for overall case decisions and end with an analysis of conjunction divergence frequency and those case outcomes.

Case Outcomes by Condition

A main effect of condition was not observed for overall case binary decisions (*conjunction problem*: 57%; *non-conjunction problem*: 54%), $\chi 2(1, N=93)=0.09$, p=.763 or

overall case probabilities (*conjunction problem*: M=58, SD=27; *non-conjunction problem*: M=56, SD=28), t(90.7)=0.38, p=.704. This was true of the binary and probability measures for each of the four elements as well. However, there was not enough power to reject a small effect size through equivalence testing for any of the measures.

Conjunction Divergence

In this study, 5.4% of the 93 cases had conjunction divergence, when every element had a winning probability, but the products of these probabilities multiplied were less than or equal to 50%. Of these 5 cases, 80% had winning case outcomes (matching elements), while the remaining 20% were losses (matching conjunction probability).

There were no differences observed across conditions in the rate of conjunction divergence (*conjunction problem*: 6.4%; *non-conjunction problem*: 4.3%), $\chi 2(1, N=93)=0.19$, p=.664, in choosing a case outcome matching the elements winning (*conjunction problem*: 100%; *non-conjunction problem*: 50%), $\chi 2(1, N=5)=1.88$, p=.171, or in overall probability answers (*conjunction problem*: M=78.3, SD=7.6; *non-conjunction problem*: M=82.5, SD=10.6), t(1.7)=0.48, p=.686. However, again, there was not enough power to reject a small effect size through equivalence testing.

Study 2 Summary

This study used textual materials to communicate a weak evidence case in favor of the plaintiff. Each case consisted of four elements, and the instruction concerning the combination of per-element evidence into a global case finding was manipulated. In one treatment, a conjunction combination rule was presented. In the other treatment, a non-conjunction combination rule was presented. The manipulation had no effect on case finding decisions,

suggesting that participants do not appreciate the difference between a conjunction requirement versus a per-element standard of proof.

Study 3: Individual Strategies

This study investigated how participants combine probabilities into one overall answer when we specified numeric probabilistic elements over multiple generic cases. Based on pilot testing, we identified a number of combination strategies to assign people into and preregistered our method of doing so.[9] In this study, we continued to look at how frequently people use or ignore multiplicative conjunction calculation, but here, we did so with a defined set of cases we created with precise probabilities, which means we were able to look at conjunction divergence behavior for every participant. Finally, in order to refine our methods, we aimed to test whether or not our background information and questions about probability would cause more people to use multiplicative conjunction calculation.[10]

Methods

We recruited 250 participants on Prolific, and they were paid $2.60 each to complete an online survey. Our final data set includes 251 participants.[11] Participants in this study saw 12 case pages in an order that was fully randomized within subjects. To examine the effects our probability page might be having on outcomes, we randomly assigned when participants would

[9] The preregistration for this study is available at https://aspredicted.org/5YQ_1Y3

[10] Our previous studies had all included this probability description and quiz, with the idea that this helps with understanding what probabilities are. However, this page was initially developed for a context where participants were asked to provide probabilities without any examples of what that looks like. In the current context, where they were going to see two probabilities stipulated by us, it might not have been necessary since they would have examples. Also, with the next study having verbal probability cues, we thought it might be helpful to skip the heavily numerical introduction, so it would be useful to first know whether we compel more conjunction math before potentially removing the quiz.

[11] We had more completed surveys than Prolific recruits, likely due to timing-out issues. We did not look for or remove the “extra” takers. We did use the same process previously described to remove completed surveys who shared an IP with another partially completed survey, to exclude participants who had potentially seen multiple versions of the survey, or completed it multiple times. All data removals were performed before any data analysis.

see the probability page in one of two between-subjects conditions: *before* versus *after* the 12 case judgment pages.

After completing generic consent and introductory information, all participants next encountered a page about independence. Since we emphasized independence between the propositions for everyone, we gave all participants this page with background information, illustrating the difference between independence and dependence with examples. At the bottom of this page was a short quiz on independence to check their understanding. Participants saw four pairs of items and were asked to choose for each pair whether the items were independent or dependent. Participants were then allowed to continue with the survey regardless of their answers. Participants in the “before” condition saw the probability page at this point. The probability page used in this study was similar to the probability quiz used in prior studies. (See Appendix for complete instructions and questions.)

The next three pages all participants saw introduced them to the case judgment task they would be performing. First, there was a “Situation Description” page where they were given high-level summary details about a hypothetical house sale scenario between Bill Buyer and Steve Seller. This was an abstracted version of the case used in prior studies, with Steve repairing a problem with a short-term fix, Bill discovering it later, and Bill taking Steve to court to recover repair costs. (See Appendix for full text of all introductory pages.)

On the next page was “Legal Instructions,” which declared that “there are two elements that need to be proven” and that the Buyer needs to prove that “each of” the elements is “more probably true than not true.” The two propositions were chosen and discussed in a way that tried to make them entirely independent from one another. The two elements were, “First, the

Seller *knowingly* made a false statement of *material* fact" and "Second, the Buyer's damages resulted from the Buyer's *reliance* on the statement." Definitions were provided for the emphasized legal terms.

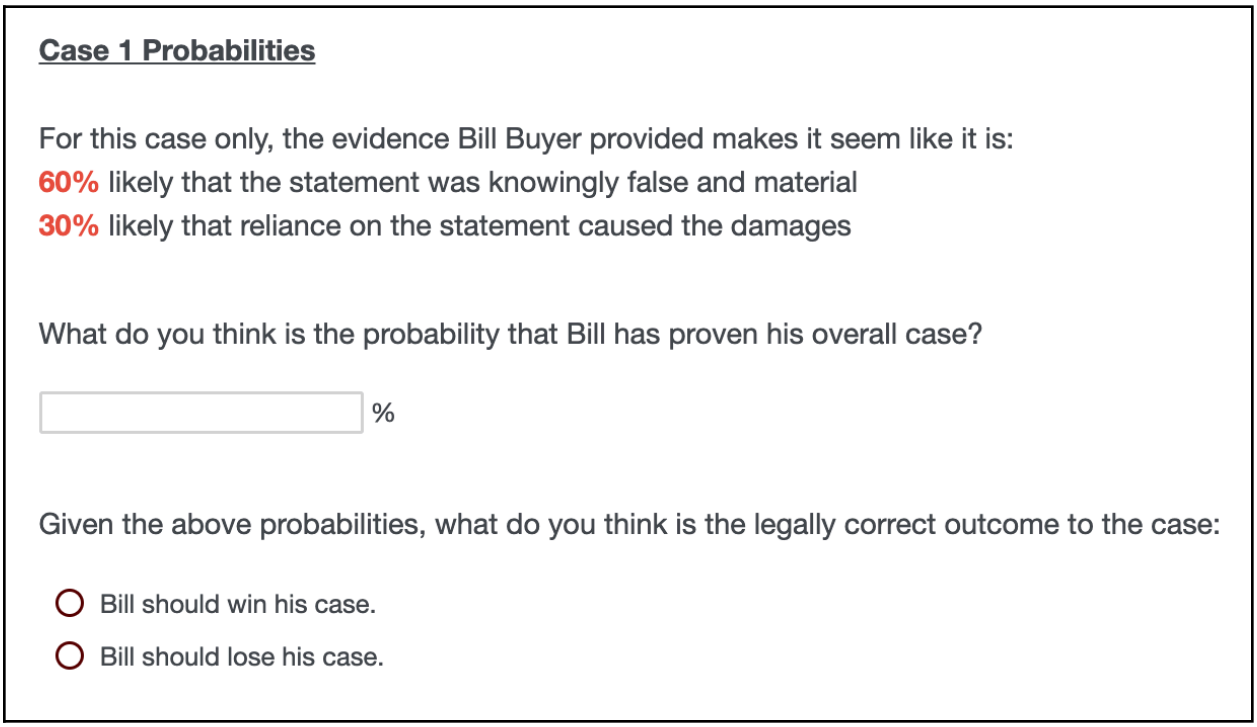
By separating the elements in this way, we hoped to disentangle the legal questions and two element probabilities from one another so that, at least theoretically, any combining of the two probabilities would not need to account for dependence. The final portion of the legal instructions stated, "In order to win, Bill Buyer needs to prove his whole case is more probably true than not true, based on consideration of all of the evidence."

The final introductory page was "Task Information," which discussed the task participants were about to perform. They were told "[t]here are many possible cases of this type" and that they would be considering 12 cases where Bill Buyer "has provided different strengths of evidence in court." We also included a statement that the elements were independent and a reminder about what that means.

Participants then saw 12 case pages with different combinations of probabilities for the first and second elements.[12] At the top of each page were links that participants could click to reread the situation description or task information. The main contents of the page were a case number header (e.g., "Bill's Court Case # 1 of 12"), a repetition of the entire legal instructions, the stipulated case probabilities, and then the case questions. The bottom portion of this page, as displayed to participants, is in Figure 2 below.

[12] As previously mentioned, the order of the 12 cases was fully randomized within subjects.

Figure 2. Study 3 Stipulated Probabilities and Questions



The numeric probability of the overall case question was always first, which limited their input to numbers 0 through 100 (with no decimals), and then the binary overall case win or lose question was second.[13] After viewing the 12 case pages, participants saw a page with an independence check question. The participants in the *after* group saw the probability page next. Finally, the last page of the survey for everyone had a question about the difficulty of the components of the task they had completed and a variety of demographic questions.

[13] Prior studies showed statistically equivalent results for having a binary or probability question first. (See "Order Effects" in Appendix for Study 1 Additional Analyses.) For this study, we decided to have everyone see probability first. There is arguably not actually one "legally correct" answer to either overall case question, but it is most out of place in the context of a question asking about the numeric probability. So, this ordering allowed us to ask about the "legally correct outcome" in the context of "the above probabilities." There is potential ambiguity about whether this includes both the probabilities we provided and the one the participant just answered. This ambiguity exists in actual cases, however, and it is part of the basis of the underlying confusion and debate, so it seems especially appropriate to allow that to exist.

Based on pilot testing, we identified specific strategies that people might adopt or anchor on. These included an average of the two probabilities (avg), multiplying the two probabilities for the statistical conjunctive answer (conj), the first element's probability (1st), the second element's probability (2nd), the highest probability presented (max), and the lowest probability presented (min). The 12 combinations were chosen to provide a mix of results to distinguish between participants using these strategies. The probability combinations used are in Table 2 below.

Table 2. Study 3 Stipulated Element Probabilities for 12 Cases

	key	*description*	1st vs 2nd	1st	2nd	conj	avg	min*	max
1	*A*	no elements win	1st < 2nd	***20%***	**40%**	8%	30%	20%	40%
2	*B*	one element wins, average loses	1st < 2nd	***20%***	**60%**	12%	40%	20%	60%
3	*B*		1st > 2nd	***60%***	**30%**	18%	45%	30%	60%
4	*C*	one element wins, average wins	1st > 2nd	***70%***	**40%**	28%	55%	40%	70%
5	*C*		1st < 2nd	***40%***	***90%***	36%	65%	40%	90%
6	*D*	[Conjunction Divergence] elements & average wins, conjunction loses	1st = 2nd	***60%***	***60%***	36%	60%	60%	60%
7	*D*		1st < 2nd	***60%***	***70%***	42%	65%	60%	70%
8	*D*		1st > 2nd	***80%***	***60%***	48%	70%	60%	80%
9	*D*		1st = 2nd	***70%***	***70%***	49%	70%	70%	70%
10	*E*	elements & average wins, conjunction wins	1st > 2nd	***90%***	***70%***	63%	80%	70%	90%
11	*E*		1st < 2nd	***80%***	***90%***	72%	85%	80%	90%
12	*E*		1st = 2nd	***90%***	***90%***	81%	90%	90%	90%

1st/2nd = the two probabilities we stipulated for the the first and second elements, conj = the (conjunction) product of the two probabilities, avg = the average of the two probabilities, min = the minimum of the two probabilities, max = the maximum of the two probabilities, gray boxes = plaintiff wins

Results

The analyses for this study will start by categorizing participants into combination strategies based on their overall case probability answers. We first describe our preregistered method of dividing people into categories, then the results of applying this method to the current data. The next section examines overall case win or lose answers and how these match a participant's own answers or the elements. Finally, we finish the results with an analysis of conjunction divergence case outcomes.[14]

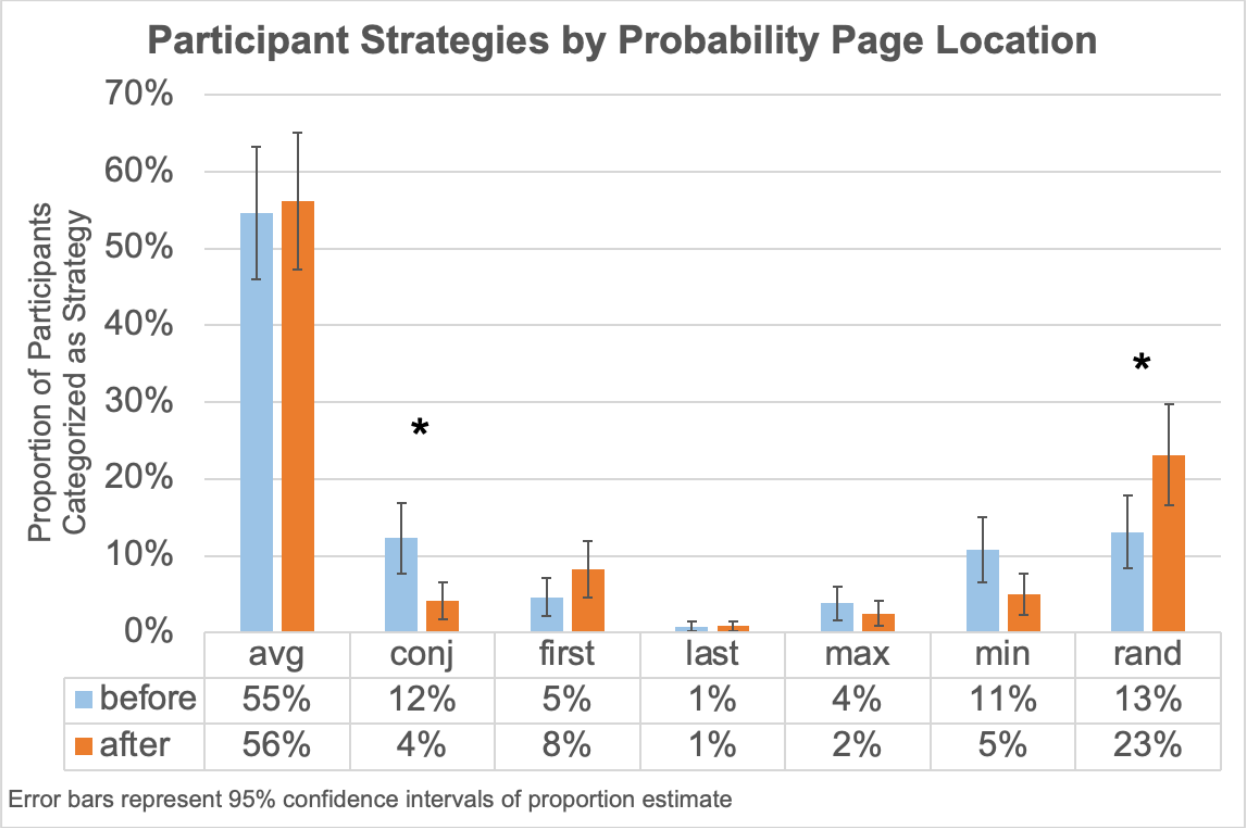
Combination Strategies by Condition

In this study, the goal was to systematically describe individual approaches to the problem of combining element probabilities. We preregistered our plan for grouping participants into strategies based on the method from pilot testing that seemed to most consistently categorize participants into clear behavior types. Our criteria was to find the lowest average absolute difference to a specific strategy. For each participant, we reviewed each of their 12 cases and calculated an average absolute difference to each strategy. For example, suppose a participant gave 64% as their answer for the seventh case, which has 60% for the first element and 70% for the second element. For that case, they would be 22 points away from conjunction, 1 point away from average, 4 points away from both first and minimum, and 6 points away from both last and maximum. These absolute differences are calculated for each strategy across all 12 cases and then averaged. Once we have the average absolute difference to each strategy, we find the minimum difference and categorize the person as utilizing that strategy. Based on the pilot testing, we also preregistered that if the minimum average absolute difference was

[14] The conjunction divergence frequency was held constant in this study across participants because we defined the element numeric probabilities.

greater than 10, the participant would be classified as random (rand) instead of one of the six main strategies. The categorizations for this study are in Figure 3 below. These are separated by the condition of seeing the probability background and quiz page *before* the cases or *after* the cases.

Figure 3. Study 3 Participant Strategy Categorizations by Condition



* indicates a difference with p < .05

We anticipated that we might find an increase in the number of people using conjunctive multiplication in the group that sees the probability page before doing the cases task, and this is what we found, $\chi 2(1, N=251)=5.46$, p=.019. The only other significant difference between proportions was one we did not predict, and that was a decrease in the number of people

categorized as random in the group that saw the probability page earlier, $\chi 2(1, N=251)=4.31$, p=.038. It seems logical that the probability instructions and quiz would encourage more precise thinking about probabilities.

Another factor we look at with categorization is how far away participants are from their best fitting strategy. In other words, we can compare the magnitude of the best average absolute difference to any of the strategies. This magnitude offers another view of closeness to fitting strategies, and the higher values for people who saw the probability page after the cases is more evidence supporting the observation that seeing the probability page makes people slightly less random. (See Table 3.)

Table 3. Study 3 Average Absolute Difference (AAD) by Probability Page Location

	number in group		AAD mean (SD)		AAD mean comparison
Smallest AAD	***before* condition**	***after* condition**	***before* condition**	***after* condition**	**t-test (*before* condition vs. *after* condition)**
avg	71	68	2.77 (2.66)	3.65 (2.60)	t(137)=1.98, p=.049
conj	16	5	1.78 (2.70)	2.92 (3.55)	t(5.5)=0.66, p=.537
first	6	10	1.15 (1.79)	5.73 (3.23)	t(14)=3.65, p=.003
last	1	1	1.67 (NA)	2.50 (NA)	no test
max	5	3	7.17 (1.54)	7.94 (2.18)	t(3.2)=0.54, p=.622
min	14	6	2.94 (3.25)	6.25 (3.25)	t(9.5)=2.09, p=.065
rand	17	28	17.75 (7.79)	16.80 (7.51)	t(32.9)=0.4, p=.689
all w/rand	**130**	**121**	**4.71 (6.35)**	**7.06 (6.93)**	**t(242.9)=2.8, p=.006**
w/o rand	**113**	**93**	**2.75 (2.83)**	**4.13 (2.91)**	**t(194.2)=3.43, p=.001**

Case Win Matching

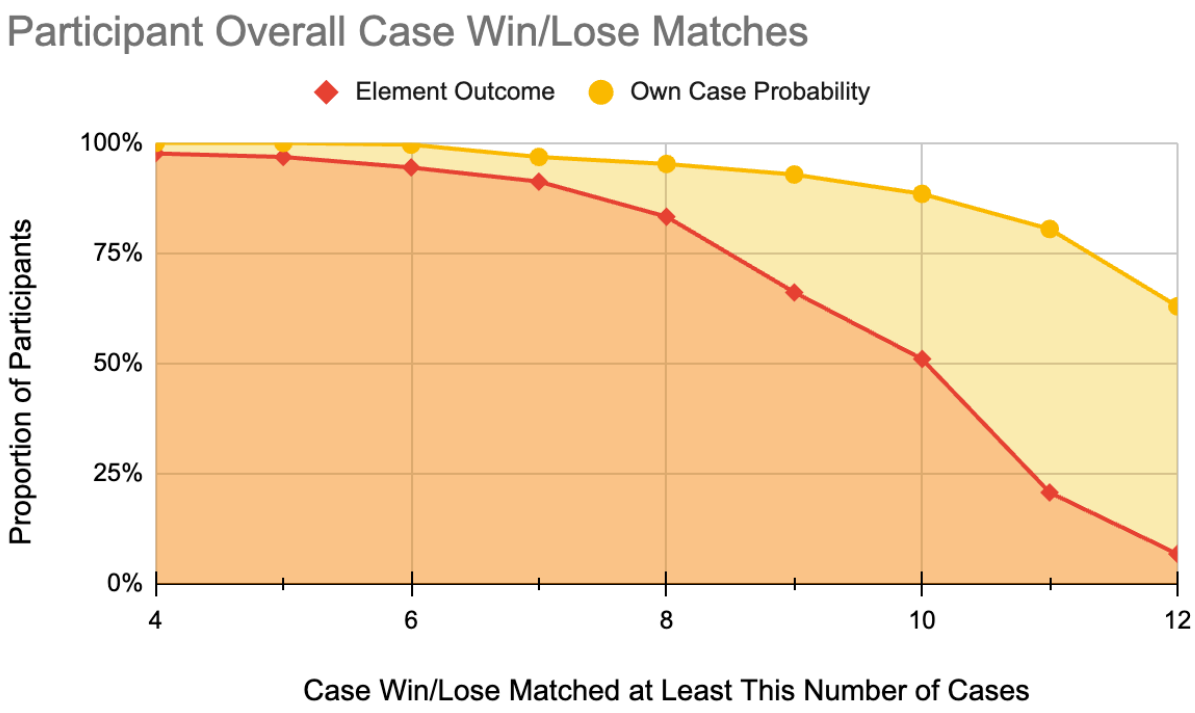
Another factor to consider is how participants answer the overall case binary question about whether the plaintiff should win or lose his case. This question is asked immediately after

the participants give a numeric probability for the plaintiff proving his overall case, so using that number to determine whether the plaintiff should win or lose is one obvious option. To examine this, we look at how often a participant's win or lose choice matches their own probability with a midpoint cutoff. Since we force round numbers, that should technically mean probabilities 51 and higher win, but since the 50 mark is somewhat ambiguous, any 50s are considered a match whether the plaintiff wins or loses.

Since participants receive ambiguous instructions, another option would be to choose an overall win only if both elements win and a loss if any element loses. We will examine this using the stipulated numeric probabilities to determine whether the elements win. We purposefully did not include any 50s in our stipulated probabilities, so this will be a direct mapping of wins for greater than 50 and losses otherwise.

For this final choice, most people seem to be following the inputs they give on the same page most frequently. Out of a maximum of 12 possible matches, participants' case win choices match their own overall case probability frequently ($M=11.2$, $SD=1.5$) and much more frequently than matching stipulated elements winning ($M=9.1$, $SD=2.0$), $t(462.6)=13.48$, $p<.001$. Figure 4 below shows how participants matched both of these outcomes.

Figure 4. Study 3 Overall Case Outcome Matches



This figure depicts the number of matches between an overall case win or lose choice and another input. When looking at numerical probabilities, a 50 is neutral and matches either a win or loss. Thus, an overall case win is matched by a probability of 50 or higher or by two elements that both win. An overall loss is matched by a 50 or lower or by at least one losing element.

Conjunction Divergence

Next, we will look at the cases that were of particular interest to us, the “D” key cases from Table 2. These four “D” cases exhibit conjunction divergence—when each element is rated as higher than the midpoint cutoff, but their product is below the midpoint. Table 4 below indicates how often participants chose case wins that were consistent with considering elements winning (elems) versus case losses that were consistent with the conjunctive product of the elements losing (conj).

Table 4. Study 3 Conjunction Divergence Case Outcomes

Matches:	elems: 0 conj: 4	elems: 1 conj: 3	elems: 2 conj: 2	elems: 3 conj: 1	elems: 4 conj: 0	n
before condition	13	6	9	16	86	130
	10%	5%	7%	12%	66%	
after condition	10	3	7	17	84	121
	8%	2%	6%	14%	69%	
all	23	9	16	33	170	251
	9%	4%	6%	13%	68%	

This table shows which strategy participants matched with their case win or loss decision, when there were different answers between a conjunction strategy and a by-element strategy.
elems = overall cases winning, consistent with a strategy of checking if each element wins
conj = overall cases losing, consistent with a strategy of checking if the conjunctive product of both elements wins

No significant differences were observed in these proportions across conditions, so we will discuss these results in terms of all the participants. In this study, 90.8% of the 251 participants chose a case outcome matching the elements winning (instead of conjunction) at least once. There is still a majority (68.0%) siding with elements, even considering solely the participants who did so every single time they saw a case with conjunction divergence.

Study 3 Summary

This study used numerical probabilities that we provided. The objective was to confirm or adjust the taxonomy of evidence combination strategies developed from a pilot study. We found similar patterns in this study, with people using a variety of combination strategies, and the majority seeming to average across the two elements. A secondary goal was determining whether some portion of individuals without the probability page will spontaneously develop the multiplicative conjunction calculation. While we did find some people still using multiplicative conjunction who had not seen the probability page before the cases, it was significantly smaller

than the portion of people who did see the probability page first. Without a prior prediction, we observe that people seem to behave more randomly when they had not first seen the probability page, as indicated both by the number of people categorized as random, as well as higher distances away from the strategies, even among followers of strategies. Finally, we observed that most people side with the elements rather than use conjunction multiplication when they consider two elements where conjunction divergence exists.

Study 4: Verbal Probability Prompts

The primary goal of Study 4 was to look at user strategies by observing behavior across multiple cases (like Study 3), with participants providing their own numeric probabilities for each element (like studies 1 and 2). To accomplish this, in Study 4, we provided participants with verbal probabilities for elements, and we attempted to get a variety of numbers by varying the verbal probability strength across multiple cases. This study also explored whether explicitly considering the elements impacts overall case results.

Methods

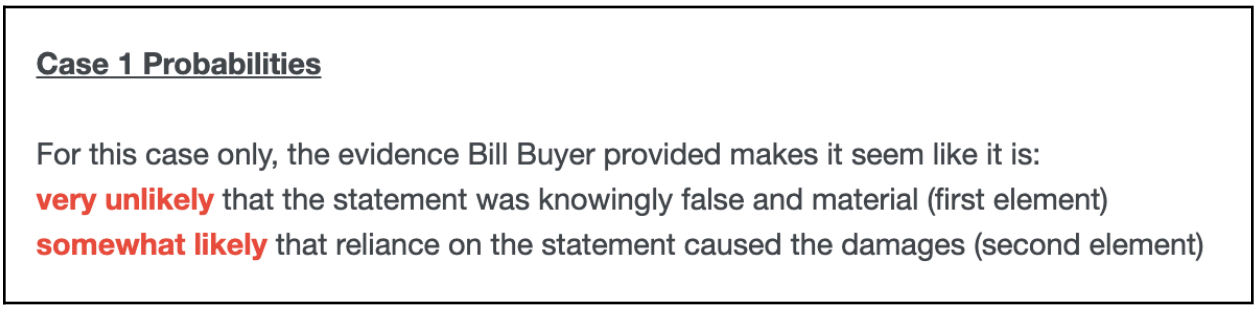
We recruited 400 participants on Prolific, and they were paid $3.20 to complete an online survey.[15] Our final data set includes 399 participants.[16] Participants in this study saw 12 case pages in an order that was fully randomized within subjects. To examine element consideration, we varied how participants were asked to respond about each of the case elements. Participants were randomly assigned to one of three between-subjects conditions: *none*, *binary*, or *numeric* questions about the elements.

Study 4 used a similar structure to Study 3 for the overall flow, but all participants saw both the background independence and background probability information and quiz pages. The general situation description and legal instructions were the same as in Study 3. We provided the full range of verbal probability options in the task information before the first case.[17]

[15] The preregistration for this study is available at https://aspredicted.org/F46_75W
[16] We used the same process previously described to remove completed surveys who shared an IP with another partially completed survey, to exclude participants who had potentially seen multiple versions of the survey or completed it multiple times. All data removals were performed before any data analysis.
[17] “For each case, we will tell you the probability that each element is true. Probabilities are provided on this scale: very unlikely, unlikely, somewhat unlikely, somewhat likely, likely, very likely” (See Appendix for full text)

The top of each of the 12 case pages was the same as in Study 3. There were links to display Situation Description or Task Information again, then a "Bill's Court Case # X of 12" followed by the plaintiff and defendant list and the full legal instructions. The "X% probable" for each element was replaced with a verbal probability descriptor, as seen in Figure 5 below.

Figure 5. Study 4 Stipulated Probabilities Example



The verbal probabilities options were a scale from very unlikely to very likely.[18] To choose specific verbal combinations, we began with a mix of desired case types from Study 3, removing one of the cases where both elements were the same and adding another case where we might be better able to examine potential element differences.[19] The 12 pairs of verbal probabilities used are in Table 5 below.

[18] We also considered using a "more probably true than not true" scale for consistency with the legal instructions, but ultimately decided to go with the simplicity and familiarity of "likely" statements, along with a better sense of how these might be mapped based on prior research.
[19] For this example, and within each key set, we balanced cases where the first element or second was the most likely, to better distinguish between participants choosing to follow one of the two elements.

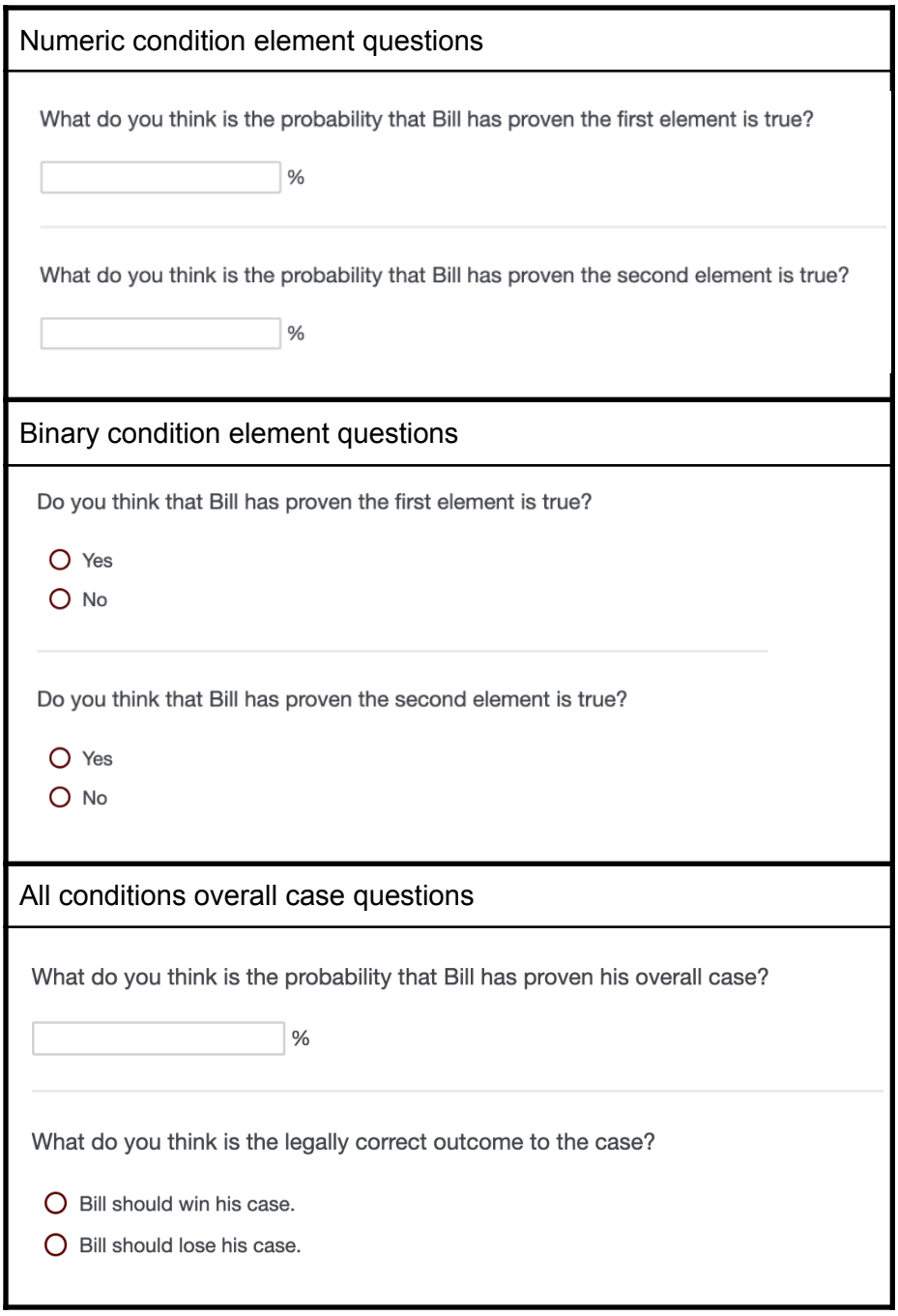
Table 5. Study 4 Stipulated Verbal Probabilities for 12 Cases

First Element	Second Element	Targeted Key Case Types		
somewhat unlikely	very unlikely	1st > 2nd	*A*	no elements win
very unlikely	somewhat likely	1st < 2nd	*B*	one element wins, average loses
somewhat likely	unlikely	1st > 2nd	*B*	
likely	somewhat unlikely	1st > 2nd	*C*	one element wins, average wins
somewhat unlikely	likely	1st < 2nd	*C*	
somewhat unlikely	very likely	1st < 2nd	*C*	
somewhat likely	somewhat likely	1st = 2nd	*D*	[Conjunction Divergence] elements & average wins, conjunction loses
somewhat likely	likely	1st < 2nd	*D*	
likely	somewhat likely	1st > 2nd	*D*	
likely	likely	1st = 2nd	*E*	elements & average wins, conjunction wins
very likely	likely	1st > 2nd	*E*	
likely	very likely	1st < 2nd	*E*	

Targeted Key Case Types were a range of results we were aiming for with the verbal probability pairs. We used Lichtenstein & Newman's (1967) mean response findings to guide the choice of phrases. See Appendix for details.

The initial questions on the case page were determined based on a participant's assigned condition. The *numeric* group's first task was to assign a numeric probability to each element. The *binary* group first assigned a binary choice about being proven for each element. The *none* group did not provide any responses about each individual element. Across all three groups, each case page ended with participants assigning a numeric probability for the overall case and then a binary win/lose for the overall case. See Figure 6 for an illustration of the questions as participants saw them.

Figure 6. Study 4 Participant Questions by Condition



After completing the 12 case pages, all participants provided a numeric probability rating for each of the six verbal probability descriptors. This mapping question was necessary to

assign categories for the *binary* and *none* conditions since they do not provide individual element probabilities. All participants then answered the questions about dependence and perceptions of parts of the study that were also used in Study 3. The final page contained optional demographic questions.

Results

The analyses for this study will be analogous to the results from Study 3, categorizing participants into combination strategies, examining matches of case win or lose decisions, and looking at conjunction divergence frequency and those case outcomes. Since users provided their own numeric probability answers for element probabilities, sometimes in multiple places, the analysis here will be slightly more complex. When results depend on element numeric probabilities, we will consider the *numeric* group first, using the probabilities participants entered separately on each case page, and then we will consider all three conditions, using the probabilities participants entered at the end of the task on the verbal probability mapping page.

Combination Strategies by Condition

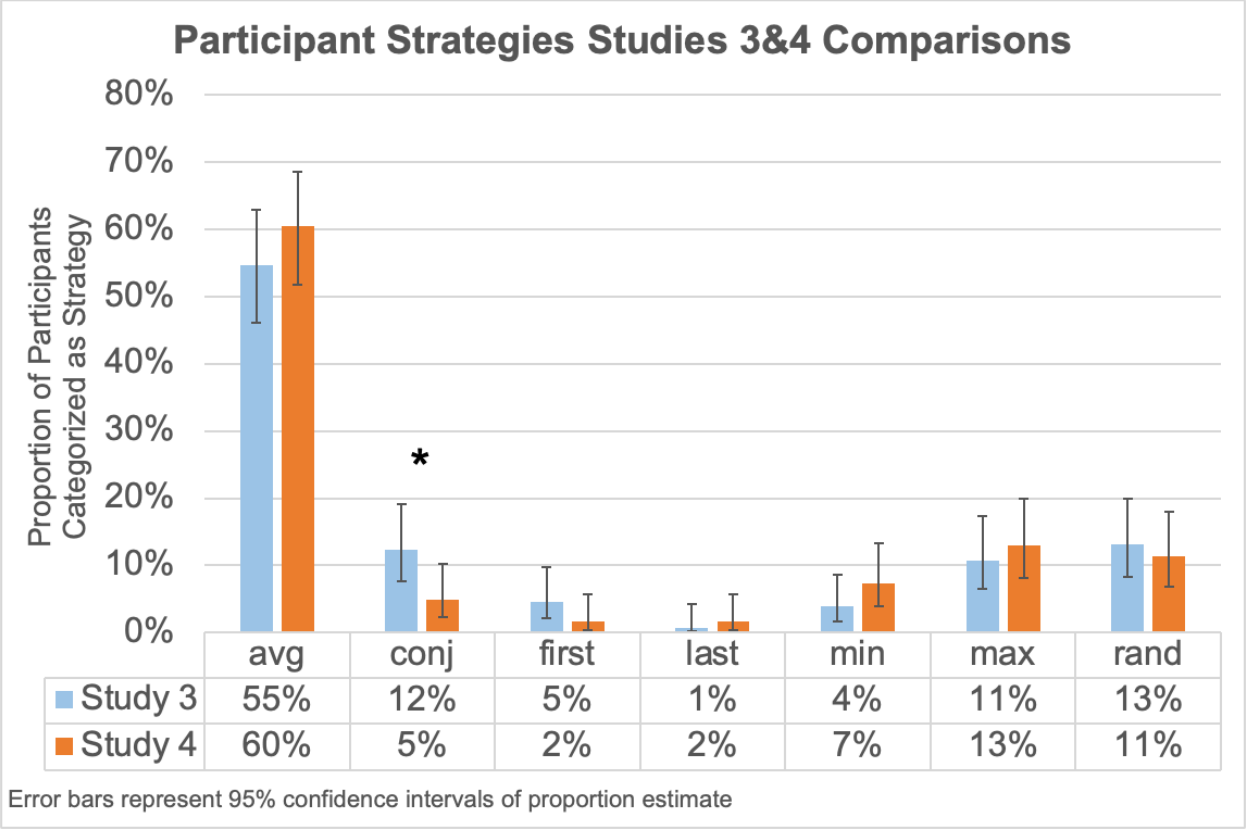
To determine how a participant is categorized, we again look to minimize the average absolute difference from a given strategy, but with this study, the strategies are based on the participants' own element numbers rather than the ones we provide. For each participant, for each of their 12 cases, we started with the numeric probabilities for the elements that the participant supplied (either the two-element probabilities for that same case for the *numeric* group or the participant's matching mapped probabilities from the end of the survey). Then, again, for each case, we constructed each strategy's answer using those element probabilities and compared the participant's overall case answer to each of these strategy answers. For

example, suppose a participant gave 70% for the first element and 90% for the second element. If they gave an overall case probability of 65%, then for that case, they would be 2 points away from conjunction, 16 points away from average, 6 points away from both first and minimum, and 26 points away from both last and maximum. This absolute difference is calculated for each strategy across all 12 cases and then averaged.

The *numeric* group was our primary group of interest in examining combination strategies because participants provided their own numeric probabilities for elements on each page, along with the overall case probability. This is the closest analogy to Study 3, where we first examined combination strategies. Figure 7 and Table 6 below show the categorization for the *numeric* condition in this study and the *before* condition in Study 3.[20]

[20] Everyone in Study 4 saw the probability background information and quiz before the case judgments, so the Study 3 condition where they also saw this information *before* is most directly comparable.

Figure 7. Study 3 versus Study 4 Participant Strategy Categorization



* indicates a difference with p < .05 Study 3 *before* condition and Study 4 *numeric* condition. All participants saw a probability page before the case task. On the case pages, Study 3 participants saw stipulated numeric probabilities for elements, and Study 4 participants saw stipulated verbal probabilities for elements and provided numeric probabilities for elements on the same page.

Table 6. Study 4 versus Study 3 Strategies

	n		Average Absolute Distance to Strategy		
			mean (SD)		
	S3	S4	Study 3	Study 4	mean comparison
avg	71	75	2.77 (2.66)	4.05 (2.70)	t(143.8)=2.89, p=.004
conj	16	6	1.78 (2.70)	2.20 (2.87)	t(8.5)=0.31, p=.764
first	6	2	1.15 (1.79)	6.67 (0.00)	t(5)=7.56, p=.001
last	1	2	1.67 (NA)	4.38 (0.88)	no test
max	5	9	7.17 (1.54)	7.09 (2.57)	t(11.8)=0.07, p=.947
min	14	16	2.94 (3.25)	5.21 (2.94)	t(26.5)=2, p=.056
rand	17	14	17.75 (7.79)	13.52 (3.16)	t(21.9)=2.05, p=.053
all w/rand	130	124	4.71 (6.35)	5.45 (4.08)	t(221.4)=1.11, p=.268
w/o rand	113	110	2.75 (2.83)	4.42 (2.86)	t(220.7)=4.4, p<.001

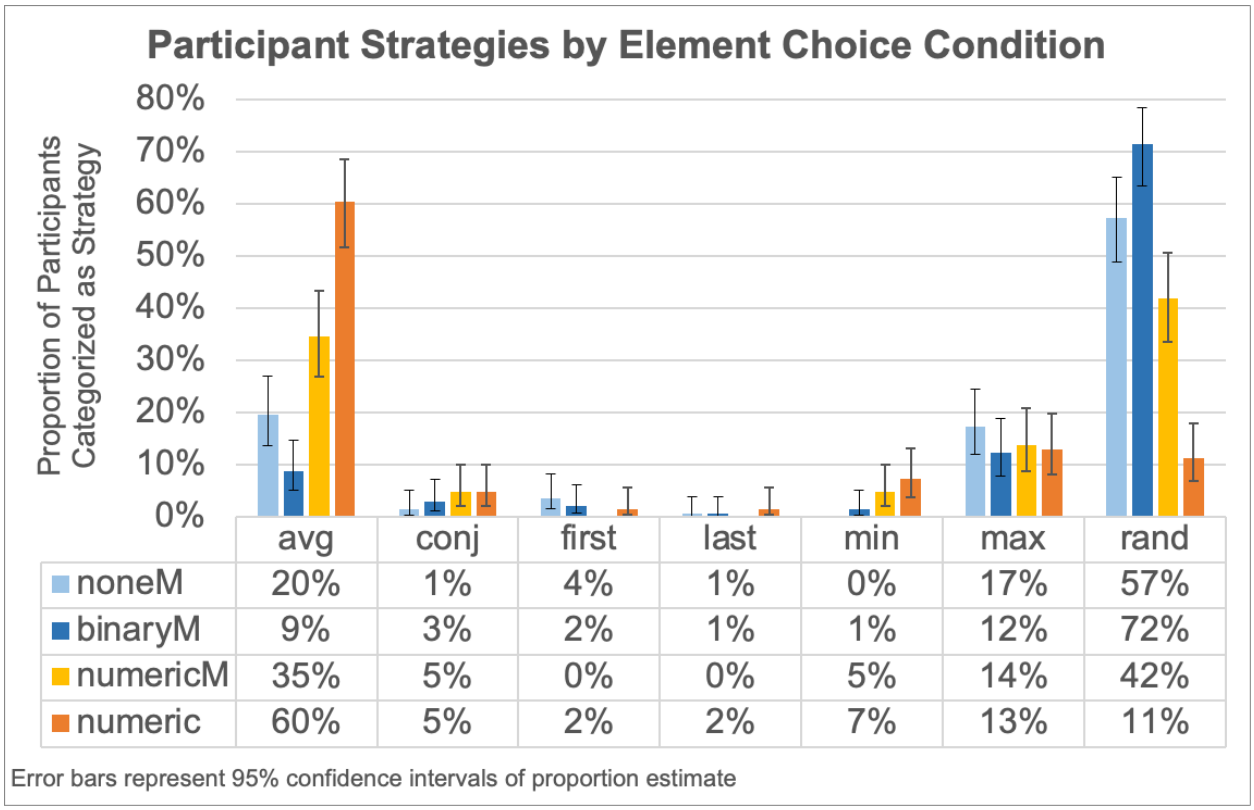
Study 3 *before* condition and Study 4 *numeric* condition.

This compares conditions across studies which means there was no random assignment into these conditions. Nonetheless, we can note the patterns observed without making causal claims. The overall pattern is very similar within these conditions from the two studies. Fewer participants spontaneously used a conjunctive math strategy (Study 3 *before* condition: 12%; Study 4 *numeric* condition: 5%), $\chi 2=4.48$, p=.034. Another observation is that participants appear to be further away from their strategies in this study than in Study 3, as seen by the average distance away from the best strategy for multiple individual categories in Table 6 above. Considering the full set of participants who were categorized into any strategy (not random), participants were closer to their strategies in Study 3 (M=2.75, SD=2.83) than they were in Study 4 (M=2.75, SD=4.42), t(220.7)=4.4, p<.001.

We can also look at strategies for the *binary* and *none* groups, but these necessarily need to use the results of the probability mapping task the participants completed only after

seeing all 12 cases. Since this is a difference in methodology, when comparing these other two conditions, we will also look at how the *numeric* group appears to behave if we similarly just use the results from their probability masking task. Figure 8 below shows categorizations using the numeric condition element probabilities from both sources.

Figure 8. Study 4 Participant Strategy Categorizations by Condition



	avg	conj	first	last	min	max	rand
noneM	20%	1%	4%	1%	0%	17%	57%
binaryM	9%	3%	2%	1%	1%	12%	72%
numericM	35%	5%	0%	0%	5%	14%	42%
numeric	60%	5%	2%	2%	7%	13%	11%

The “M” designation after condition names means these strategies were determined based on the mapped verbal probabilities provided at the end of the study after all 12 cases had been completed.

Here we are interested in how random participants are. Fewer people in the *numeric* condition are categorized as random when looking at probabilities they provided on each page compared to the probabilities they provided at the end of the study (*numeric* case page probabilities: 11%; *numeric* mapped probabilities: 42%), $\chi 2(1, N=248)=29.81$, $p<.001$. However,

there are still significantly fewer people categorized as random in *numeric* using mapped probabilities than in either the *none* (57%), χ2(1, N=262)=6.12, p=.013 or *binary* (72%), χ2(1, N=261)=23.33, p<.001, conditions. There is also a significant difference between these last two, with people in the *binary* group being categorized as random more frequently than the *none* group, χ2(1, N=275)=6.12, p=.013.[21]

Case Win Matching

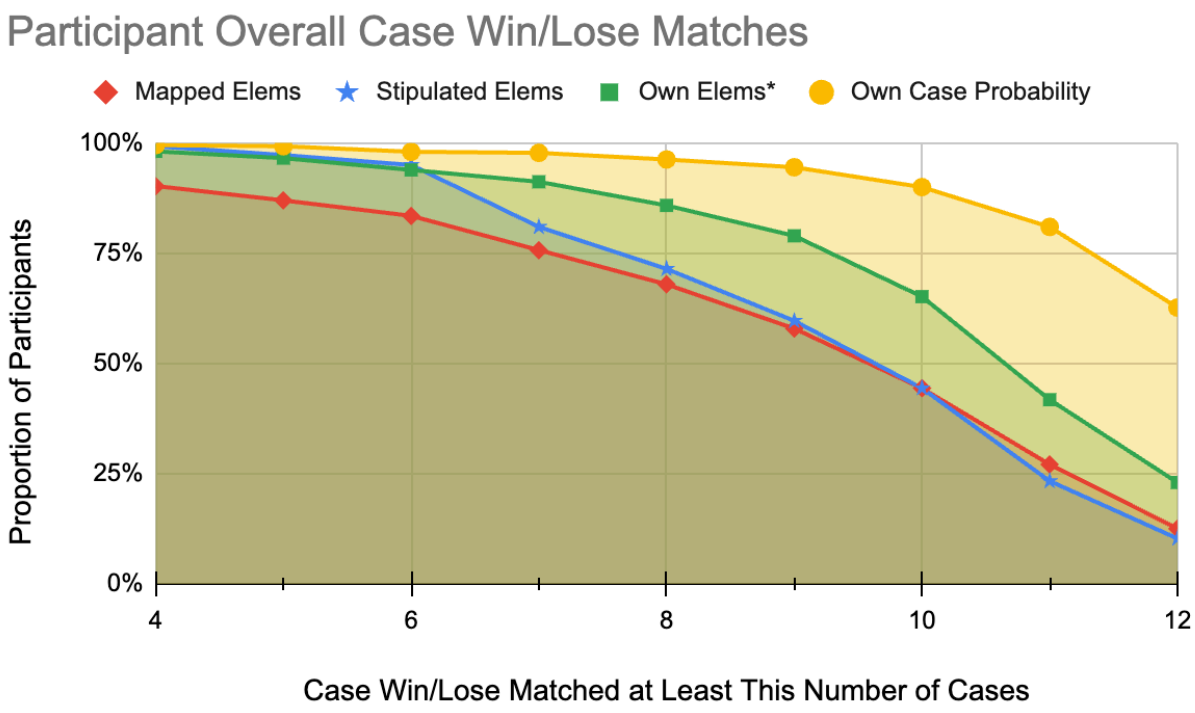
As in Study 3, we will again consider how participants answer the overall case binary question about whether the plaintiff should win or lose his case. We first look at how often participants' winning or losing choices match their own probability, with a midpoint cutoff.[22]

Since participants receive ambiguous instructions, another option would be to choose an overall win only if both elements win and a loss if any element loses. We will examine this option using a few different measures to determine whether elements win. For everyone, we can check on whether the verbal probability stipulated elements are on the “likely” side and whether a participant's mapped numeric probabilities for the verbal prompts are above the midpoint. For the participants in conditions that made individual element choices, we can also look at those. We can see whether the numeric group’s probabilities are above the midpoint and whether the binary group chose that the elements were proven true.

[21] As a robustness check, this hierarchy mostly holds true if we use a higher cutoff for randomness, such as 15. The least random is numeric case page probabilities (3%) as compared with all other conditions including numeric mapped probabilities: (27%), χ2(1, N=248)=27.97, p<.001. The most random is binary (42%) as compared with all other conditions including none (30%), χ2(1, N=275)=4.21, p=.040. However, the difference between numeric mapped and none is no longer significant, χ2(1, N=262)=0.29, p=.591.
[22] As noted in Study 3, since we force round numbers, that should technically mean probabilities 51 and higher win, but since the 50 mark is somewhat ambiguous, any 50s will be considered a match whether the plaintiff chooses win or lose.

For the final choice, most people seem to be following the inputs they give on the same page most frequently. Out of a maximum of 12 possible matches, participants case win choices match their own overall case probability frequently (M=11.2, SD=1.6) and much more frequently than any of the element options, including stipulated elements being likely (M=8.8, SD=2.1), $t(725.4)=17.88$, $p<.001$, mapped probabilities winning (M=8.3, SD=3.1), $t(586.4)=16.54$, $p<.001$, and their own element choices winning (M=9.7, SD=2.2), $t(425.2)=9.25$, $p<.001$. The other options form a hierarchy of preferences that are significant as well. Participants match choices to their own elements more frequently than stipulated elements, $t(543)=5.23$, $p<.001$, and mapped probabilities, $t(652.7)=6.8$, $p<.001$, and they match the stipulated elements more than the mapped probabilities, $t(708.5)=2.63$, $p=.009$. Figure 9 below shows this relationship at each cutoff for matches.

Figure 9. Study 4 Overall Case Outcome Matches



This figure depicts the number of matches between an overall case win or lose choice and another input. When looking at numerical probabilities, a 50 is neutral and matches either a win or loss. Thus, an overall case win is matched by a probability of 50 or higher or by two elements that both win. An overall loss is matched by a 50 or lower or by at least one losing element. *Own Elems includes the 261 participants in the *binary* and *numeric* conditions. The participants in the *none* condition did not make individual element decisions.

This hierarchy of matching preference makes cognitive sense in terms of distance from each of the other factors. The most recent number typed is the most influential, followed by decisions made immediately before that, prompts read immediately before that, and then a mapped probability that probably wasn't in mind until being asked this specifically at the end of the survey.

Conjunction Divergence

Next, we again look at instances where there is conjunction divergence—when each element is rated as higher than the midpoint cutoff, but their product is below the midpoint—to

see how frequently the conjunction problem could occur and also how frequently it does occur. For the purposes of this analysis, the cutoff will be the stricter legal one, where only numbers above 50 (not 50 itself) count as winning. Since the numeric task had a different methodology for eliciting element probabilities, results will be reported using two data sets—the numeric condition participants using their entered element probabilities on each page and all participants using only their mapped probabilities.

Tables 7 and 8 below indicate how often participants had conjunction divergence (CD), and from within those CD cases, how often they chose case wins that were consistent with considering elements (elems) winning versus case losses which were consistent with the conjunctive product (conj) of the elements losing.

Table 7. Study 4 Conjunction Divergence Cases and Outcomes, *Numeric* Condition

	1 CD	2 CDs	3 CDs	4 CDs	5 CDs	6 CDs	7 CDs	8 CDs	Sum
0 conj (all elems)	11	22	27	12	4	1	1	1	79
1 conj (rest elems)	2	1	1	3	1	0	0	0	8
2 conj (rest elems)		4	0	2	0	0	0	0	6
4 conj (rest elems)				1	0	0	0	0	1
Sum	13	27	28	18	5	1	1	1	94

This table shows which strategy participants matched with their case win or loss decision when there were different answers between a conjunction strategy and a by-element strategy.
elems = overall cases winning, consistent with a strategy of checking if each element wins
conj = overall cases losing, consistent with a strategy of checking if the conjunctive product of both elements wins
gray boxes = every conjunction divergence (CD) case was consistent with a conjunctive product strategy

Looking at the numeric group's case element probabilities from each individual case page is the most similar use case to the prior studies. Conjunction divergence occurred at least once for 75.8% of the 124 participants in the numeric condition when using the case page probabilities. Within the 94 numeric condition participants with conjunction divergence, only 7.4% always gave answers consistent with a conjunctive multiplied product strategy. The remaining 92.6% chose a case outcome matching the elements winning at least once. The majority is still large (84.0%), even if we consider only the participants who were consistent with matching elements winning every time they had conjunction divergence in their data.

Table 8. Study 4 Conjunction Divergence Cases and Outcomes Using Mapped Probabilities

	1 CD	2 CDs	3 CDs	4 CDs	6 CDs	Sum
0 conj (all elems)	57	1	105	29	3	195
1 conj (rest elems)	13	1	23	10	2	49
2 conj (rest elems)		0	6	2	2	10
3 conj (rest elems)			7	4	0	11
4 conj (rest elems)				2	0	2
5 conj (rest elems)					1	1
6 conj (rest elems)					1	1
Sum	70	2	141	47	9	269

This table shows which strategy participants matched with their case win or loss decision, when there were different answers between a conjunction strategy and a by-element strategy.
elems = overall cases winning, consistent with a strategy of checking if each element wins
conj = overall cases losing, consistent with a strategy of checking if the conjunctive product of both elements wins
gray boxes = every conjunction divergence (CD) case was consistent with a conjunctive product strategy

The results for all participants using the mapped verbal probabilities at the end of the study are very similar to what we saw looking at the numeric group. Conjunction divergence occurred at least once for 67.4% of the 399 participants when using mapped probabilities. Within the 269 participants with conjunction divergence, 91.4% chose a case outcome matching the elements winning at least once. The majority is still large (72.5%), even if we just consider the participants who matched elements winning every time. Every condition displayed this pattern, and there were no statistically significant differences between them in the rate of siding with elements in cases with conjunction divergence.[23]

[23] Always sided with elements within Conjunction Divergence: (numeric: 77%; binary: 67%; none 73%), numeric vs. binary: $\chi 2(1, N=173)=1.82$, p=.177, numeric vs. none: $\chi 2(1, N=186)=0.35$, p=.556, binary vs. none: $\chi 2(1, N=179)=0.63$, p=.426.
Sided with elements at least once within Conjunction Divergence: (numeric: 93%; binary: 88%; none 93%), numeric vs. binary: $\chi 2(1, N=173)=1.49$, p=.222, numeric vs. none: (numeric: 93%; none: 93%), $\chi 2(1, N=186)=0.03$, p=.867, binary vs. none: $\chi 2(1, N=179)=1.17$, p=.279.
Rates of Conjunction Divergence: (numeric: 73%; binary: 61%; none: 73%), numeric vs. binary: $\chi 2(1, N=261)=4.19$, p=.041, numeric vs. none: $\chi 2(1, N=220)=0$, p=.956, binary vs. none: $\chi 2(1, N=233)=3.81$, p=.051. There is one significant result here, with the lower rate of binary people having cases with

Study 4 Summary

This study replicated the design from Study 3 but with non-numerical (and more realistic) verbal descriptions of the per-element strength of evidence. The primary objective was to explore the distribution of participants' heterogeneous combination strategies when case materials were presented verbally, without numbers, that might encourage the adoption of ad hoc calculation rules. Additional objectives were to look at conjunction divergence outcomes with the verbal materials and to look at whether explicitly considering elements shifted behavior.

This study showed a similar pattern of strategy choices as the prior study. Many people seem to be averaging, while small numbers of people use other strategies like conjunction or whether the minimum element wins, and some people behave randomly. Across conditions, there were more participants categorized as random when they did not provide element probabilities on each page. Although one explanation for this might be that it is easier to be consistent in the *numeric* condition because they provide all three numeric probabilities together, this pattern holds even when we ignore the per-case element probabilities and just use the *numeric* participant's mapped probabilities, like the other two conditions. Even with the same temporal distance in providing the numeric information for elements, the *numeric* group retains significantly fewer participants who are random. Participants don't seem to be adopting a consistent strategy relative to a set of numbers they have in mind for each of the verbal probability phrases. The act of thinking about numbers for each element seems to be allowing participants to develop strategies more readily.

---

conjunction divergence. While it is possible this is a real result, this is a relatively high p-value in the middle of a large number of tests. Regardless, any findings about whether a participants' mapped probabilities leave room for conjunction divergence or not are outside of the scope of this paper.

This final study gave us the most opportunities to look at conjunction divergence with participants providing unique numbers they generated themselves. As with the other studies, we saw the same pattern that in conjunction divergence cases, people are generally choosing case wins, corresponding to both elements winning, rather than the losses that would be necessitated by the conjunctive probability answer.

General Discussion

The participants in Study 1 did not seem to combine element probabilities as conjunctions, given that they are insensitive to the two-element versus four-element case manipulation, and their answers differ from the calculated conjunction probabilities. They seem to be attentive to the details provided (given the differences in strong versus weak case evidence conditions), but they are not combining each of the elements in a probabilistic way, even though we ask them for probabilities.

In Study 2, the effect of legal instructions was considered. There were no observed differences among case variables or siding with elements rates, regardless of whether or not participants saw the conjunction problem legal instructions.

Rates of conjunction divergence in these first two studies were relatively rare, with only about 5% of participants giving element probabilities where the outcome of an element-by-element consideration and multiplying the probabilities would result in different answers. Within conjunction divergence cases, the vast majority of the time, participants choose that the case wins rather than loses, which is the answer consistent with an element-by-element consideration.

Study 3 provided more data about the conjunction problem by presenting participants with 12 cases with stipulated numeric probability element ratings, including multiple cases with conjunction divergence. By analyzing participant behaviors over multiple cases, we were able to divide participants into categories representing which strategy was closest to how they were combining element probabilities into one overall answer. While some people spontaneously multiplied elements for their overall case answer, we found that our probability introductory

information and quiz page did encourage more participants to apply this probabilistic process. Encountering the probability page also seemed to make participants behave a little less randomly overall, with fewer participants categorized as random and participants applying their chosen strategies more narrowly.

Study 4 used a similar 12-case design but with stipulated verbal probability element strengths. This allowed participants to experience the elements in a non-numerical way while still providing a variety of cases and an ability to categorize participants into strategies. The participants who immediately converted the verbal probabilities into numbers behaved most similarly to participants from Study 3, though they followed the strategies a little less precisely, with larger differences between their ratings and the answers prescribed by the strategies. When grouping into strategies by using conversions to numbers after the case task, all the participants appeared much more random, but the immediate converters still exhibited more systematic behavior.

Together, the last two studies examined many more cases of conjunction divergence than in Studies 1 and 2. Throughout these cases, the vast majority of participants decide case outcomes as wins, consistent with an element-by-element winning check, instead of losses, which would be consistent with a conjunctive multiplied product of probabilities.

Together, these studies add support to the idea that the conjunction problem might be present in actual juror judgments, and it is likely larger than just being present in those jurisdictions that incorporate the conjunction problem into their instructions. It is still arguable whether the correct answer would be to use probabilistic math in most real-world cases with interdependent sources of evidence and overlapping issues and conditional inferences.

However, in our studies, we minimize these questions with precise instructions and a stipulation that the elements were independent, including an explanation of the concepts of dependence and independence. If our stipulations were to be believed and our instructions were followed faithfully, overall case probability answers should have been much closer to the multiplicative conjunction answers than to averaging. However, the data show a clear tendency to do the opposite. Our win and loss instructions included more ambiguity, so we don't claim that there is a normative correct answer, but these answers also presented a clear pattern with a majority of people following their own probability answer for conjunction divergence cases, which corresponds to an answer based on satisfying individual elements rather than mathematical probabilities.

The present studies also provide a new demonstration where people's intuitions lead them towards averaging, which is the strategy most frequently adopted by our participants. These studies also provide some interesting insights about multi-element strategy development related to how much participants are engaging in consideration of probabilities. Seeing the probability introductory page induced more strategy following, and participants' answers were closer to those in their dominant strategy. Making decisions about element numeric probabilities had the same effects, even when strategies were assigned based on numbers provided only at the end of the survey. Together, this provides some evidence that the act of thinking about probabilities encourages more participants to develop a strategy and/or to follow that strategy more closely.

Future Research

The findings of Study 1 could be expanded upon through additional studies varying the number of elements in multi-element cases. In particular, studying some cases where certain case elements were “stipulated as true” by the lawyers in the case versus experimenters stipulating a probability rating of “absolutely certain” or “100%” would be interesting. It is possible that the extra high probability ratings would shift overall case probability answers higher because of participants using averaging.

Study 2’s examination of legal instructions should be followed up with a higher-powered study before being able to make supported inferences related to the lack of observed differences. In addition, it would be of practical value to test alternative wordings of instructions to identify instructions that increase conformity to the mathematical conjunction principle. It would also be useful to develop and test simple software solutions that could facilitate reasoning consistent with whichever rule, element-by-element versus conjunction, is deemed appropriate for the case and jurisdiction.

The finding that thinking about probabilities leads to following strategies more closely could be studied within other contexts, such as two supposedly “unrelated” studies where the first study involves more intensive experience with numerical probabilities and the second study shifts to making judgments about verbal probabilities, like in Study 4. Future studies could also examine reliability through repeating questions within a larger set of cases.

Finally, studies incorporating more interdependent elements would be more informative about judgments in the real-world setting where this work could have practical and policy

implications and would contribute to the debate about whether the conjunction problem is really a “problem” that needs fixing or not.

Appendix 1: Study Materials

Studies 1&2 Case Information

[case strength = strong is in square brackets] (case strength = weak is in parentheses)

Case Information

Steve was about to put his house on the market for sale. He discovered a major plumbing problem in the house and hired a plumber. The plumber put in a temporary fix, [and](but) he told Steve the fix [would probably last for about one year] (was permanent).

Bill was searching for a home to buy for him and his family. After a week of checking out various properties he looked at Steve's house. Steve's house was in a convenient location and had all the features Bill's family was looking for, and was at the top of the price range that Bill could afford. Bill made an offer on the house that day, which Steve accepted.

Over the next week Bill hired [an excellent, licensed] (a cheap, unlicensed) home inspector to check out the house. There were a few minor repair items that the inspector found that Steve agreed to fix before the sale finalized. The inspector did not find the plumbing problem [, but it would have been very difficult to discover.] (, but saw some suspicious patchwork and warned Bill to be careful with the purchase.)

Meanwhile, Steve filled out [a mandatory] (an optional) "Seller Disclosure" form. [All defects must be disclosed on this form.] Steve stated that the appliances, roof, foundation, plumbing, electricity, and more were all in "excellent" condition. Bill's agent [carefully reviewed the disclosure form with him, and they were confident it was accurate.] (explained the disclosure form was almost always filled out this way, and that it usually included some inaccurate statements.) The house sale finalized a month later.

A week after Bill's family moved into the home they experienced a backing up problem in the bathroom. Bill called a plumber who, after investigating the issue, explained that there were roots from trees growing into the pipes under the house. Furthermore, he said it looked like somebody else had already discovered the problem, because the pipes had a temporary fix applied to them. He said he could redo the temporary fix for

$1000, but that in the next few months they would need to address the problem more permanently. He quoted $15,000 for the full repairs.

Bill let the plumber do the temporary fix then called Steve's agent. Steve called Bill back personally to say he didn't realize the roots would be an ongoing problem. Steve refused to help Bill pay for the cost of repairs.

Bill took Steve to court to recover the repair costs. During the case, a real estate expert testified that the plumbing issue [would have significantly decreased] (would not have affected) the sales price.

The attorneys agree that the above facts are accurate.

Please imagine you a jury member in this case. On the next page are legal instructions.

Study 1 Legal Instructions

Legal Instructions

The lawyers have stipulated that listing the plumbing as "excellent" on the Seller's Disclosure form was a false statement.

To win his case, Bill needs to prove that it is more probably true than not true that Steve committed fraud by making this false statement. The plaintiff needs to prove that each of the following propositions is more probably true than not true:

[number of elements = 4]
First, The false statement was of a material fact. [e1]

Second, Steve knew the statement was false. [e2]

Third, Bill reasonably believed the statement. [e3]

Fourth, Bill's damages resulted from his reliance on the statement. [e4]

Definitions:
Material means that it could have affected the buyer’s purchase decision or the sale price.
Knowing means consciously with understanding of the facts or circumstances.
Reasonably believed means that a reasonable person in the same situation would have believed the statement.
Reliance on a statement means being dependent on the statement, and acting based on the statement.

[number of elements = 2]
First, [e1 or e2]

Second, [e3 or e4]

Definitions:
[definitions corresponding to their respective elements]

Study 2 Legal Instructions

[instructions = conjunction problem is in square brackets]
(instructions = non-conjunction problem is in parentheses)

Legal Instructions

The lawyers have stipulated that listing the plumbing as "excellent" on the Seller's Disclosure form was a false statement.

[To win his case, Bill needs to prove that each of the following propositions is more probably true than not true:] (To win his case, Bill needs to prove that it is more probably true than not true that Steve committed fraud by making this false statement. The plaintiff needs to prove that each of the following propositions is more probably true than not true:)

First, The false statement was of a **material** fact.

Second, Steve **knew** the statement was false.

Third, Bill **reasonably believed** the statement.

Fourth, Bill's damages resulted from his **reliance** on the statement.

[If you find from your consideration of all the evidence that all of these propositions are more probably true than not true, then your verdict should be for Bill.

On the other hand, if you find from your consideration of all the evidence that any of these propositions has not been proved as required in this instruction, then your verdict should be for Steve.]
Definitions:
Material means that it could have affected the buyer’s purchase decision or the sale price.
Knowing means consciously with understanding of the facts or circumstances.
Reasonably believed means that a reasonable person in the same situation would have believed the statement.
Reliance on a statement means being dependent on the statement, and acting based on the statement.

Studies 1&2 Probability Quiz

A probability for purposes of this study means a statement of your belief that a state of the world is true, stated as a percent.

If you were absolutely certain that an event would NOT happen, you would give that event a probability of "0%" On the other hand, if you were absolutely certain that the event would occur, you would say the probability is "100%"

For example, you might state the probability that a fair coin will come up "heads" when tossed is "50%" This would mean you believe it is equally likely that a head or a tail would come up when the coin is tossed. Or you might be asked by a friend, "Will it rain this afternoon?" and you might reply, "I think so, I would say the probability is about 90%" This would mean you think it is very likely that it will rain this afternoon.

Following are a few probability questions. In the 0-100 percent range, give your best answer to each question.

1. Suppose you have a fair six-sided die with the numbers 1, 2, 3, 4, 5, 6 printed on each side of the die. You toss that die ...

What is the probability that the die will land with an even number (2 or 4 or 6) facing upwards?

2. Suppose you toss a fair coin twice, what is the probability it comes up "heads" both times?

3. Ken will get to go out with his friends tonight only if both his mom and dad say yes. There is a 60% chance his mom will say yes and an 80% chance his dad will say yes. What is the probability that Ken will get to go out tonight?

4. What is the approximate probability that it will rain where you live before midnight tonight?

Studies 3&4 Independence and Dependence Page

Background: Independence and Dependence

Legal cases usually involve multiple elements. In some cases, evidence can help prove more than one element, and these elements are dependent on each other. In other cases, the evidence is separate, and these elements are independent from each other.

For example, if you were trying to prove these three things: 1. Sally carried an umbrella today. 2. It rained in Sally's city today. 3. Sally's mother received mail today.

1 & 2 would be considered dependent, because they are likely to happen or not happen together, and some evidence (like "the ground is wet") would help prove both 1 & 2.

1 & 3 would be considered independent, because one happening is unrelated to the other happening, and separate evidence would be needed for each of them.

In other words, if you know it is true that "Sally carried an umbrella today" you might also think it is more likely to be true that "It rained in Sally's city today." But, if you know it is true that "Sally carried an umbrella today" it does not tell you anything about whether "Sally's mother received mail today."

Following are a few examples of pairs of elements. Think about whether these are likely to be dependent or independent and give your best answer to each question.

[all questions have two multiple choice options: O Dependent O Independent]
[order was adjusted between studies, Study 4 ordering is below, Study 3 # is in {curly brackets}]

1 {4}. - The driver is driving faster than the speed limit.
- The driver is late for an appointment. *[answer: Dependent]*

2 {1}. - A tossed six-sided die lands with an even number (2 or 4 or 6) facing upwards.
- It will rain where you live before midnight tonight. *[answer: Independent]*

3 {2}. - The first toss of a fair coin comes up "heads"
- The second toss of a fair coin comes up "heads" *[answer: Independent]*

4 {3}. - Ken's mom gives him permission to go out tonight.
- *[3]* Ken's dad gives him permission to go out tonight. *[arguable, not scored]*
- *[4]* Ken's mom gave him permission to go out last night *[answer: Dependent]*

Studies 3&4 Probability Page

Background: Probability

A numeric probability for purposes of this study means a statement of your belief that a state of the world is true, stated as a percent.

If you were absolutely certain that an event would NOT happen, you would give that event a probability of "0%" On the other hand, if you were absolutely certain that the event would occur, you would say the probability is "100%"

For example, you might state the probability that a fair coin will come up "heads" when tossed is "50%" This would mean you believe it is equally likely that a head or a tail would come up when the coin is tossed. Or you might be asked by a friend, "Will it rain this afternoon?" and you might reply, "I think so, I would say the probability is about 90%" This would mean you think it is really likely that it will rain this afternoon.

Following are a few probability questions. In the 0-100 percent range, give your best answer to each question.

1. Suppose you have a fair six-sided die with the numbers 1, 2, 3, 4, 5, 6 printed on each side of the die. You toss that die ...
What is the probability that the die will land with an even number (2 or 4 or 6) facing upwards?

2. Suppose you toss a fair coin twice, what is the probability it comes up "heads" both times?

3. Ken will get to go out with his friends tonight only if both his mom and dad say yes. There is a 60% chance his mom will say yes and an 80% chance his dad will say yes. What is the probability that Ken will get to go out tonight?

4. Janet is searching for a birthday present for her friend. Either a nice box of candy or a pretty bouquet of flowers would make a good present. From past experience, if she goes to the local shopping center there is a 40% chance she will find nice box of candy, and a 70% chance she will find a pretty bouquet. A) What is the overall chance she will succeed in finding at least one good birthday present for her friend when she goes to the shopping center: either a nice box of candy or a pretty bouquet?

B) And, what is the chance that she will find both birthday presents: both a nice box of candy and a pretty bouquet?

Studies 3&4 Situation Description

Please imagine the following situation:

Situation Description

Steve Seller was getting ready to put his house on the market for sale. Steve discovers a structural problem in the house and hires a repair person to fix the problem. The repair that was done was only a short-term fix.

Bill Buyer was searching for a home to buy. Steve's house was in a convenient location and had features Bill was looking for. Bill made an offer on the house, which Steve accepted.

Bill hired a home inspector, but the structural problem in the house was not discovered. Steve filled out a "Seller Disclosure" form and stated that everything about the house was in "excellent" condition. The house sale was finalized.

Shortly after Bill moved into the home the structural problem reappeared and then the short-term fix was discovered. Bill's repair person quoted a very large cost for the full permanent repairs.

Bill Buyer reached out to Steve Seller about the issue, but Steve refused to help Bill pay for the cost of repairs. Now Bill is taking Steve to court with a claim of real estate fraud. If Bill can prove his case, Steve will likely need to pay for some or all of the permanent repair costs.

Studies 3&4 Legal Instructions

Bill's Court Case

Plaintiff: Bill Buyer
vs.
Defendant: Steve Seller

Legal Instructions

For this real estate fraud case, there are two elements that need to be proven. The Buyer needs to prove that each of the following elements is more probably true than not true:

First, the Seller *knowingly* made a false statement of *material* fact.

Second, the Buyer's damages resulted from the Buyer's *reliance* on the Seller's statement.

Definitions:
Knowing means consciously with an understanding of the facts or circumstances.
Material means that it could have affected the Buyer's purchase decision or the sale price.
Reliance on a statement means being dependent on the statement, and acting based on the statement.

In order to win, the Buyer needs to prove their whole case is more probably true than not true, based on consideration of all of the evidence.

Studies 3&4 Task Information

Task Information

There are many possible cases of this type, with different facts, documentation, experts, and other unique considerations. On the following pages, you will consider 12 cases where “Bill Buyer” has provided different strengths of evidence in court.

Note: For these cases, please assume the truth of the elements are independent. That means they were decided based on entirely separate sets of evidence. (For example, knowing that the first proposition is true does not give you any information about the second proposition.)

[Study 3 only] For each case, we will tell you the probability that each element is true. Probabilities can range from 0% (definitely false) to 100% (definitely true).

[Study 4 only] For each case, we will tell you the probability that each element is true. Probabilities are provided on this scale: very unlikely, unlikely, somewhat unlikely, somewhat likely, likely, very likely

For each case, you will be asked to provide a judgment about how the overall case should be decided.

Study 3 Case Questions Page

Click here to display the Situation Description again.
Click here to display the Task Information again.

Bill's Court Case # X of 12

Plaintiff: Bill Buyer
vs.
Defendant: Steve Seller

Legal Instructions
For this real estate fraud case, there are two elements that need to be proven. The Buyer needs to prove that each of the following elements is more probably true than not true:

First, the Seller *knowingly* made a false statement of *material* fact.

Second, the Buyer's damages resulted from the Buyer's *reliance* on the statement.

Definitions:
Knowing means consciously with understanding of the facts or circumstances.
Material means that it could have affected the buyer's purchase decision or the sale price.
Reliance on a statement means being dependent on the statement, and acting based on the statement.

In order to win, the Buyer needs to prove their whole case is more probably true than not true, based on consideration of all of the evidence.

Case X Probabilities
For this case only, the evidence Bill Buyer provided makes it seem like it is:
X% likely that the statement was knowingly false and material
X% likely that reliance on the statement caused the damages

What do you think is the probability that Bill has proven his overall case?
____% [number text entry]

Given the above probabilities, what do you think is the legally correct outcome to the case:
O Bill should win his case
O Bill should lose his case

Study 4 Case Questions Page

Click here to display the Situation Description again.
Click here to display the Task Information again.

Bill's Court Case # X of 12

Plaintiff: Bill Buyer
vs.
Defendant: Steve Seller

Legal Instructions

For this real estate fraud case, there are two elements that need to be proven. The Buyer needs to prove that each of the following elements is more probably true than not true:

First, the Seller *knowingly* made a false statement of *material* fact.

Second, the Buyer's damages resulted from the Buyer's *reliance* on the Seller's statement.

Definitions:
Knowing means consciously with an understanding of the facts or circumstances.
Material means that it could have affected the buyer's purchase decision or the sale price.
Reliance on a statement means being dependent on the statement, and acting based on the statement.

In order to win, the Buyer needs to prove their whole case is more probably true than not true, based on consideration of all of the evidence.

Case X Probabilities

For this case only, the evidence Bill Buyer provided makes it seem like it is:
<verbal probability> that the statement was knowingly false and material (first element)
<verbal probability> that reliance on the statement caused the damages (second element)

[NUMERIC] What do you think is the probability that Bill has proven the first element is true?
___ % [number text entry] [repeated again with "second element"]

[BINARY] Do you think that Bill has proven the first element is true?
O Yes O No [repeated again with "second element"]

[NOELEMS doesn't have first/second element questions.]

[ALL] What do you think is the probability that Bill has proven his overall case?
___ % [number text entry]

[ALL] What do you think is the legally correct outcome to the case?
O Bill should win his case.
O Bill should lose his case.

Study 4 Stipulated Verbal Probabilities

We used Lichtenstein & Newman's (1967) mean response findings to guide the choice of verbal probability pairs. First we mapped the verbal probabilities into the numeric probability equivalent from Lichtenstein & Newman. We then used these numerical probabilities to calculate each strategy's answer and categorized combinations into key sets based on whether these strategies win or lose. We chose pairs of words that mapped into the desired case set. These mappings are detailed in Table 9 below.

Table 9. Study 4 Stipulated Verbal Probabilities for 12 Cases

Stipulated Verbal Probabilities That Participants Will See		Lichtenstein & Newman (1967) translations		Strategy's Answers if using on Translated Verbal Probabilities				Targeted Key Case Types		
first element	second element	1st elem	2nd elem	conj	avg	min	max	1st vs. 2nd	key	description
somewhat unlikely	very unlikely	31%	9%	3%	20%	9%	31%	>	*A*	no elements win
very unlikely	somewhat likely	9%	59%	5%	34%	9%	59%	<	*B*	one element wins, average loses
somewhat likely	unlikely	59%	18%	11%	39%	18%	59%	>	*B*	
likely	somewhat unlikely	72%	31%	22%	52%	31%	72%	>	*C*	one element wins, average wins
somewhat unlikely	likely	31%	72%	22%	52%	31%	72%	<	*C*	
somewhat unlikely	very likely	31%	87%	27%	59%	31%	87%	<	*C*	
somewhat likely	somewhat likely	59%	59%	35%	59%	59%	59%	=	*D*	[Conjunction Divergence] elements & average wins, conjunction loses
somewhat likely	likely	59%	72%	42%	66%	59%	72%	<	*D*	
likely	somewhat likely	72%	59%	42%	66%	59%	72%	>	*D*	
likely	likely	72%	72%	52%	72%	72%	72%	=	*E*	elements & average wins, conjunction wins
very likely	likely	87%	72%	63%	80%	72%	87%	>	*E*	
likely	very likely	72%	87%	63%	80%	72%	87%	<	*E*	

gray boxes = plaintiff wins, 1st/2nd elem = the first/second element's verbal phrase translated into a numeric probability, conj = the (conjunction) product of the two probabilities, avg = the average of the two probabilities, min = the minimum of the two probabilities, max = the maximum of the two probabilities. These were illustrative only for guiding the experimental verbal choices. In the study, each participant provided their own numeric translation.

Studies 3&4 Dependence Check Question

The case you looked at had these elements:
First, the Seller knowingly made a false statement of material fact.
Second, the Buyer's damages resulted from the Buyer's reliance on the Seller's statement.

If you know the first element was definitely true:
100% likely that the statement was knowingly false and material
Would this change your perception of the second element?
___% likely that reliance on the statement caused the damages

O The second element would now be less likely (lower percent)
O The second element would not be changed (same percent)
O The second element would now be more likely (higher percent)

Studies 3&4 Difficulty Question

What were your perceptions of the various parts of this study? (Select at least one answer per statement.)

checkbox grid options: easy, difficult, interesting, confusing, neutral, not applicable

The Background Information on Independence and Dependence.

The Background Information on Probabilities. [Study 3 *start* condition, Study 4]

The Situation Description story about Bill and Steve.

The Legal Instructions with elements and definitions.

Thinking about whether an element was true or not.

Choosing a probability for the overall case.

Choosing if Bill should win or lose the overall case.

Choosing the dollar amount of repair costs Bill would owe. [attention check]

Choosing different answers for 12 different cases.

The Probability Questions. [Study 3 *end* condition only]

Study 4 Probability Mapping Question

Numeric probabilities can range from 0% (definitely false) to 100% (definitely true). Please provide your own numeric probability ratings for each of these words:

very unlikely	____%	[number text entry, for each, 0-100]
unlikely	____%	
somewhat unlikely	____%	
somewhat likely	____%	
likely	____%	
very likely	____%	

Appendix 2: Study 1 Additional Analyses

Power Calculations

The sample size was determined using multiple power analysis procedures, and taking the maximum.

When required for the calculations, the values used were:

Significance level: .05
Power: .80
Effect size (Cohen's d and h): .3 (considered "small")

Sample means, standard deviations, and proportions were based on results from prior pilot studies.

For the probability of plaintiff proving the case (0-100):
Cohen (1988) for t tests for means suggested a sample size of n=138.
Lakens (2017) for equivalence test for the difference between two independent means suggested sample size of n=195 (for "weak" strength) or n=190 (for "strong" strength.)

For binary finding for the plaintiff or defendant (1 or 0):
Cohen (1988) for differences between proportions suggested a sample size of n=137.
Lakens (2017) for equivalence test for the difference between two proportions suggested sample size of n=193 (for "weak" strength) or n=145 (for "strong" strength.)

(All sample sizes are per cell.)

Element Decisions and Case Decision Consistency

22% of participants (127/571) who said the plaintiff should win selected at least one element as not being won. As was the case with overall probabilities, this error was more frequent in the participants who saw the weak evidence cases versus the strong evidence cases (strong: 16%; weak: 33%), $\chi 2(1, N=571)=19.27$, $p<.001$. (See Table 10.)

Table 10. Overall case decision errors among participants who chose that plaintiff should win

	Neither error	Probability error only (<51%)	Elements error only	Both types of errors
All (n=571)	418, 73%	26, 5%	85, 15%	42, 7%
Strong evidence (n=362)	294, 81%	9, 2%	45, 12%	14, 4%
Weak evidence (n=209)	124, 59%	17, 8%	40, 19%	28, 13%
strong vs weak proportions (all $\chi 2$ with df=1, N=571)	$\chi 2$=31.25, p<.001	$\chi 2$=8.47, p=.004	$\chi 2$=4.19, p=.041	$\chi 2$=16.29, p<.001

When looking instead at participants who decided the plaintiff should lose, patterns were similar, except error rates were always directionally (and sometimes significantly) higher in the strong evidence cases versus the weak evidence cases. (See Table 11.)

Table 11. Overall case decision errors among participants who chose that plaintiff should lose

	Neither error	Probability error only (>=51%)	Elements error only	Both types of errors
All (n=195)	139, 71%	21, 11%	23, 12%	12, 6%
Strong evidence (n=24)	6, 25%	5, 21%	5, 21%	8, 33%
Weak evidence (n=171)	133, 78%	16, 9%	18, 11%	4, 2%
strong vs weak proportions (all $\chi 2$ with df=1, N=195)	$\chi 2$=26.12, p<.001	$\chi 2$=1.81, p=.178	$\chi 2$=1.27, p=.259	$\chi 2$=29.85, p<.001

Overall, 27% (209/766) of participants had one of these errors related to the binary case finding. (See Table 12.)

Table 12. Overall case decision errors among all participants

	Neither error	Probability error only	Elements error only	Both types of errors
All (n=766)	557, 73%	47, 6%	108, 14%	54, 7%

The more probably true than not true cutoff can also be examined for each individual element. Overall error rates here were between 8% and 13%. The same strong versus weak evidence cases pattern was directionally present in each of the results, but rarely significantly. Overall, 19% (149/766) of participants had an error related to at least one of the binary elements findings. (See Tables 13 – 15.)

Table 13. Individual element probability errors, when element was true
(Participants who chose that element was true in binary decision, but probability <51%)

	e1	e2	e3	e4
All	55/471, 12%	23/406, 6%	40/531, 8%	36/484, 7%
Strong evidence	28/260, 11%	11/274, 4%	16/265, 6%	13/264, 5%
Weak evidence	27/211, 13%	12/132, 9%	24/266, 9%	23/220, 10%
strong vs weak proportions	χ2(1, N=471)= 0.29, p=.591	χ2(1, N=406)= 3.4, p=.065	χ2(1, N=531)= 1.3, p=.255	χ2(1, N=484)= 4.56, p=.033

Table 14. Individual element probability errors, when element was false

	e1	e2	e3	e4
All	18/98, 18%	21/170, 12%	9/43, 21%	14/87, 16%
Strong evidence	5/21, 24%	5/22, 23%	6/18, 33%	8/30, 27%
Weak evidence	13/77, 17%	21/170, 12%	3/25, 12%	6/57, 11%
strong vs weak proportions	χ2(1, N=98)= 0.17, p=.683	χ2(1, N=192)= 1.01, p=.314	χ2(1, N=43)= 1.73, p=.188	χ2(1, N=87)= 2.69, p=.101

Table 15. Individual element probability errors, combined

	e1	e2	e3	e4
All	73/569, 13%	44/576, 8%	49/574, 9%	50/571, 9%

Combining both binary case finding and elements findings errors together, 34% (264/766) of participants committed at least one consistency error. In the individual differences section, some differences between participants who did and did not make an error are discussed.

Order Effects

When looking at overall case judgments there was only one observed significant effect of order. When participants saw the elements questions before the overall case questions, their case probability answers were higher (elements first: M=70, SD=29; overall case first: M=65, SD=27), t(763.1)=2.49, p=.013. This was true for both the strong evidence (elements first: M=85, SD=17; overall case first: M=78, SD=19), t(383.9)=4.15, p<.001, and weak evidence (elements first: M=56, SD=30; overall case first: M=50, SD=27), t(377)=2.09, p=.037 subsets.

There was not an observed difference in case probability answers based on whether they saw the case probability (M=68, SD=28) or case binary (M=67, SD=28) question first, t(761.7)=0.5, p=.614, and these results were statistically equivalent with the power to detect a small effect (ΔL=-8.4, d=-0.3; ΔU=8.4, d=0.3), t(764)=3.65, p<.001. When participants saw the element questions before the overall case questions, there was no observed difference in rates of finding in favor of the plaintiff (elements first: 77%; overall case first: 72%), χ2(1, N=766)=1.77, p=.183, and these results were statistically equivalent with the power to detect a small effect (ΔL=-0.11, h=-0.3; ΔU=0.11, h=0.3), z=2.27, p=.012. Similarly there was not an observed difference in finding in favor of the plaintiff based on whether they saw the case probability or case binary question first, (case prob first: 77%; case binary first: 72%), χ2(1, N=766)=3.01, p=.083, and these results were statistically equivalent with the power to detect a small effect (ΔL=-0.11, h=-0.3; ΔU=0.11, h=0.3), z=1.81, p=.035.

There were not any consistently observed order effects on elements. There were a couple of scattered results that were significant at the .05 level (element 1 probability, strong evidence strength, elements vs. case first; and element 2 probability, weak evidence strength, elements vs. case first). However, given the large number of tests (48) to look for element order effects, and relatively high p-values for these two results (p=.030, p=.028), they do not seem to be indicative of true effects of ordering, and are instead close to the number of false positives we would expect given the number of tests and significance level. Additionally, many of the equivalence tests were significant, indicating there were no effects larger than a small size for order differences, and every equivalence test using the whole element population (instead of subsetting by weak or strong case evidence) resulted in significant equivalence tests. (The subsets may have been underpowered to detect equivalence.)

Strong vs weak Difference Testing

The pattern observed for overall cases was also true for all but one of the eight element questions, with the element 3 binary proportions difference being in the expected direction, but both proportions were very high and the difference was not statistically significant. This pattern was the same when considering only the two elements or only the four elements for strong vs. weak case evidence.

element 1 probability (strong: M=83, SD=25; weak: M=69, SD=30), t(555.3)=6.09, p<.001
element 1 binary (strong: 92%; weak: 73%), χ2(1, N=574)=36.01, p<.001
element 2 probability (strong: M=89, SD=19; weak: M=51, SD=40), t(399.7)=14.56, p<.001
element 2 binary (strong: 93%; weak: 48%), χ2(1, N=580)=142.24, p<.001
element 3 probability (strong: M=88, SD=21; weak: M=81, SD=23), t(573.3)=3.86, p<.001
element 3 binary (strong: 94%; weak: 91%), χ2(1, N=577)=1.05, p=0.152
element 4 probability (strong: M=86, SD=23; weak: M=75, SD=29), t(531.8)=4.95, p<.001
element 4 binary (strong: 90%; weak: 80%), χ2(1, N=577)=12.09, p<.001

Overall probability, two elems (strong: M=81, SD=20; weak: M=52, SD=28), t(340.2)=11.83, p<.001
Overall probability, four elems (strong: M=82, SD=17; weak: M=54, SD=29), t(302.3)=11.57, p<.001
Overall binary, two elems (strong: 93%; weak: 57%), χ2(1, N=387)=65.47, p<.001
Overall binary, four elems (strong: 95%; weak: 53%), χ2(1, N=379)=87.04, p<.001

Two vs four Difference Testing

The pattern observed for overall cases was also true for all but one of the eight element questions, with the element 3 probability means difference being in the unexpected direction, and significantly different (see Appendix). This pattern was the same when

considering only the weak case strength or only the strong case strength for two vs. four elements, except the element 3 probability difference only appeared significant with weak case strength.

Element 1 probability (two: M=74, SD=29; four: M=77, SD=29), t(380.8)=1.33, p=.185
element 1 binary (two: 80%; four: 84%), χ2(1, N=574)=1.31, p=.874
element 2 probability (two: M=71, SD=36; four: M=70, SD=38), t(414.4)=0.41, p=.683
element 2 binary (two: 73%; four: 69%), χ2(1, N=580)=0.94, p=.166
element 3 probability (two: M=81, SD=25; four: M=86, SD=21), t(338.2)=2.3, p=.022
element 3 binary (two: 89%; four: 94%), χ2(1, N=577)=4.7, p=.985
element 4 probability (two: M=78, SD=27; four: M=81, SD=27), t(383.6)=1.27, p=.20
element 4 binary (two: 86%; four: 85%), χ2(1, N=577)=0.12, p=.365

Overall probability, strong (two: M=81, SD=20; four: M=82, SD=17), t(378.6)=0.55, p=.581
Overall probability, weak (two: M=52, SD=28; four: M=54, SD=29), t(377.3)=0.71, p=.478
Overall binary, strong (two: 93%; four: 95%), χ2(1, N=386)=0.63, p=.785
Overall binary, weak (two: 57%; four: 53%), χ2(1, N=380)=0.82, p=.182

Two vs four Equivalence Testing

E1prob (ΔL=-8.6, d=-0.3; ΔU=8.6, d=0.3), t(572)=2.06, p=.020
E2prob (ΔL=-11.1, d=-0.3; ΔU=11.1, d=0.3), t(578)=3.02, p=.001
E3prob (ΔL=-6.7, d=-0.3; ΔU=6.7, d=0.3), t(575)=0.98, p=.165
E4prob (ΔL=-8, d=-0.3; ΔU=8, d=0.3), t(569)=2.11, p=.018
E1find (ΔL=-0.1, h=-0.3; ΔU=0.1, h=0.3), z=1.92, p=.027
E2find (ΔL=-0.12, h=-0.3; ΔU=0.12, h=0.3), z=2.08, p=.019
E3find (ΔL=-0.07, h=-0.3; ΔU=0.07, h=0.3), z=0.95, p=.171
E4find (ΔL=-0.09, h=-0.3; ΔU=0.09, h=0.3), z=2.46, p=.007

s$pprob (ΔL=-5.5, d=-0.3; ΔU=5.5, d=0.3), t(384)=2.4, p=.009
w$pprob (ΔL=-8.6, d=-0.3; ΔU=8.6, d=0.3), t(378)=2.21, p=.014
s$pwins (ΔL=-0.06, h=-0.3; ΔU=0.06, h=0.3), z=1.54, p=.062
w$pwins (ΔL=-0.14, h=-0.3; ΔU=0.14, h=0.3), z=1.89, p=.029

s$e1prob (ΔL=-7.5, d=-0.3; ΔU=7.5, d=0.3), t(283)=2.26, p=.012
s$e2prob (ΔL=-5.9, d=-0.3; ΔU=5.9, d=0.3), t(296)=2.09, p=.019
s$e3prob (ΔL=-6.4, d=-0.3; ΔU=6.4, d=0.3), t(283)=1.38, p=.085
s$e4prob (ΔL=-7, d=-0.3; ΔU=7, d=0.3), t(292)=2.27, p=.012

w$e1prob (ΔL=-9, d=-0.3; ΔU=9, d=0.3), t(287)=0.77, p=.221
w$e2prob (ΔL=-12.2, d=-0.3; ΔU=12.2, d=0.3), t(280)=2.04, p=.021
w$e3prob (ΔL=-6.8, d=-0.3; ΔU=6.8, d=0.3), t(290)=0.2, p=.423
w$e4prob (ΔL=-8.6, d=-0.3; ΔU=8.6, d=0.3), t(275)=0.56, p=.286

s$e1find (ΔL=-0.06, h=-0.3; ΔU=0.06, h=0.3), z=1.71, p=.043
s$e2find (ΔL=-0.05, h=-0.3; ΔU=0.05, h=0.3), z=1.39, p=.082
s$e3find (ΔL=-0.07, h=-0.3; ΔU=0.07, h=0.3), z=0.54, p=.294
s$e4find (ΔL=-0.06, h=-0.3; ΔU=0.06, h=0.3), z=1.16, p=.122
w$e1find (ΔL=-0.13, h=-0.3; ΔU=0.13, h=0.3), z=1.1, p=.135
w$e2find (ΔL=-0.15, h=-0.3; ΔU=0.15, h=0.3), z=1.62, p=.053
w$e3find (ΔL=-0.08, h=-0.3; ΔU=0.08, h=0.3), z=0.82, p=.206
w$e4find (ΔL=-0.11, h=-0.3; ΔU=0.11, h=0.3), z=1.88, p=.030

Two vs four Equivalence Testing with Minimum Product Difference

For this test, a conjunction product was calculated for each participant from their element probabilities. The difference in these products between two and four elements was compared, and the minimum difference was the small end of the 95% confidence interval.

Calculated products (from participants element probabilities)
weak (two: M=45, SD=35; four: M=30, SD=34), difference: M=15.5 (d=0.45), 95% CI [8.5, 22.5], t(378)=4.37, p<.001
strong (two: M=76, SD=31; four: M=66, SD=35), difference: M=10.3 (d=0.32), 95% CI [3.8, 16.9], t(375.1)=3.09, p=.002
all (two: M=61, SD=36; four: M=48, SD=39), difference: M=12.9 (d=0.34), 95% CI [7.6, 18.2], t(757.8)=4.75, p<.001

W (ΔU=8.5, d=0.3), t(378)=3.63, p<.001
S (ΔU=3.8, d=0.2), t(384)=2.56, p=.005
All just using mean difference in all: (ΔU=7.6, d=0.27), t(764)=4.53, p<.001

Also significant is use 3.8 value, as overall min: (ΔU=3.8, d=0.13), t(764)=2.64, p=.004

Element Correlations

An alternate reason that two-element and four-element cases could have similar overall results would be if the elements were very strongly correlated. If the element correlation was one, then the elements would be dependent on each other, and conjunction multiplication wouldn't be required. For example, P(e1) = P(e2) = P(e1 & e2). To rule out this explanation the correlations between elements are analyzed. There are no obvious patterns in the data, as indicated in Figures 10 – 13 below. Also, the individual correlations are at most about 50%, as indicated in Tables 16 – 21 below.

Additionally, considering the model where element probabilities predict overall probabilities, the variance inflation factors (VIF) were tested. The VIF tests did not detect collinearity among the elements.

Figure 10. Correlation Plots, Two Elements, Strong Case Evidence

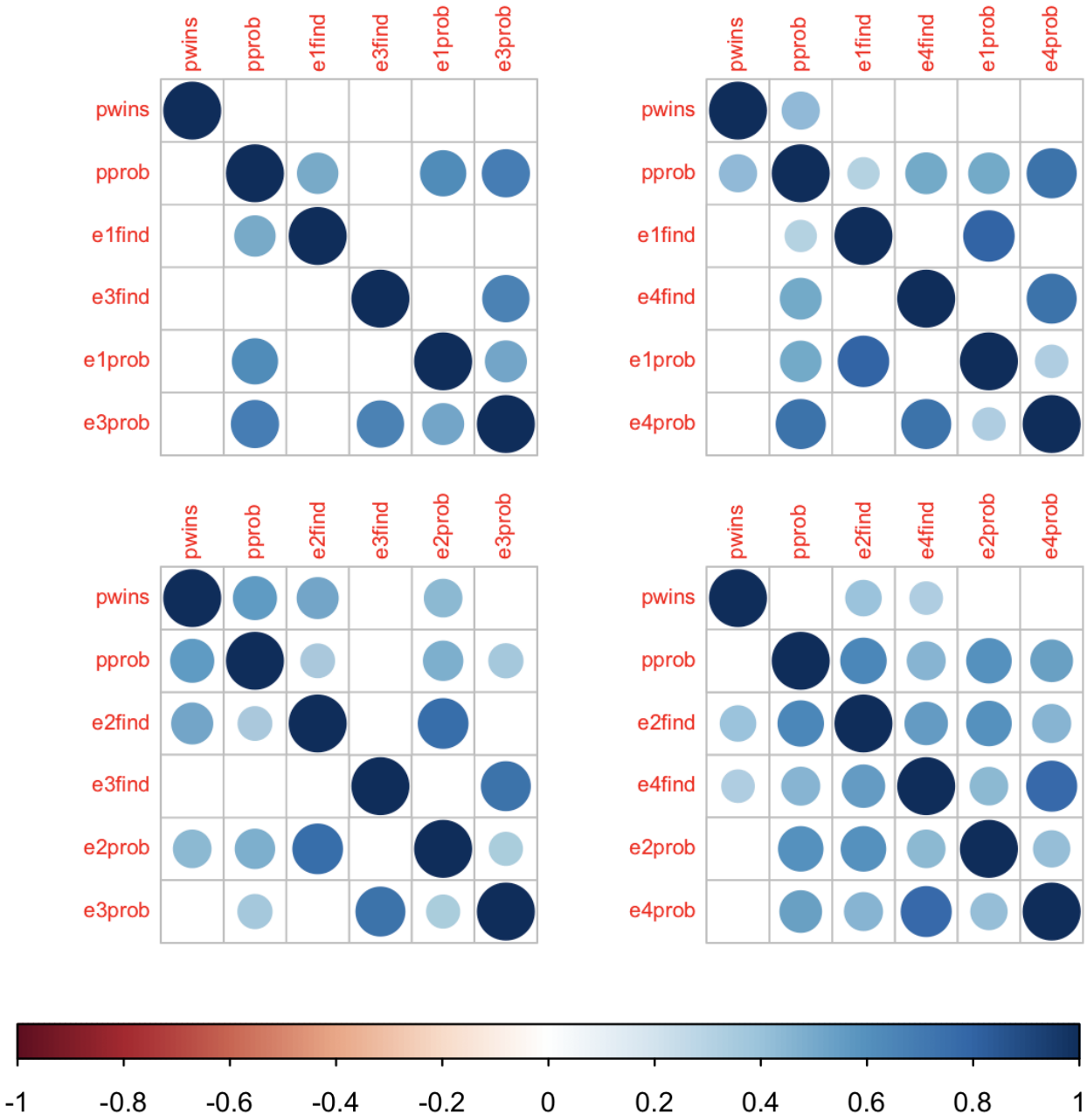


Figure 11. Correlation Plots, Two Elements, Weak Case Evidence

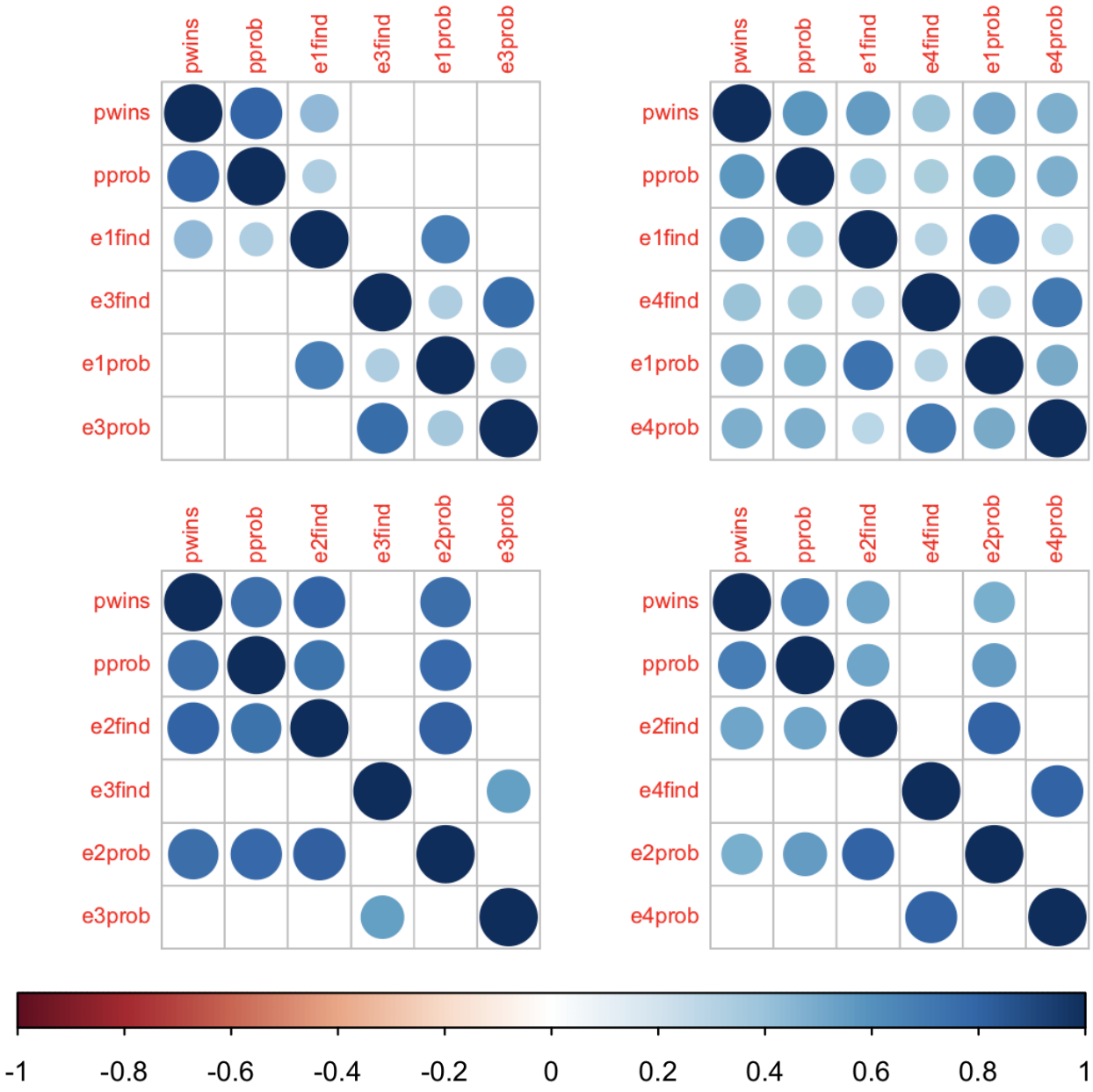


Figure 12. Correlation Plot, Four Elements, Strong Case Evidence

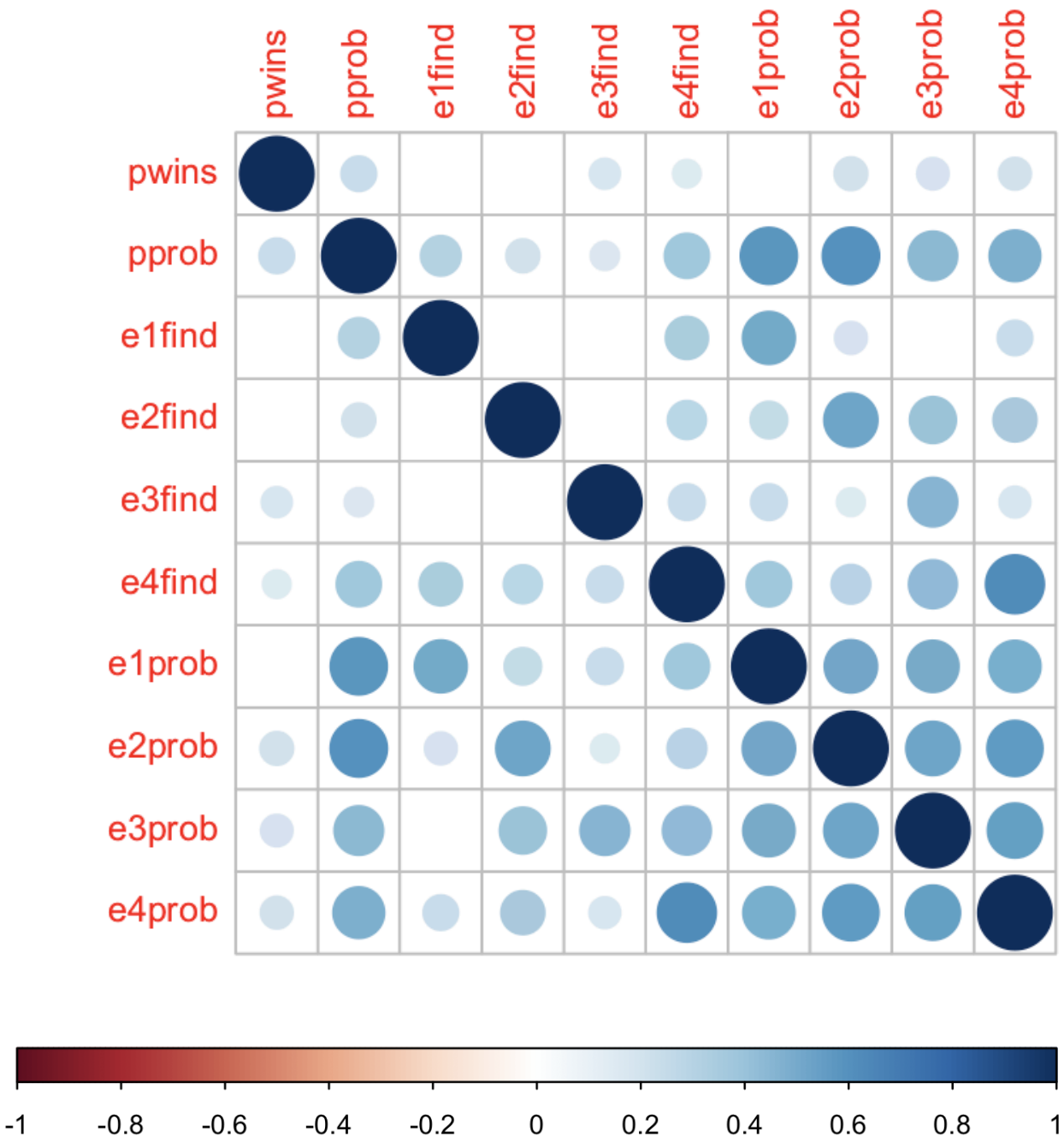


Figure 13. Correlation Plot, Four Elements, Weak Case Evidence

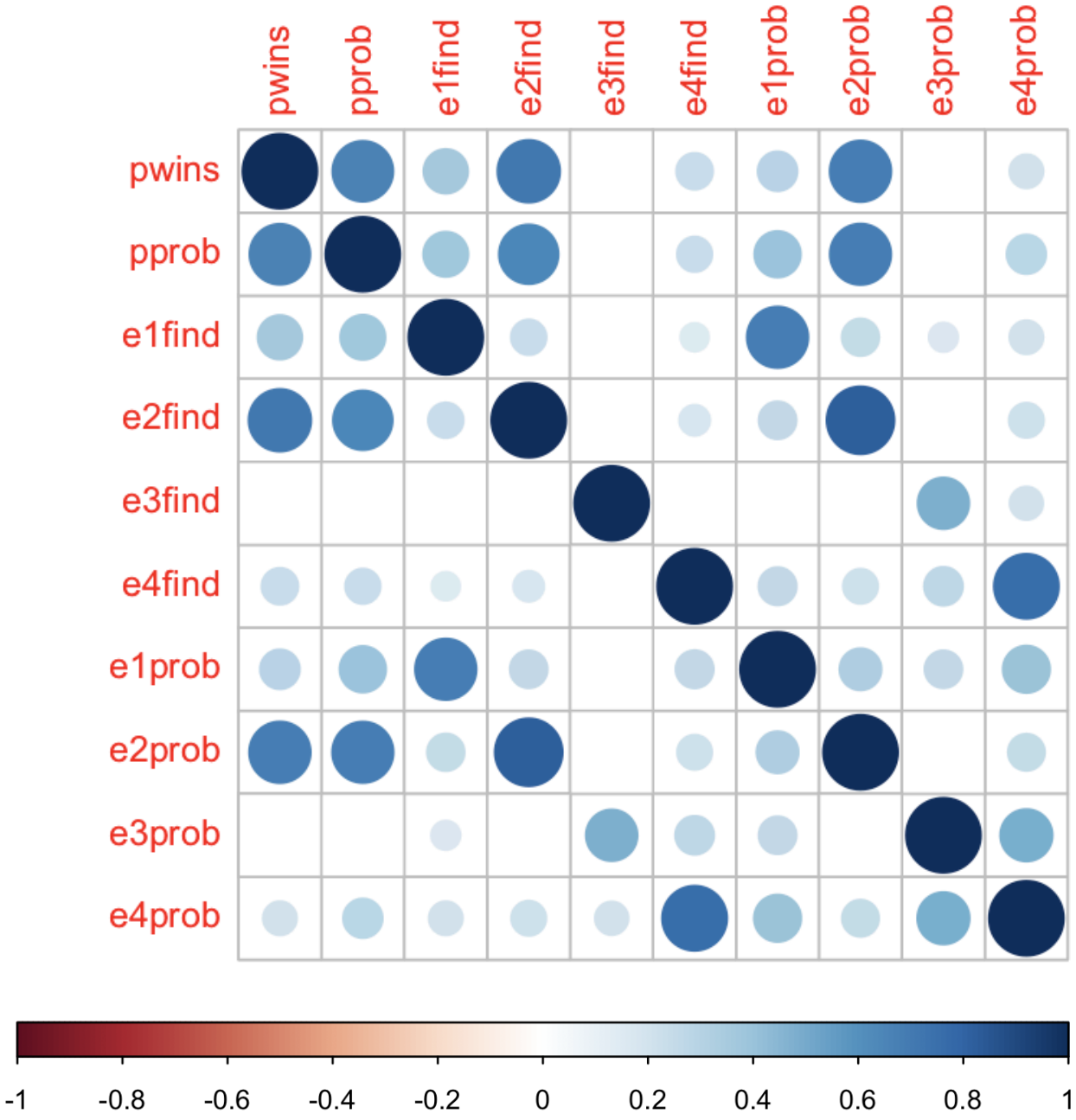


Table 16. Element Correlations, Four Elements - Strong

	e1prob	e2prob	e3prob
e1prob			
e2prob	0.51***		
e3prob	0.49***	0.53***	
e4prob	0.49***	0.57***	0.55***

p < .001 '***', p < .01 '**', p < .05 '*'

Table 17. Element Correlations, Two Elements - Strong

	e1prob	e2prob	e3prob
e1prob			
e2prob	NA		
e3prob	0.51***	0.33*	
e4prob	0.32*	0.40**	NA

p < .001 '***', p < .01 '**', p < .05 '*'

Table 18. Element Correlations, All Elements - Strong

	e1prob	e2prob	e3prob
e1prob			
e2prob	0.51***		
e3prob	0.49***	0.48***	
e4prob	0.45***	0.53***	0.55***

p < .001 '***', p < .01 '**', p < .05 '*'

Table 19. Element Correlations, Four Elements - Weak

	e1prob	e2prob	e3prob
e1prob			
e2prob	0.31***		
e3prob	0.25***	0.07	
e4prob	0.40***	0.24***	0.49***

p < .001 '***', p < .01 '**', p < .05 '*'

Table 20. Element Correlations, Two Elements - Weak

	e1prob	e2prob	e3prob
e1prob			
e2prob	NA		
e3prob	0.36*	0.10	
e4prob	0.49***	0.04	NA

p < .001 '***', p < .01 '**', p < .05 '*'

Table 21. Element Correlations, All Elements - Weak

	e1prob	e2prob	e3prob
e1prob			
e2prob	0.31***		
e3prob	0.27***	0.08	
e4prob	0.43***	0.21**	0.49***

p < .001 '***', p < .01 '**', p < .05 '*'

Appendix 3: Study 2 Additional Analyses

Main Effects

Binary Overall Case (Conjunction Problem: 57%; Non-Conjunction Problem: 54%), χ2(1, N=93)=0.09, p=.763

Probability Overall Case (Conjunction Problem: M=58, SD=27; Non-Conjunction Problem: M=56, SD=28), t(90.7)=0.38, p=.704

Binary Element 1 (Conjunction Problem: 81%; Non-Conjunction Problem: 83%), χ2(1, N=93)=0.05, p=.826

Binary Element 2 (Conjunction Problem: 49%; Non-Conjunction Problem: 48%), χ2(1, N=93)=0.01, p=.915

Binary Element 3 (Conjunction Problem: 87%; Non-Conjunction Problem: 89%), χ2(1, N=93)=0.08, p=.777

Binary Element 4 (Conjunction Problem: 74%; Non-Conjunction Problem: 74%), χ2(1, N=93)=0, p=.951

Probability Element 1 (Conjunction Problem: M=66, SD=30; Non-Conjunction Problem: M=67, SD=33), t(89.5)=0.13, p=.897

Probability Element 2 (Conjunction Problem: M=49, SD=37; Non-Conjunction Problem: M=50, SD=38), t(90.7)=0.2, p=.843

Probability Element 3 (Conjunction Problem: M=72, SD=25; Non-Conjunction Problem: M=70, SD=29), t(88.9)=0.44, p=.658

Probability Element 4 (Conjunction Problem: M=60, SD=30; Non-Conjunction Problem: M=65, SD=31), t(90.6)=0.77, p=.445

Equivalence Tests

Probabilities
Overall (ΔL=-8.2, d=-0.3; ΔU=8.2, d=0.3), t(91)=1.06, p=.145
E1 (ΔL=-9.5, d=-0.3; ΔU=9.5, d=0.3), t(91)=1.32, p=.096
E2 (ΔL=-11.3, d=-0.3; ΔU=11.3, d=0.3), t(91)=1.25, p=.108
E3 (ΔL=-8.1, d=-0.3; ΔU=8.1, d=0.3), t(91)=1, p=.160
E4 (ΔL=-9.2, d=-0.3; ΔU=9.2, d=0.3), t(91)=0.68, p=.249

Binary
Overall (ΔL=-0.14, h=-0.3; ΔU=0.14, h=0.3), z=1.09, p=.139
E1 (ΔL=-0.1, h=-0.3; ΔU=0.1, h=0.3), z=1.06, p=.145
E2 (ΔL=-0.15, h=-0.3; ΔU=0.15, h=0.3), z=1.32, p=.093
E3 (ΔL=-0.08, h=-0.3; ΔU=0.08, h=0.3), z=0.94, p=.173
E4 (ΔL=-0.12, h=-0.3; ΔU=0.12, h=0.3), z=1.24, p=.108

Appendix 4: Study 3 Additional Analyses

Individual Case Effects

We did not have any predictions about specific cases, but we can also look at case decisions made across each of the cases. (See Table 22.)

Table 22. Study 3 Individual Case Results by Condition

case key	elements		plaintiff probabilities mean(sd)			plaintiff wins proportion		
	1st	2nd	start page	end page	diffs	start page	end page	diffs
A	***20%***	**40%**	28.1 (14.8)	30.1 (15.3)		12%	19%	
B	***20%***	**60%**	35.9 (17.5)	39.6 (15.3)		24%	24%	
B	***60%***	**30%**	42.8 (14.4)	47.8 (14.1)	y	29%	36%	
C	***70%***	**40%**	51.8 (15.9)	55.9 (14.7)	y	65%	70%	
C	***40%***	**90%**	54.7 (18.1)	57.0 (13.9)		55%	62%	
D	***60%***	**60%**	57.5 (11.5)	58.6 (12.4)		78%	78%	
D	***60%***	**70%**	61.6 (13.1)	62.3 (13.5)		76%	79%	
D	***80%***	**60%**	68.4 (12.7)	67.3 (15.0)		85%	88%	
D	***70%***	**70%**	67.3 (13.6)	67.8 (13.5)		80%	89%	y
E	***90%***	**70%**	77.5 (13.7)	77.7 (18.2)		95%	91%	
E	***80%***	**90%**	81.0 (14.7)	80.2 (18.5)		94%	90%	
E	***90%***	**90%**	86.6 (16.4)	87.5 (17.0)		93%	92%	

y = statistical significant difference was observed ($p < .05$)

For the most part people responded similarly, as an entire group within each condition, to the cases. A few differences were observed. The numeric probability was higher in the *end* condition for the 60% & 30% case, $t(248.3)=2.79$, $p=.006$, and the 70% and 40% case, 55.9 (14.7), $t(249)=2.14$, $p=.034$. The proportion of wins was higher in the *end* condition for the 70% and 70% case, $\chi 2(1, N=251)=4.09$, $p=.043$. However the differences are relatively small, and

1.2 false positives would be expected with this many comparisons, so more evidence would be required to indicate meaningful shifts in some of these cases.

Appendix 5: Study 4 Additional Analyses

Element Consideration Case Effects

Here we will consider differences across the three element consideration conditions: *numeric*, *binary*, and *none* when it comes to individual cases. We are interested in whether explicitly contemplating each element impacts overall case outcomes. We will compare the overall binary win proportions across the conditions. We will also compare each case's binary win and overall case probability. (See Table 23.) Given that participants were also randomly assigned to considering elements winning and losing or not, we were also especially interested in the results of "B" and "C" key group questions, when one element wins but the other loses. In these cases, regardless of conjunctive math, the plaintiff should lose, but there are a number of wins selected. Since our element consideration conditions made for some earlier interesting comparisons, we added the extra "c5" case to Study 4, for a total of three "C" key group cases.

Table 23. Study 4 Individual Case Results by Condition

case key	plaintiff probabilities mean(sd)				plaintiff wins proportion			
	none	binary	numeric	diffs	none	binary	numeric	diffs
a1	25.7 (21.7)	29.2 (28.1)	31.3 (25.7)		19%	27%	28%	
b2	35.1 (20.3)	39.2 (26.3)	38.3 (24.8)		22%	36%	23%	b,c
b3	41.8 (20.4)	40.2 (24.0)	41.5 (20.7)		33%	46%	34%	b,c
c4	50.0 (17.9)	50.4 (22.6)	50.7 (17.3)		46%	47%	60%	a,c
c5	45.9 (19.0)	45.6 (21.4)	50.7 (18.6)	b,c	46%	39%	47%	
c6	50.7 (18.2)	50.2 (21.7)	54.6 (19.0)		50%	52%	59%	
d7	58.3 (15.0)	59.0 (21.9)	59.1 (18.0)		78%	76%	77%	
d8	63.7 (15.2)	64.4 (19.6)	63.5 (18.6)		87%	84%	80%	
d9	66.3 (15.8)	63.6 (19.8)	61.7 (18.2)	b	88%	82%	83%	
e10	72.0 (17.5)	71.9 (19.5)	66.8 (20.6)	b,c	91%	91%	90%	
e11	78.9 (14.6)	73.5 (21.2)	74.1 (18.9)	a,b	93%	90%	89%	
e12	77.4 (15.0)	74.2 (21.7)	72.9 (20.6)	b	93%	92%	91%	

a,b,c = statistical significant difference was observed ($p < .05$)
a=none vs. binary, b=none vs. numeric, c=binary vs. numeric

The two “B” cases had one barely winning element (“somewhat likely”) and a stronger losing element (“very unlikely” and “unlikely”), with the idea that for commonly used numbers for these phrases, even the average of both would be a losing case. For both of these cases, people in the *binary* condition chose a case overall win more often than people in the *none* (b2 none: 22%; binary: 36%), $\chi 2(1, N=275)=7.26$, p=.007, (b3 none: 33%; binary: 46%), $\chi 2(1, N=275)=4.6$, p=.032 or *numeric* (b2 numeric: 23%), $\chi 2(1, N=261)=6.02$, p=.014, (b3 numeric: 34%), $\chi 2(1, N=261)=3.97$, p=.046 conditions. Given that case instructions stipulate winning only if “each element” is proven, it is somewhat unexpected that people considering whether each element wins or loses are doing worse at following this legal instruction, than people not considering the elements via an explicit question about them. One explanation would be if the

binary group was indicating individual elements were both wins when they decided the case should win. The element answers from these participants are in Tables 24 and 25 below, grouped by how they decided the overall case. In both cases only a portion of the participants (54% in b2 and 33% in b3) choosing that the case should win also choose both elements winning. These types of logical violations will be discussed further in the Consistency section below. For now, we can say the binary group's decisions to let the plaintiff win more frequently aren't driven entirely by mis-deciding how the elements should turn out.

Table 24. Study 4 Case b2 binary condition answers

case b2 binary	overall case lose			overall case win			totals		
	e2 lose	e2 win	totals	e2 lose	e2 win	totals	e2 lose	e2 win	totals
e1 lose	42	44	86	3	16	19	45	60	105
e1 win	1	0	1	4	27	31	5	27	32
totals	43	44	87	7	43	50	50	87	137

case b2: e1 = very unlikely (lose), e2 = somewhat likely (win)
dark gray boxes = correct answers, light gray boxes = correct element choices

Table 25. Study 4 Case b3 binary condition answers

case b3 binary	overall case lose			overall case win			totals		
	e2 lose	e2 win	totals	e2 lose	e2 win	totals	e2 lose	e2 win	totals
e1 lose	36	2	38	6	1	7	42	3	45
e1 win	36	0	36	35	21	56	71	21	92
totals	72	2	74	41	22	63	113	24	137

case b3: e1 = somewhat likely (win), e2 = unlikely (lose)
dark gray boxes = correct answers, light gray boxes = correct element choices

The three "C" cases had one barely losing element ("somewhat unlikely") with stronger winning elements ("likely" and "very likely"), with the idea that for commonly used numbers for these phrases, the average of both would be a winning case. In these cases the individuals in *numeric* were the one that became differentiated, by giving out better results for the plaintiff. In

case c4, participants in the *numeric* condition chose that the plaintiff should win at a higher rate than those in the *none* condition (none: 46%; numeric: 60%), $\chi 2(1, N=262)=5.22$, p=.022, and *binary* condition (binary: 47%), $\chi 2(1, N=261)=4.45$, p=.035. In case c5, participants in the *numeric* condition gave the plaintiff a higher probability of winning compared to those in the *none* condition (*none*: M=45.9, SD=19.0; *numeric*: M=50.7, SD=18.6), t=-2.06, df=258.1, p=.040, and *binary* (*binary*: M=45.6, SD=21.4), t=-2.06, df=258.5, p=.041 condition. This pattern emerged as well for case c6, but the results were only marginally significant (none: M=50.7, SD=18.2; numeric: M=54.6, SD=19.0), t=-1.67, df=254.3, p=.096), (binary: M=50.2, SD=21.7), t=-1.75, df=258.7, p=.082. This suggests that by typing in higher numbers for the one winning element in these cases, the *numeric* condition participants are being swayed further in the direction of deciding for the plaintiff than the other conditions.

The final pattern to emerge from these results is that in the highest valued cases the participants in the numeric condition consistently give lower overall case probability numbers than those in the none condition. This is true in case d9 (*none*: M=66.3, SD=15.8; *numeric*: M=61.7, SD=18.2), t(245.1)=2.21, p=.028, e10 (*none*: M=72.0, SD=17.5; *numeric*: M=66.8, SD=20.6), t(242.5)=2.21, p=.028, e11(*none*: M=78.9, SD=14.6; *numeric*: M=74.1, SD=18.9), t(231.2)=2.31, p=.022, and e12 (*none*: M=77.4, SD=15.0; *numeric*: M=72.9, SD=20.6), t(222.7)=2, p=.047. This suggests that converting the verbal probabilities into numeric ones is driving down overall case probabilities. This is consistent with recent literature comparing the combination of verbal and numeric probabilities in a forecasting setting (Mislavsky & Gaertig, 2022).

In these four cases there is also a significant difference with the binary condition, joining the higher none condition in e10, and the lower numeric condition in e11, but the lack of

consistency here and the amount of tests being ran influencing significance don't suggest any obvious pattern or reasoning about the binary group here.

Consistency

Some consistency has been mentioned previously, like 11.2 out of 12 matches on average, between participants' overall case win choices and their case probability choices. Within participants, the average correlation between these two answers is 0.76 (SD=0.18) which also shows internal consistency. Of the 399 participants, 62.7% answered these two questions consistently for every case they saw. Of the remaining 149 participants who sometimes did not match, we can look at how their errors skewed. There were 75 (50.3%) participants who only had false wins (probabilities > 50 with a loss), 58 (38.9%) participants who only had false losses (probabilities < 50 with a win), and 16 (10.7%) participants who had a mix of both. Comparing these proportions, there were potentially more participants skewing in the direction of false wins, $\chi 2(1, N=298)=3.92$, p=.048[24].

For the *binary* and *numeric* condition participants we can also look for consistency between a participant's element answers and their overall case win or loss. We won't consider false losses here because there isn't a clear normative answer that dictates a win for most of the 12 cases.[25] However, it is clear that when at least one element loses, the overall case should lose, so we can look at false wins of this type. (See Table 26.)

[24] This is an admittedly high p-value for a non pre-registered comparison, but it is also consistent with other studies where participants tend to skew towards wins.
[25] This project would not exist if it was clear that, for example, a "slightly likely" first element and "slightly likely" second element should always win.

Table 26. Study 4 Overall Case False Wins Based on Elements Losing

# of false wins	0	1	2	3	4+	n
binary	32%	21%	16%	11%	20%	137
numeric	24%	18%	26%	15%	18%	124
Sum	28%	20%	21%	13%	19%	261

While matching one's own probability was at about two thirds doing perfectly, here there's under a third of participants matching elements perfectly. However, most people who exhibit this inconsistency are doing so for only a small number of cases.

Bibliography

Allen, R. J., & Pardo, M. S. (2019). Clarifying relative plausibility: A rejoinder. *The International Journal of Evidence & Proof*, 23(1-2), 205-217.

Allen, R. J., & Pardo, M. S. (2019). Relative plausibility and its critics. *The International Journal of Evidence & Proof*, 23(1-2), 5-59.

Anderson, N.H. (1981). *Foundations of Information Integration Theory*. New York: Academic Press.

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational behavior and human performance*, 9(3), 396-406.

Benjamin, D.J. (2019). Errors in probabilistic reasoning and judgment biases. In B.D. Bernheim, S. DellaVigna, & D. Laibson (Eds.), *Handbook of Behavioral Economics* (Vol. 2, pp. 69-126). New York: Elsevier, North Holland.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2(1), 6-10.

Charness, G., Karni, E., & Levin, D. (2010). On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior*, 68(2), 551-556.

Clermont, K. M. (2019). The silliness of magical realism. *The International Journal of Evidence & Proof*, 23(1-2), 147-153.

Clermont, K. M. (2015). Trial by traditional probability, relative plausibility, or belief function. *Case Western Reserve Law Review*, 66(2), 353-392.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Doyle, J. K. (1997). Judging Cumulative Risk 1. *Journal of Applied Social Psychology*, 27(6), 500-524.

Fan, Y., Budescu, D. V., & Diecidue, E. (2019). Decisions with compound lotteries. *Decision*, 6(2), 109-133.

Goldsmith, R.W. (1978). Assessing probabilities of compound events in a judicial context. *Scandinavian Journal of Psychology, 19(1),* 103-110.

Hastie, R. (2019). The case for relative plausibility theory: Promising, but insufficient. *The International Journal of Evidence & Proof*, 23(1-2), 134-140.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1(2), 259-269.

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4), 355-362.

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, 9, 563-564.

Mislavsky, R., & Gaertig, C. (2022). Combining probability forecasts: 60% and 60% is 60%, but likely and likely is very likely. *Management Science*, 68(1), 541-563.

Nance, D. A. (2019). The limitations of relative plausibility theory. *The International Journal of Evidence & Proof*, 23(1-2), 154-160.

Nance, D. A. (2016). *The Burdens of Proof: Discriminatory Power, Weight of Evidence, and the Tenacity of Belief*.

Pardo, M. S., & Allen, R. J. (2008). Juridical proof and the best explanation. *Law and Philosophy*, 27(3), 223-268.

Peterson, C.R., & Beach, L.R. (1967). Man as an intuitive statistician. *Psychological Bulletin, 68(1),* 29-46.

Phillips, L.D., & Edwards, W. (1966). Conservativism in a simple probability inference task. *Journal of Experimental Psychology, 72(3),* 346-354.

Schwartz, D. S., & Sober, E. (2019). What is relative plausibility?. *The International Journal of Evidence & Proof*, 23(1-2), 198-204.

Schwartz, D. S., & Sober, E. (2017). The Conjunctions Problem and the Logic of Jury Findings. *Wm. & Mary L. Rev.*, 59, 619.

Shafer, G. (1990). Perspectives on the theory and practice of belief functions. *International Journal of Approximate Reasoning*, 4(5-6), 323-362.

Shafer, G. (1976). *A mathematical theory of evidence* (Vol. 42). Princeton university press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science Magazine, 185*, 1124-1131

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4), 293-315.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129-133.

Wiggins, E. C., & Breckler, S. J. (1990). Special verdicts as guides to jury decision making. *Law & Psychol. Rev.*, 14, 1.

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594–604.