The Probabilities of Might and Would Counterfactuals

Michael J. Shaffer (St. Cloud State University) and James R. Beebe (University at Buffalo)

1. Might/Would Duality, Imaging, and the Probabilities of Counterfactuals

David Lewis famously defended the following analysis of the might conditional (i.e. "Might/Would Duality"):

(MWD) $\neg(p \mathrel{\square\!\!\rightarrow} \neg q) \equiv p \mathrel{\Diamond\!\!\rightarrow} q$.[1]

In other words, the negation of 'If p were true, q would not be true' is equivalent to 'If p were true, q might be true.' MWD is both an interesting and prima facie plausible principle about might and would counterfactuals. Following Adams, many logicians have also entertained the interesting and prima facie plausible view that the probability of a conditional is a conditional probability:

(CCCP) $P(p > q) = P(q \mid p)$ for all p, q in the domain of P such that $P(p) > 0$.[2]

Here '>' is being used to represent a generic conditional that includes indicatives. So understood CCCP does not necessarily have an obvious application to counterfactuals. More importantly, Lewis (1976) explicitly rejected CCCP with respect to many types of conditionals and suggested that the probabilities of these conditionals should rather be understood as policies for *feigned* minimal belief revision. On this view, the probability of such a conditional should be understood to be the probability of the consequent, given the minimal revision of $P(\cdot)$ that makes the probability of the antecedent of the conditional equal to 1. Formally, Lewis (1976 and 1986a) understands imaging as follows:

(IMAGE) $(p > q) = P'(q)$, if p is possible.

---

[1] See Lewis 1973a and 1973b.

[2] See Adams 1965, Adams 1975, Bennett 2003, Hájek 1994, Hájek and Hall 1994 and Arló-Costa 2014.

Here $P'(\cdot)$ is the minimally revised probability function that makes $P(p) = 1$. Lewis tells us that we are to understand $P(\cdot)$as a function defined over a finite set of possible worlds, with each world having a probability $P(w)$. Furthermore, the probabilities defined on these worlds sum to 1, and the probability of a sentence, p for example, is the sum of the probabilities of the worlds where it is true. In this context the image on p of a given probability function is obtained by 'moving' the probability of each world over to the p-world closest to w.[12] Finally, the revision in question is supposed to be the minimal revision that makes p certain. In other words, the revision is to involve only those alterations necessary for making $P(p) = 1$. What is interesting is that Lewis (1976, 308-312) explicitly believes that IMAGE correctly applies to Stalnaker conditionals and that Stalnaker's account of conditionals is basically correct for counterfactuals.[3] This very strongly suggests that Lewis is implicitly if not explicitly committed to IMAGE as an account of the probabilities of *counterfactuals*, particularly as he construes them in his 1973a and 1973b.

However, we maintain that jointly adopting MWD and IMAGE is deeply problematic and that one or both of them must go. In order to see this let us look at what these two claims imply about the probabilities of might counterfactuals. First, the probability calculus tells us that:

(PR) $P(\neg p) = 1 - P(p)$.[4]

By MWD, $P(p \Diamond\!\!\rightarrow q)$ is logically equivalent to $P(\neg(p \Box\!\!\rightarrow \neg q))$. By PR, $P(\neg(p \Box\!\!\rightarrow \neg q))$ is equal to $1 - P(p \Box\!\!\rightarrow \neg q)$. Finally, applying IMAGE, $1 - P(p \Box\!\!\rightarrow \neg q)$ is equivalent to $1 - P'(\neg q)$. Thus we derive the following crucial theorem:

(PMC) $P(p \Diamond\!\!\rightarrow q) = 1 - P'(\neg q)$.

[3] See Lewis 1976, p. 308 f.n. 8.
[4] See Howson and Urbach 1993.

This all looks very straightforward, but PMC seems to be deeply problematic when we take a closer look at what we think is the obviously correct way to understand might counterfactuals.

We can see this quite clearly by introducing a basic urn model as follows. In urn_1 there are 99 white balls and 1 black ball. All draws from all urn_1 are replaced. Let D_i represent the proposition that a draw is made from urn_i, let W_i represent the proposition that a white ball is drawn from urn_i and let B_i be the proposition that a black ball is drawn from urn_i. Now consider the following claims:

(c1) If I were to draw a ball from urn_1, then it might be a white ball.

(c2) If I were to draw a ball from urn_1, then it might be a black ball.

(c3) If I were to draw a ball from urn_1, then it would be a white ball.

(c4) If I were to draw a ball from urn_1, then it would be a black ball.

(c1) and (c2) can be regimented as follows:

(c1') $D_1 \Diamond\!\!\rightarrow W_1$.

(c2') $D_1 \Diamond\!\!\rightarrow B_1$.

According to PMC and the description of urn_1 the probabilities of (c1) and (c2) are supposed to be as follows:

(Pc1) $P(D_1 \Diamond\!\!\rightarrow W_1) = 1 - P'(\neg W_1) = .99$.

(Pc2) $P(D_1 \Diamond\!\!\rightarrow B_1) = 1 - P'(\neg B_1) = .01$.

These values can be determined because both $P'(\neg B_1)$ and $P'(\neg W_1)$ are fully fixed by the constitution of urn_1. Note too that they are not just equivalent to the probabilities of the consequents independent of the antecedents and so the conditionality involved is important and makes a difference. If no ball were drawn then there would be no chance it would be white and no chance it would be black (i.e. we would have probability 0 in both cases). So, by IMAGE, if

we feign a revision of belief such that $P(D_1) = 1$ (i.e. we feign that we are certain that a ball is drawn from urn_1), it is clear both that $P'(\neg B_1) = .99$ and that $P'(\neg W_1) = .01$. Moreover, according to IMAGE and given the urn_1 model the probabilities of (c3) and (c4) are also as follows:

(Pc3) $P(D_1 \mathrel{\square\!\!\rightarrow} W_1) = P'(W_1) = .99$.

(Pc4) $P(D_1 \mathrel{\square\!\!\rightarrow} B_1) = P'(B_1) = .01$.

Despite the seemingly odd result that (c1) and (c3) have the same probability and that (c2) and (c4) have the same probability, this all looks to be quite straightforward and is simply a consequence of jointly endorsing MWD and IMAGE.

However, on careful inspection, the probabilities of (c1) and (c2) so determined are at odds with what seems to be the correct way to understand them. The relevant English correlates of (Pc3) and (Pc4) are as follows:

(Pc3') The probability that if I were to draw a ball from urn_1, then it would be a white ball is .99.

(Pc4') The probability that if I were to draw a ball from urn_1, then it would be a black ball is .01.

These two expressions and their associated probabilities seem reasonable. But, this does not seem to be true in the case of the relevant English versions of (Pc1) and (Pc2):

(Pc1') The probability that if I were to draw a ball from urn_1, then it might be a white ball is .99.

(Pc2') The probability that if I were to draw a ball from urn_1, then it might be a black ball is .01.

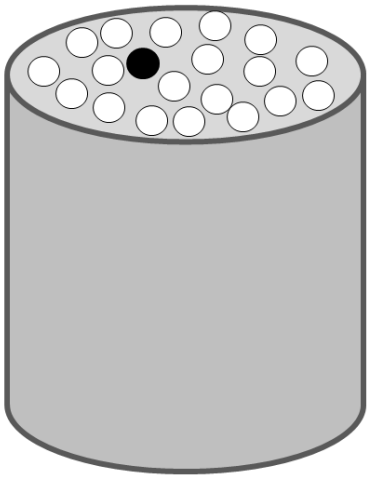
Unlike (Pc3) and (Pc4), these sentences that specify the probabilities of might counterfactuals do not seem to be correct. The probabilities of (c1) and (c2) should not be .99 and .01. The

probability that I *might* draw a white ball upon drawing a ball from urn_1 should be 1. This is because it is possible that I might do so. The same thing applies in the case of the probability that I *might* draw a black ball upon drawing a ball from urn_1. So, something appears to be wrong with the principles that have been used to derive PMC. Notice that this does not depend on the constitution of the urns in terms of the proportion of white to black balls where both are contained in an urn. Upon any draw from any urn in which there are both white and black balls, the probability *that I might draw a ball of a given color* is 1. This is because it is true that these outcomes might happen in those chance set-ups. So no matter the proportion of white to black balls in any such urn it is certain that both possible outcomes, drawing a white ball and drawing a black ball, are possible outcomes. So, it appears to be the case that the English usage of might in these conditional contexts is sensitive *only* to modal factors and that in English might counterfactual 'might' is not sensitive to probabilistic considerations whereas 'would' appears to be sensitive to such factors. But this is not reflected in systems that incorporate both MWD and IMAGE. Thus, from our perspective, MWD and IMAGE are incompatible and one or both of them must go.[5]

2. Study 1

[5] In the postscript to Lewis 1979 from Lewis 1986b—for rather different reasons—Lewis entertains the possibility that MWD is false and that there may be an alternate reading of some might counterfactuals. Thus, he contrasts what he calls the "not-would-not" (i.e. nwn) analysis of might counterfactuals with what he calls the "would-be-possible" (wbp) analysis (Lewis 1986c, 64). On this alternate analysis c1 would be analyzed as follows: If I were to draw a ball from urn_1, then it would be possible that it is a white ball. Similarly, c2 would be analyzed as follows: If I were to draw a ball from urn_1, then it would be possible that it is a black ball. Our intuitions are more consonant with this alternative reading, but Lewis officially endorses the nwn analysis captured by MWD in various works despite the apparent correctness of the wbp analysis. Moreover, he notes his reservations about this alternative (see 1986c, 63). In any case, we suggest then that independent of Lewis' own official commitments, there are deeply interesting questions about the acceptability of MWD that are highlighted in looking at how it interacts with IMAGE.

We wanted to investigate whether our intuitions about the probabilities of might and would counterfactuals matched those of ordinary individuals (i.e., non-philosophers). So in Study 1, we presented 60 participants (average age = 32, 37% female, 73% Caucasian) on Amazon's Mechanical Turk (www.mturk.com) with the instructions found in Table 1. All of the participants in each of our studies had at least a 98% approval rating on at least 5000 tasks on MTurk. They were paid $.40 each for approximately three minutes of their time. Participation in more than one study reported in this paper was prevented.



In the jar above there are 99 white balls and 1 black ball. The questions below ask you to reflect on the chance that certain statements about this jar are true. It is important to note that each question asks how probable you think the ENTIRE UNDERLINED STATEMENT is. So, be sure to think about each question in relation to the entire statement and not merely in relation to each if-clause or each then-clause.

Table 1. Participant instructions in Study 1.

Each participant was asked to respond to all four of the following counterfactual statements, the order of which was counterbalanced across the participant pool:

Might White. If you were to draw a ball from the jar, you might draw a white ball.

Might Black. If you were to draw a ball from the jar, you might draw a black ball.

Would White. If you were to draw a ball from the jar, you would draw a white ball.

Would Black. If you were to draw a ball from the jar, you would draw a black ball.

After each statement, we asked 'On a scale from 0 to 100, where 0 means 'absolutely impossible' and 100 means 'absolutely certain,' how likely do you think the underlined statement above is true?' We initially hypothesized that we would find support for our perspective on might and would counterfactuals in the intuitions of ordinary individuals. As we will see, this turned out to be only partially the case. Participant responses are summarized in Figure 1.

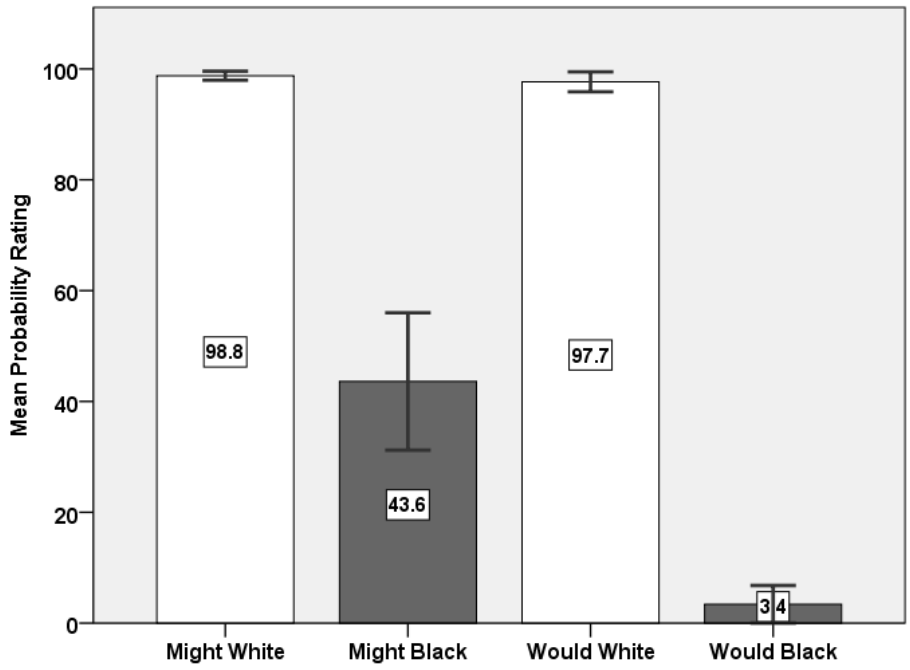


Figure 1. Mean probability ratings of Might White, Might Black, Would White, and Would Black in Study 1. Error bars represent 95% confidence intervals in all figures.

There was strong agreement among participants about how to assess the would counterfactuals. For Would White, almost all of the participants (52) gave 99 as their answer,[6] and in response to Would Black, 52 out of 60 participants gave 1 as their answer.[7] Thus, participants' probability assessments of Would White and Would Black correspond to what we think the correct answers are. These answers are also the ones that PMC deems to be correct.

In regard to the might counterfactuals, however, almost half of the participants agreed with our perspective and almost half agreed with PMC. For Might White, 26 participants gave 99 as their answer, and 29 gave 100 as theirs.[8] As we noted above, we believe that the correct probability assessment of Might White is 100, while according to PMC it should be 99. Because the answers of 99 and 100 are so close, one might think that if half of participants chose one answer and half chose the other, such a slight divergence should be theoretically insignificant. However, because participants' assessments of the two would counterfactuals were overwhelmingly in accord with what we think is the correct answer and because (as we will see) that participant responses to Might Black are also roughly split between those that accord with our view and those that accord with PMC, we do not think the observed differences regarding Might White should be passed over lightly.

[6] The remaining answers for Would White were 50, 80, 90, and 95. One participant did not give a valid answer to Would White. It is noteworthy that the foregoing responses are not from nine separate participants. The participant who responded with 80 to Might White responded with 90 to Would White, and one of the participants who responded with 90 to Might White gave 50 as their answer to Would White. One and the same participant gave the 95 answer to both questions.

[7] The remaining answers for Would Black were 0, 5, 5, 10, 10, 20, 99, and 99. The two participants who gave answers of 99 to this question spent less time on the entire task (51 and 60 seconds) than just about any other participants. The median time spent on the task as a whole was 141 seconds, and the average time spent was 331 seconds. So, it is not likely that the two outlying values of 99 are the result of deep, philosophical reflection on would counterfactuals.

[8] The remaining four answers for Might White were 80, 90, 90, 95, and 98.

The most significant finding from Study 1 is that the distribution of participant responses to Might Black was bimodal, with two large clusters at the far ends of the spectrum. 27 participants chose 1 as their answer, and 23 chose 100.[9] We believe that 100 is the correct probability assessment for Might Black, while according to PMC, 1 is the correct answer. In the case of Might Black, the divergence between the two sets of answers (1 and 100) is much greater than the two sets for Might White (99 and 100). It seems that our results call into question PMC to some degree, insofar as half of our participants failed to give the answer that PMC says is correct. Of course, our results also present an interesting challenge for our own view of how might counterfactuals should be understood, because half of the participants failed to agree with us.

A two-way repeated-measures ANOVA shows that there was clearly a significant main effect for modality (might vs. would) and a significant main effect for color (black vs. white), with a significant interaction between modality and color.[10] In other words, (i) it mattered whether the statement in question was a might or a would counterfactual, (ii) it mattered whether the statement in question concerned white or black balls, and (iii) the effect of modality and color was not uniform across all four statements (with participant answers to Might Black standing out from the others). All effect sizes were quite large. Outcomes (i) and (ii) both seem to be predicted by MWD and IMAGE, but (iii) does not seem to be predicted by either.

3. Study 2

Because of the bimodal distribution of responses we received to Might Black, we ran a second study that sought to examine more closely how participants understood the four counterfactual

[9] The remaining answers for Might Black were 2, 5, 5, 10, 10, 10, 10, 50, 70, and 75.

[10] Modality: $F(1, 58) = 38.31$, $p < .0001$, $r = .63$ (large effect size). Color: $F(1, 58) = 573.43$, $p < .0001$, $r = .95$ (very large effect size). Modality * color: $F(1, 58) = 34.90$, $p < .0001$, $r = .61$ (large effect size).

statements used in Study 1. In particular, we wondered whether some of the participants who gave surprising answers to Might Black might have failed to evaluate the probability of the entire conditional statement as we had instructed, ignoring the conditionality of the statement in the process. Therefore, in Study 2, we first presented participants with the following two conjunctive statements before having them respond to the four counterfactual conditionals used in Study 1[11]:

And White. You draw a ball from the jar, and it is white.

And Black. You draw a ball from the jar, and it is black.

The question after each statement was the same as in Study 1, viz., 'On a scale from 0 to 100, where 0 means 'absolutely impossible' and 100 means 'absolutely certain,' how likely do you think the underlined statement above is true?' The order of And White and And Black was counterbalanced, and the order of the remaining four statements was counterbalanced as well. We hypothesized that having participants reflect on the probabilities of these unconditional statements before asking them to evaluate conditional statements could improve their understanding of the latter.

Participants were 60 workers from MTurk (average age = 36, 38% female, 88% Caucasian). They were paid $.40 each for their participation. Their responses are summarized in Figure 2.

[11] The instructions in Table 1 were modified slightly to read 'In the jar above there are 99 white balls and 1 black ball. The questions below ask you to reflect on the probability that certain statements about this jar are true.'

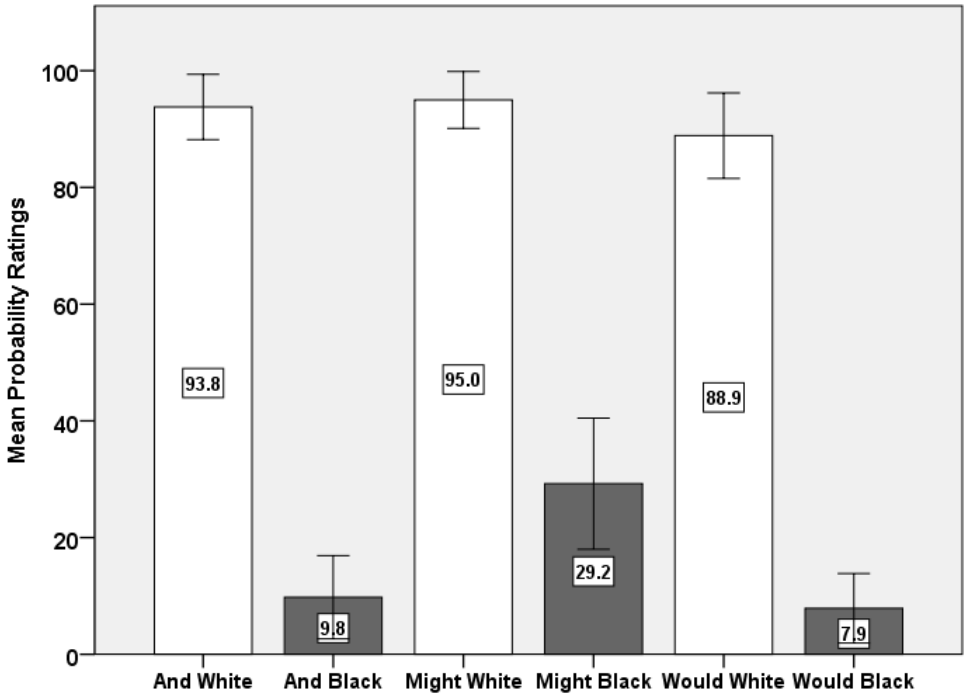


Figure 2. Mean probability ratings of Might White, Might Black, Would White, Would Black, And White, and And Black in Study 2.

As in Study 1, participants' probability evaluations of the two would counterfactuals corresponded almost exactly to what PMC and we both maintain are the correct answers. 47 of 60 participants gave 99 as their answer to Would White[12], and 48 gave 1 as their answer to Would Black[13]. Participant responses to Might White were again split into two (numerically close) groups, with 38 participants giving 99 as their answer and 15 giving 100.[14] However, a larger proportion of participants in Study 2 agreed with PMC's verdict on Might White than in Study 1, while a smaller proportion agreed with our verdict. This difference between participant responses in the two studies is statistically significant.[15]

[12] The remaining answers for Would White were 0, 0, 0, 1, 1, 50, 50, 90, 94, 97, 98, 98, and 100.

[13] The remaining answers for Would Black were 0, 0, 0, 2, 2, 3, 20, 50, 50, 99, 99, and 100.

[14] The remaining answers for Might White were 0, 1, 50, 95, 95, 98, and 98.

[15] A chi-squared analysis of the proportion of 99 and 100 answers to Might White in Studies 1 and 2 reveals a significant difference: χ^2 (1, N = 108) = 6.677, p < .05, Cramér's V = .25.

Also as in Study 1, we observed two groups of responses to Might Black at the extreme ends of the spectrum. Again, however, we observed more participants in Study 2 agreeing with PMC than in Study 1. In Study 1, approximately half of participants agreed with PMC and half agreed with us. In Study 2, however, over half of participants (36 out of 60) gave 1 as their probability rating of Might Black (which PMC deems to be the correct answer), while only 14 gave our preferred answer of 100.[16] Comparing the frequency of 1 and 100 answers (and ignoring all other answers), we found that the difference between the frequencies of these answers in Studies 1 and 2 approached statistical significance.[17]

Looking at participants' responses to the conjunctive statements we introduced in Study 2, we found that 48 of 60 participants gave the correct answer of 99 to And White[18], and 51 gave the correct answer of 1 to And Black.[19] When we excluded the 12 participants who did not answer both of these questions correctly, we found that the percentage of participants who agreed with PMC's verdict on Might Black increased and the percentage who agreed with our own verdict decreased (cf. Table 2).

	% choosing 1	**% choosing 100**
Study 1	45.0%	38.3%
Study 2 (all participants)	60.0%	23.3%
Study 2 (all participants who answered And White and And Black correctly)	64.6%	20.8%

Table 2. Percentages of participants who chose probability ratings of either 1 or 100 in response to Might Black in Studies 1 and 2.

[16] The remaining answers for Might Black were 2, 5, 8, 10, 10, 10, 25, and 50.
[17] χ^2 (1, N = 100) = 3.48, p = .062, Cramér's V = .19.
[18] The remaining answers were 0, 1, 1, 80, 97, 98, 98, 100, 100, 100, 100, and 100.
[19] The remaining answers were 0, 3, 5, 30, 99, 99, 100, 100, and 100.

Despite the fact that the distributions of participant responses to the four counterfactual statements from Studies 1 and 2 may look rather similar in Figures 1 and 2, a 2 x 2 x 2 (modality x color x study) mixed ANOVA reveals a significant difference between them.[20] In other words, having participants first reflect on And White and And Black made a statistically significant difference to the way they evaluated the other four statements. The largest such difference was found in their evaluation of Might Black, where the mean probability rating decreased from 43.6 in Study 1 to 29.2 in Study 2.[21]

We found these results to be surprising, inasmuch as it seems obvious to use that the probability rating of Might Black should be 100. Furthermore, and contrary to our prediction, having participants reflect on And White and And Black before evaluating might and would counterfactuals did not lead them to embrace our perspective in greater numbers. The results of Study 2 are positive for PMC, insofar as ordinary usage seems to accord better with its predictions than our own. However, it remains the case that not as many individuals agree with PMC's verdict on might counterfactuals as agree with its verdict on would counterfactuals. Nonetheless, we continue to maintain that we do not see how the value of *P*(If I were to draw a ball from urn, then it might be a black ball) could be anything less than maximal.

4. Study 3

In order to further examine whether participants were properly appreciating the conditionality of the counterfactual statements at the heart of our research project and to investigate the extent to

[20] There was a main effect for study, $F(1, 117) = 7.40$, $p < .01$, $r = .24$ (small effect size).

[21] Focusing on participant responses to Might White, Might Black, Would White, and Would Black, a 2 x 2 (modality x color) mixed measures ANOVA revealed (as in Study 1) significant main effects for modality (might vs. would) and color (white vs. black) and a significant interaction between them. Modality: $F(1, 59) = 17.42$, $p < .001$, $r = .48$ (medium to large effect size). Color: $F(1, 59) = 149.71$, $p < .001$, $r = .85$ (very large effect size). Modality * color: $F(1, 59) = 9.43$, $p < .01$, $r = .37$ (medium effect size).

which participants evaluated the conditionals using the feigned minimal belief revision hypothesized by Lewis (1976), we performed a third study that began with the following instructions:

> In the jar above there are 99 white balls and 1 black ball. Suppose you were to blindly draw a ball from the jar. What is the probability that the following statements would be true?

Participants were then shown the following four statements, which are the consequents of Might White, Might Black, Would White, and Would Black:

> *Might White 2*. You might draw a white ball.
>
> *Might Black 2*. You might draw a black ball.
>
> *Would White 2*. You would draw a white ball.
>
> *Would Black 2*. You would draw a black ball.

After each statement, participants were given the same instructions as above ('On a scale from 0 to 100, where 0 means 'absolutely impossible' and 100 means 'absolutely certain,' how likely do you think the statement above is true?'). As before, the order of the four statements was counterbalanced. We hypothesized that by (i) removing the conditionality of the might and would counterfactuals we have been investigating from the (antecedents of the) statements themselves, (ii) placing the conditionality in the instructions, and (iii) asking participants to evaluate unconditional statements that formerly appeared as consequents of the counterfactuals, participants might be led to provide more accurate assessments of claims involving 'might' and would.' Furthermore, this study design would allow us to test the accuracy of IMAGE as a model of ordinary usage, which predicts that participant responses in Studies 1 and 3 should not differ significantly.

Participants were 60 workers from MTurk (average age = 36, 47% female, 75% Caucasian). They were paid $.40 each for their participation. Their responses are summarized in Figure 3.

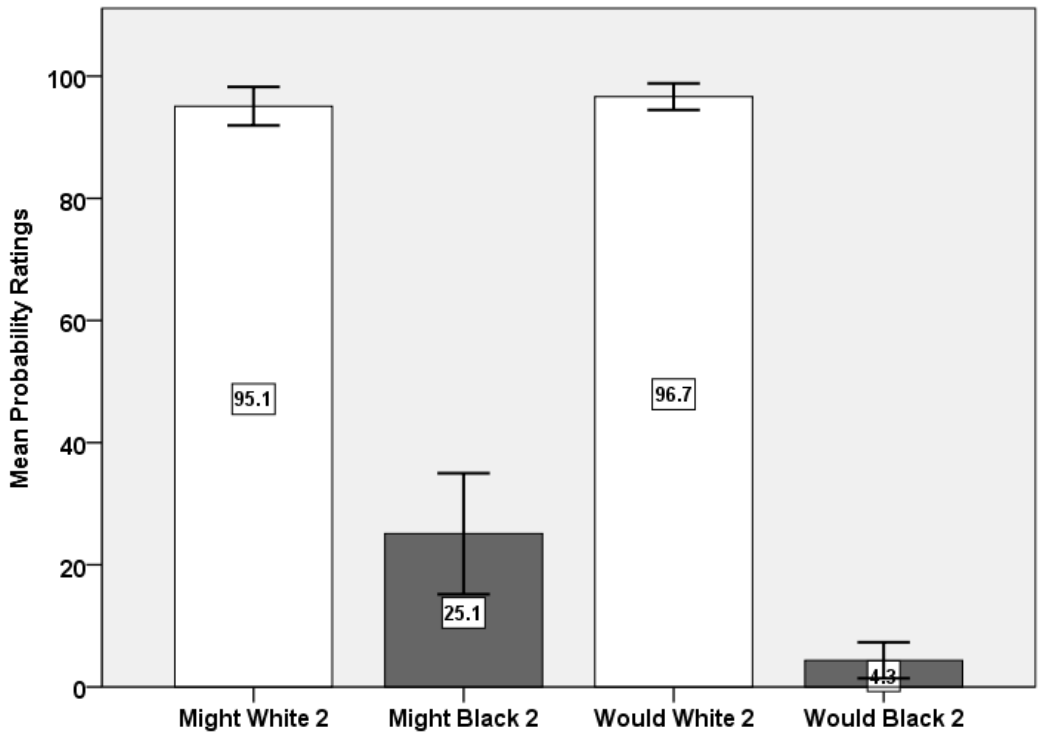


Figure 3. Mean probability ratings of Might White 2, Might Black 2, Would White 2, and Would Black 2 in Study 3.

We again observed strong agreement among participants regarding both would statements. 45 of 60 participants gave 99 as their answer to Would White 2, and 46 gave 1 as their answer to Would Black 2. In response to Might White 2, 29 participants gave 99 as their answer, and 17 gave 100 as their answer. Exactly half of participants gave 1 as their answer to Might Black 2, while only 9 gave 100. The remaining answers for Might Black 2 fell more

toward the lower end of the scale.[22] The same statistically significant differences observed in Studies 1 and 2 were found in Study 3 as well.[23] Although participant responses in Studies 1 and 3 shared certain similarities, there was nevertheless a statistically significant difference between them.[24] The main differences were that slightly fewer participants in Study 3 gave 99 and 1 as their responses to the would counterfactuals and fewer agreed with our assessment of the might counterfactuals. Thus, the manipulations that we introduced into Studies 2 and 3 in an effort that we thought might lead more participants to agree with our assessment of the probabilities of Might Black and Might Black 2 had the opposite effect. This also means that the data did not square fully with what IMAGE would predict, at least insofar as IMAGE is understood as a model of ordinary uses of 'might' and 'would.'

5. Study 4

As we noted above, in Studies 1 through 3, we asked participants 'On a scale from 0 to 100, where 0 means 'absolutely impossible' and 100 means 'absolutely certain,' how likely do you think the underlined statement above is true?' John Turri (personal communication) suggested to us that one reason why we might have obtained the surprising results reported above is that the verbal anchors we provided for participants' probability judgments are not naturally interpreted as opposites. In other words, the opposite of 'certain' may not be 'impossible.' Turri remarked:

> For instance, someone who focused more on the fact that '0' meant 'absolutely impossible' might then treat '100' to mean 'absolutely possible.' And someone who

[22] The remaining answers for Might Black 2 were 2, 3, 4, 5, 5, 10, 10, 10, 20, 20, 20, 20, 20, 40, 50, 50, 76, 80, and 80.

[23] There were significant main effects for modality and color and a significant interaction between them. Modality: $F(1, 57) = 14.26$, $p < .001$, $r = .45$. Color: $F(1, 57) = 606.77$, $p < .001$, $r = .96$. Modality * color: $F(1, 57) = 18.24$, $p < .001$, $r = .49$.

[24] A 2 x 2 x 2 (modality x color x study) mixed ANOVA revealed a significant main effect for study: $F(1, 115) = 7.45$, $p < .01$, $r = .25$. There were no significant interactions between study and the other variables.

> focused more on the fact that '100' meant 'absolutely certain' might then treat '0' as 'absolutely uncertain.' This could explain the bimodal pattern [of responses to Might Black], if people were roughly equally likely to interpret the entire scale by reference to the anchor more salient to them.

We agree that the verbal anchors we provided for participants' numeric answers were not ideally suited to be opposites. So, in Study 4 we reran Studies 1 and 3 using the same research materials but instead of using the question above, we asked participants 'On a scale from 0% to 100%, what is the chance that the underlined statement above is true?' We provided a blank marked '__%' in which they were to record their answers.

Participants in Study 4a (the modified replication of Study 1) were 60 MTurk workers (37% female, average age = 36, 82% Caucasian). They were paid $.40 each for their participation. Participant responses did not differ significantly from those of Study 1 (cf. Figure 4).[25]

[25] A 2 x 2 x 2 (modality x color x study) mixed ANOVA failed to reveal a significant main effect for study: $F(1, 117) = .955$, $p > .05$. There were no significant interactions between study and the other variables.

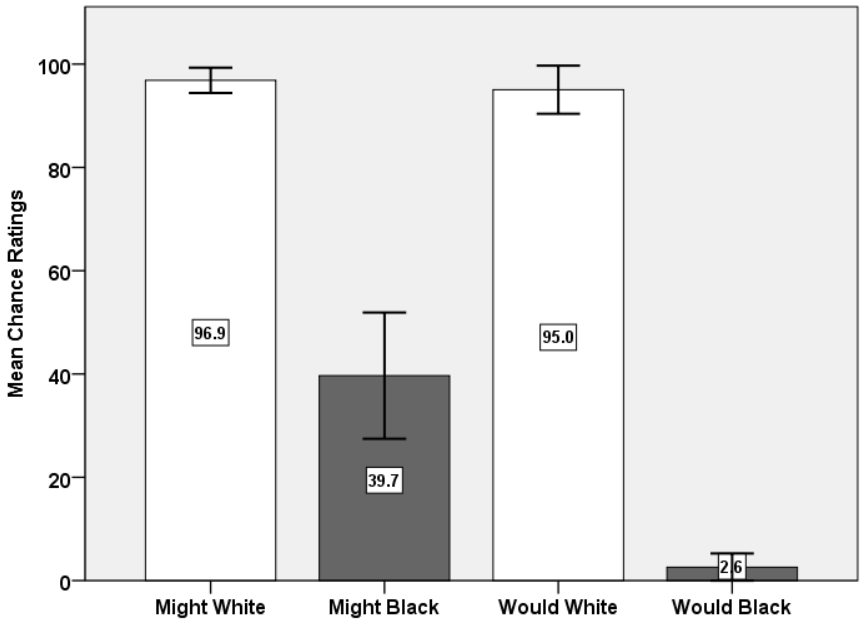


Figure 4. Mean chance ratings of Might White, Might Black, Would White, and Would Black in Study 4a.

As in Study 1, there was strong agreement among participants about how to assess the would counterfactuals, but participants were divided in their responses to the might counterfactuals. For Would White, 51 of 60 participants gave 99 as their answer[26], and 53 gave 1 as their answer for Would Black.[27] For Might White, 29 of 60 gave 99 as their answer and 22 gave 100.[28] 30 of 60 participants gave 1 as their answer to Might Black, while 21 answered 100.[29]

Participants in Study 4b (the modified replication of Study 3) were 60 MTurk workers (average age = 36, 38% female, 82% Caucasian).[30] They were paid $.40 each for their participation. Just as in Study 4a, rewording our central question did not significantly affect

[26] The remaining answers for Would White were 0, 0, 80, 90, 90, 97, 98, 98, and 100.
[27] The remaining answers for Would Black were 0, 0, 0, 2, 10, 10, and 80.
[28] The remaining answers for Might White were 50, 50, 75, 90, 90, 95, 95, 98, and 98.
[29] The remaining answers for Might Black were 3, 3, 5, 5, 10, 10, 50, 75, and 90.
[30] We used the same instructions as in Study 3, except that we change the word 'probability' to 'chance' in the following statement: 'In the jar above there are 99 white balls and 1 black ball. Suppose you were to blindly draw a ball from the jar. What is the probability that the following statements would be true?'

participants' responses.[31] 52 of 60 participants gave 99 as their answer to Would White 2, and 52 gave 1 as their answer to Would Black 2. In response to Might White 2, 36 gave 99 as their answer, and 18 gave 100. 36 participants responded with 1 as their answer to Might Black 2, and 16 responded with 100. None of these patterns of answers differed significantly from those in Study 3.[32] It does not seem, then, that the idiosyncrasies of the verbal anchors we provided in Studies 1 and 3 were responsible for generating the surprising participant responses we observed to Might Black and Might Black 2.

[31] A 2 x 2 x 2 (modality x color x study) mixed ANOVA failed to reveal a significant main effect for study: $F(1, 114) = .48$, $p > .05$. There were no significant interactions between study and the other variables.

[32] Chi-squared analyses were used to compare frequencies of 1, 99, and 100 answers. Might White 2: χ^2 (1, $N = 100$) = .14, $p = .71$, Cramér's V = .04. Might Black 2: χ^2 (1, $N = 91$) = .66, $p = .42$, Cramér's V = .09.

6. Study 5

Studies 1 through 4 all had within-subjects designs, which is to say that every participant responded to all four of the statements used in each study. We wondered whether being presented with all four statements might lead participants to the erroneous conclusion that the probability of any statement about drawing a black ball is .01. Consider the following facts:

(1) The probability of drawing a white ball from the urn is .99.

(2) The probability of drawing a black ball from the urn is .01.

(3) The probability of Would White ('If you were to draw a ball from the jar, you would draw a white ball') is .99.

(4) The probability of Would Black ('If you were to draw a ball from the jar, you would draw a black ball') is .01.

(5) The probability of And White ('You draw a ball from the jar, and it is white') is .99.

(6) The probability of And Black ('You draw a ball from the jar, and it is black') is .01.

Thus, for many of the statements used in our studies, there is a connection between a probability value of .99 and statements about white balls and a probability value of .01 and statements about black balls. We wondered whether this association might have led participants to unreflectively think that the probability of Might Black ('If you were to draw a ball from the jar, you might draw a black ball') was .01 as well simply because it was a statement about a black ball.

Therefore, in Study 5 we presented participants with only Might Black or Might Black 2, in the thought that isolating these statements would control for any possible undue influence from the patterns just described. In one condition, which we will call Study 5a, we reran Study 4a using only Might Black. In Study 5b, we reran Study 4b using only Might Black 2. Employing a between-subjects design, we recruited 100 MTurk workers (average age = 35, 42%

female, 78% Caucasian) and paid each one $.25 for their work. Participant responses are summarized in Table 3.

Study	*M*	% Choosing 1	% Choosing 100
5a	57.6	34%	54%
5b	37.6	54%	32%

Table 3. Summary statistics from Study 5.

Contrary to what we wondered might be the case, isolating Might Black and Might Black 2 in Studies 5a and 5b did not lead participants to overwhelmingly choose 100 as the correct answer to the question 'On a scale from 0% to 100%, what is the chance that the underlined statement above is true?' Only 54% (Study 5a) and 32% (Study 5b) chose our preferred answer.

7. Study 6

In an investigation of epistemic modals, John Turri (forthcoming) encountered surprising results with statements involving 'might' that bear some similarities to our own findings, although he investigated indicative unconditional statements while we investigated counterfactual conditionals. Turri found that individuals did not always treat statements such as 'Seth might not have plagiarized his paper' (in light of a certain body of information) and 'It's possible that Seth did not plagiarize his paper' (in light of that same information) as logically equivalent, whereas according to the consensus in philosophy they should have done so. On the received view (e.g., Kratzer 1977, von Fintel & Gillies 2007, MacFarlane 2011), 'It might be that *p*' is true just in case *p* is true in at least one (relevant) possibility that is consistent with a relevant set of available information. Some of the time, however, Turri's participants acknowledged that it was possible

that Seth did not plagiarize his paper but denied that he might not have done so. After telling participants that a smartphone engineer conducted a test and discovered that a smartphone was 80% likely to have spyware installed, Turri asked them to indicate the extent to which they agreed or disagreed that the smartphone might not have spyware. On a scale from 1 (strongly disagree) to 7 (strongly agree), the mean participant answer was 3.91, which was nearly indistinguishable from the neutral midpoint.

In light of Turri's results, we wanted to see if individuals' assessments of *what is possible* in regard to our urn of white and black balls matched their assessments about *what might happen* when a ball is drawn. Therefore, in Study 6, we asked participants to evaluate the probabilities of the following statements:

(P1) If you were to draw a ball from the jar, it is possible for you to draw a black ball.

(P2) It is possible for you to draw a black ball.

(P1) and (P2) are modeled after Might Black and Might Black 2, but the 'you might' in the originals was replaced by 'it is possible for you to.' We used the same instructions as in Study 4, viz., 'On a scale from 0% to 100%, what is the chance that the underlined statement above is true?'

We predicted that participants would overwhelmingly choose 100 as the correct response to both (P1) and (P2), although of course by this point we knew not to be overly confident in our predictions. The grounds for our prediction were that it seems intuitively obvious that 100 is the correct answer to both questions and because Turri found that participants' assessments of what is possible sometimes diverged from their assessments of what might happen. Participants in Study 6 were 100 MTurk workers (average age = 34, 41% female, 76% Caucasian) who were

paid $.25 for a two minute task. In a between-subjects design, participants were presented with either (P1) or (P2). Their responses are summarized in Table 4.

Condition	*M*	% Choosing 1	% Choosing 100
P1	44.9	48%	40%
P2	41.5	52%	36%

Table 4. Summary statistics from Study 6.

Once again, we were surprised by what we observed. Participants failed to give high probability ratings to statements (P1) and (P2), even though it seems clear that one can know with certainty that they are true. Furthermore, in contrast to some (but not all) of Turri's findings, our participants gave probability estimates of statements about the possibility of a black ball being drawn that matched their estimates of the probability that a black ball might be drawn.

8. Study 7

In our final study, we again departed from asking participants to evaluate the probabilities of might and would statements. In Study 7, we simply asked participants to indicate the extent to which they agreed or disagreed that Might White and Might Black were true. Participants were given the answer choices 'Completely Disagree,' 'Mostly Disagree,' 'Slightly Disagree,' 'Neither Agree nor Disagree,' 'Slightly Agree,' 'Mostly Agree,' and 'Completely Agree.' Participants were 100 MTurk workers (average age = 35, 40% female, 75% Caucasian) who were paid $.25 for their time. Study 7 had a between-subjects design, with participants receiving only Might White or Might Black.

Assigning a value of 1 to Completely Disagree, 2 to Mostly Disagree, and so on, and averaging participant responses resulted in a mean agreement rating of 6.38 (out of 7) for Might White and 5.28 for Might Black. Both of these means fell significantly above the neutral midpoint of 4 (with large effect sizes).[33] 92% of participants selected 'Slightly Agree,' 'Mostly Agree,' or 'Completely Agree' for Might White, and 72% did so for Might Black. On the whole, then, participants agreed that the two might counterfactual were true, although not in equal proportions. This result is rather puzzling, given that so many participants assigned Might Black and its cousin, Might Black 2, low probability values.

9. General Discussion

In Section 1 we showed that MWD—the thesis of might/would duality—and IMAGE—the thesis that the probabilities of counterfactual conditionals should be understood as policies for feigned minimal belief revision—jointly entail PMC. We showed that PMC implies that the probabilities of Would White and Would Black would be .99 and .01, respectively. We agreed that this was the correct verdict and proceeded to show that the intuitions of ordinary individuals were very solidly in accord with this viewpoint (cf. Table 5).

Study	**% Choosing Intuitively Correct Answer**	
	Would White	**Would Black**
1	86.7%	86.7%
2	78.3%	80.0%
3	75.0%	76.7%
4a	85.0%	88.3%
4b	86.7%	86.7%

Table 5. Percentages of participants in Studies 1 through 4 who chose the intuitively correct answer of .99 for Would White and .01 for Would Black.

[33] Might White: $t(49) = 11.79$, $p < .001$, $r = .86$. Might Black: $t(49) = 4.18$, $p < .001$, $r = .51$.

We also showed that PMC implies that the probabilities of Might White and Might Black should be .99 and .01. We argued that these probabilities should instead both be 1. If you know there are 99 white balls and 1 black ball in an urn, and you know it is possible to draw either a white or a black one, it seems that you can be certain that you might draw a white one and certain that you might draw a black one. Following the consensus in philosophy, we assumed that there was a simple and straightforward connection between what is possible and what might happen and between being certain that something is possible and being certain that something might happen.

When we investigated folk assessments of these notions, however, we encountered some surprising results. In Studies 1 through 4, we observed between 18% and 42% of participants assigning a maximal probability value or chance rating of 100% to a proposition that concerned something participants knew might happen. Only in Study 5a did this percentage top 50%. After testing a variety of experimental manipulations designed to better guide participants in their tasks and focus their attention on the relevant features of the cases, we failed to find an increased percentage of participants who gave what seems to us to be the intuitively correct verdict regarding might counterfactuals and related statements involving 'might.'

Nevertheless, when we look at correlations between participants' level of education and their responses to Might Black and Might Black 2, we find some measure of vindication. We asked participants in all of our studies to indicate their level of education, giving them the answer choices 'Some high school,' 'High school graduate,' 'Some college, no degree,' 'Associates degree,' 'Bachelors degree,' and 'Graduate degree (Masters, Doctorate, etc.).' Combining data from Studies 1 through 5, a significant positive correlation was found between level of education

and higher probability assessments of Might Black and Might Black 2 (r = .11, p < .05).[34] In other words, increased education made participants more likely to agree with our assessment of Might Black and less likely to agree with PMC.

Turri (forthcoming) believes that the results of his investigation of epistemic modals constitute "the first empirical evidence that the strong view [that a proposition must be true just in case it is true in all the possibilities consistent with the available information and might be true just in case it is true in at least one possibility consistent with the available information] fails to capture the ordinary meaning of epistemic modals." Although we did not take a stand on 'must' statements, we believe that what Turri calls the strong view of epistemic modals gives the correct verdict of might statements. There are various reasons why we do not follow Turri in thinking that the empirical evidence falsifies the strong view of might statements. One is that Might Black, Might Black 2, (P1), and (P2) seem to be obviously correct on the basis of the ordinary meanings of the relevant terms and not merely on a theoretically loaded interpretation of those terms. A second reason is that in both Turri's studies and our own, participants often failed to agree that an event was possible when they were explicitly told that it was clearly possible. This suggests an important lack of competence (or at least a very significant performance error) in handling basic modal notions (at least in the contexts that have been investigated).

Perhaps the most important reason why we are reluctant to agree that our data (and some of Turri's) falsifies our preferred interpretation of the probabilities of might and would counterfactuals is that participants in both sets of studies seemed to be especially poor at assigning probabilities to non-atomic and non-indicative statements. They seemed largely capable of providing correct assessments of the probabilities of atomic, non-conditional

[34] We also found a significant correlation between gender and Might Black responses, with females being more likely to give lower probability estimates, r = -.14, p < .01. There were no gender differences on any other measures. A positive correlation of .10 between age and Might Black responses approached significance, p = .057.

statements like '*x* is *F*.' But they seemed rather unable to handle conditional statements like 'If *x* is *F*, then *x* is *G*,' modal statements like '*x* might be *F*,' or—worst of all—might counterfactuals like 'If *x* were *F*, then *x* might be *G*.' For example, even when participants acknowledged that it was possible for a given *x* to be *F*, they often assigned statements such as *P*(It is possible for *x* to be *F*) near zero probability ratings. In asking participants to assess the probabilities of might counterfactual, we are asking them to think about probabilities, non-indicative modalities, and the conditional structure of test statements all at the same time. We suspect that this combination of task demands resulted in research materials that often exceeded the competence of our participants.[35]

We believe there are several things that we have accomplished in this article. First, in the more theoretical part of our paper, we believe we have formulated a significant challenge to a certain combination of views about the duality of might and would counterfactuals. Secondly, in the empirical portion of our paper, we have added to the store of unexpected findings concerning ordinary individuals' assessments of 'might' statements. Our data did not fully fit with either PMC or the received view among theorists about how these terms behave. We believe that the unexpected results we report will provide the occasion for further theory building and empirical investigation of folk conceptions of modalities.

Turri's investigation of 'might' has led him to the hypothesis that 'might' functions as the dual to 'knows' in something like the following fashion:

---

[35] When Turri provided participants with probabilistic information from the 'inside view' (Kahneman & Tversky 1982), which pertains to dispositions of particular individuals, events, or entities, rather than 'outside' or background base rates, participants performed even worse on probabilistic assessment. The probabilistic information that we provided counts as coming from the outside view. An 'inside view' version of our experiments would have had us telling participants the following: "John has drawn a ball from the urn, but he has not yet looked at what color it is. There is a 99% chance that it is white, and a 1% chance that it is black. Please tell us whether you think the following statement is true: 'The ball that John drew might be black.'" If Turri's results are any indication, participants would have performed even worse on this question than on the questions we actually gave them. It is noteworthy that Kahneman (2011, ch. 23) ardently maintains that the outside perspective should be preferred to the inside perspective.

(T) 'It might be that *p*' is true (given a certain body of information available to *S*) ≡ *S* does not know that not-*p*.

We do not at present believe there is sufficient evidence for endorsing Turri's hypothesis. One reason is that his participants did not always attribute knowledge or assess 'might' statements in a way that conformed to (T). Secondly, when their judgments did conform more closely to (T), it was when Turri provided them with probabilistic information from Daniel Kahneman and Amos Tversky (1982) call 'the inside view,' which pertains to dispositions of particular individuals, events, or entities, rather than from 'the outside view,' which concerns base rates or broad distributions of data.[36] Kahneman (2011, ch. 23), however, ardently maintains that people are more likely to make better and more accurate judgments when they are given information from the outside view or when they adopt an outside view in regard to the information that they have available.

The explanation for our unexpected data concerning might statements may simply be that lack of training and experience made the assessment of the probabilities of conditional modal statements rather difficult. But whether this simple hypothesis is true or whether a revisionist understanding of modal notions, such as the one suggested by Turri is correct, we believe that empirical investigations such as ours can help scholars ascertain how ordinary individuals reason with and make judgments about such notions.

[36] The probabilistic information that we provided counts as coming from the outside view. An 'inside view' version of our experiments would have had us telling participants the following: "John has drawn a ball from the urn, but he has not yet looked at what color it is. There is a 99% chance that it is white, and a 1% chance that it is black. Please tell us whether you think the following statement is true: 'The ball that John drew might be black.'" If Turri's results are any indication, participants would have performed even worse on this question than on the questions we actually gave them.

References

Adams, E. (1965). "The Logic of Conditionals," *Inquiry* 8, 166-197.

Adams, E. (1975). *The Logic of Conditionals*. Dordrecht: Reidel.

Arló-Costa, H. (2014). "The Logic of Conditionals," *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2014/entries/logic-conditionals/>.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Eells, E. and B. Skyrms, eds., (1994) *Probability and Conditionals: Belief Revision and Rational Decision*, Cambridge: Cambridge University Press.

von Fintel, K., & Gillies, A. S. (2007). "An Opinionated Gguide to Epistemic Modality." In T. S. Gendler & J. Hawthorne (Eds.), *Oxford Studies in Epistemology*, Vol. 2. Oxford: Oxford University Press, pp. 32–62.

Hájek, A. (1994) "Triviality on the Cheap?" In E. Eells and B. Skyrms (1994), 113-141.

Hájek, A. and N. Hall (1994) "The Hypothesis of Conditional Construal of Conditional Probability", in E. Eells and B. Skyrms (1994), 75-113.

Howson, C. and P. Urbach (1993). *Scientific Reasoning: The Bayesian Approach*, 2nd ed. Open Court, Chicago.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: Farrar, Straus, and Giroux.

Kahneman, D., & Tversky, A. (1982). "Variants of Uncertainty." *Cognition* 11: 143–57.

Kratzer, A. (1977). "What Must and Can Must and Can Mean." *Linguistics and Philosophy* 1: 337–55.

Lewis, D. (1973a). *Counterfactuals*. Harvard University Press, Cambridge.

Lewis, D. (1973b). “Counterfactuals and Comparative Possibility,” *Journal of Philosophical Logic* 4: 418-446.

Lewis, D. (1976). “Probabilities of Conditionals and Conditional Probabilities,” *Philosophical Review* 85, 297-315.

Lewis, D. (1979). “Counterfactuals and Time’s Arrow,” *Nous* 13: 455-476.

Lewis, D. (1986a). “Probabilities of Conditionals and Conditional Probabilities II,” *Philosophical Review* 95, 581-589.

Lewis, D. (1986b). *Philosophical Papers* Vol. II. Oxford: Oxford University Press.

Lewis, D. (1986c). “Postscripts to Counterfactuals and Time’s Arrow,” in Lewis 1986b, 52-66.

MacFarlane, J. (2011). “Epistemic Modals are Assessment-Sensitive.” In A. Egan & B. Weatherson (eds.), *Epistemic Modality*. Oxford: Oxford University Press, pp. 144–78.

Turri, J. (forthcoming). “Epistemic Modals and Alternative Possibilities.”