01 Jan 2022

Measurement Effects in Decision-Making

Devin Michael Burns
Missouri University of Science and Technology, burnsde@mst.edu

Charlotte Hohnemann

Recommended Citation

Burns, D. M., & Hohnemann, C. (2022). Measurement Effects in Decision-Making. *Journal of Behavioral Decision Making* Wiley.
The definitive version is available at https://doi.org/10.1002/bdm.2311

Received: 17 January 2022 | Revised: 31 October 2022 | Accepted: 1 November 2022
DOI: 10.1002/bdm.2311

RESEARCH ARTICLE

WILEY

Measurement effects in decision-making

Devin M. Burns[1] | Charlotte Hohnemann[2]

[1]Psychological Science, Missouri University of Science and Technology, Rolla, Missouri, USA
[2]Work and Organizational Psychology, Bergische Universität Wuppertal, Wuppertal, Germany

Correspondence
Devin M. Burns, Missouri S&T, 500 W. 14th St., H-SS 113, Rolla, MO 65409.
Email: burnsde@mst.edu

Abstract

When participants are shown a series of stimuli, their responses differ depending on whether they respond after each stimulus or only at the end of the series, in what we call a measurement effect. These effects have received paltry attention compared with more well-known order effects and pose a unique challenge to theories of decision-making. In a series of two preregistered experiments, we consistently find measurement effects such that responding to a stimulus reduces its impact on later stimuli. While previous research has found such effects in noncumulative tasks, where participants are instructed only to respond to the most recent stimulus, this may be the first demonstration of these effects when participants are asked to combine information across either two or four stimuli. We present modeling results showing that although several extant classical and quantum models fail to predict the direction of these effects, new versions can be created that can do so. Ways in which these effects can be described using either quantum or classical models are discussed, as well as potential connections with other well-known phenomena like the dilution effect.

KEYWORDS
constructive effects, decision-making, dilution, measurement effects, order effects, quantum

1 | INTRODUCTION

It is said that the most trustworthy memory is the one which is never recalled, because memories get reconstructed and somehow tainted every time they are brought to mind. In a similar fashion, opinions and evaluations of facts or arguments can be influenced by the simple act of providing a response. When studying decision-making processes, it can always seem like a good idea to collect more data: If we want to watch the decision-making process unfold, it seems obvious that we should have participants provide responses after each new piece of evidence is presented. However, such repeated measurement can systematically distort the very data we are trying to collect!

The idea that the method by which responses are collected can have a formative impact on the resulting opinions, sometimes referred to as a *measurement* or *constructive* effect, has received attention in more applied fields, especially marketing (Morwitz & Fitzsimons, 2004). The effect has been discussed for measuring attitudes like customer satisfaction and purchase intentions (Morwitz et al., 1993) but has received comparably little attention in more cognitive tasks. Researchers argued that in cases when the critical attitude or response is not existing at the time of measurement, it will be created and directly influenced by this measurement (Feldman & Lynch, 1988). But even when the attitude of interest is already existing, the structure of the measurement can still have an influence on the response (Feldman & Lynch, 1988). With a blind choice paradigm, Sharot et al. (2010) addressed the common criticism that the measurement does not change the attitudes and, instead, only reflects the preferences. Their results supported the existence of measurement effects in affective judgments by overruling alternative explanations.

But a measurement effect can also describe more than just measurement error. It has been debated whether attitudes are constructed every time they are needed or just one time and then recalled (Schwarz, 2007). While the classical approach holds the proposition that it is difficult to change attitudes, the constructive view is based on the assumption that it is difficult to not influence attitudes. In his review, Schwarz provided support for the constructive

https://doi.org/10.1002/bdm.2311

perspective. First, he pointed out that only context-sensitive evaluations are able to adapt to the current situation and guide behavior with regard to actual conditions. Additionally, he explained how temporally constructed attitudes can account for variability when the context is changing, as well as for stability when the context remains the same.

There has been some awareness of these effects in more basic research as well. Hogarth and Einhorn (1992) drew an important distinction between two different data collection methodologies: a Step-by-Step (SbS) task where responses are given after every piece of evidence and an End-of-Sequence (EoS) task where a response is not given until all pieces of evidence have been provided. They noted that in 43 published experiments, those using an EoS response mode tended to show primacy effects (at least for relatively simple arguments), while those using a SbS format overwhelmingly produced recency effects.

Hogarth and Einhorn (1992) proposed that both of these types of tasks can be modeled through a sequential anchoring and adjustment process where the impact of each piece of evidence depends on the current belief state. SbS processing updates belief with every piece of evidence, while EoS processing uses the first piece of evidence to set the anchor and then performs a single adjustment step combining all the remaining pieces of evidence. Hogarth and Einhorn (1992) point out that while SbS tasks require the updating of belief after every piece of evidence by requesting incremental responses, EoS tasks cannot similarly require participants to withhold from updating their beliefs until the end: They may implicitly update their beliefs in an SbS manner, especially if there are many pieces of evidence or they are complex in nature.

1.1 | Quantum models for order and measurement effects

Quantum probability theory has been used as an alternative method for describing the complexities of human decision-making. As quantum probability theory readily includes order effects and measurement effects, it was successfully used to explain these in several experiments. Most studies have focused on the explanation of question order effects (Boyer-Kassem et al., 2016), where the order in which two ostensibly independent questions are asked is shown to have an impact on the responses that people provide. It has been shown that quantum models predict a previously unknown balance between such order effects (regardless of parameterization) referred to as the QQ-equality, and empirical data from more than 70 studies have been shown to conform to these predictions (Wang & Busemeyer, 2013; Wang et al., 2014). It should be noted that subsequent work has demonstrated that a family of classical models produces similar predictions (Kellen et al., 2018).

Kvam et al. (2015) and Busemeyer et al. (2019) have explored measurement effects in a perceptual information accumulation context using the random dot motion task. Participants were asked to provide a fast response to the random dot stimulus and then rate their confidence in the direction of motion after continuing to view it a little longer. This was compared with a condition where the initial response was replaced with a simple motor task (click the mouse). They showed interference from the first response, such that confidence judgments were less extreme following a choice than in the control condition (though direction of motion accuracy remained constant). They presented a quantum random walk model as an alternative to a classical Markov random walk model and showed that it could produce the requisite interference patterns that the Markov model could not (without additional augmentations).

One of the only other attempts to apply quantum models for describing measurement effects has been by White et al. (2014, 2017) with a paradigm where participants used a 9-point scale to indicate how happy various advertisement images made them feel. In one version (experiment 2), either a positively or negatively valenced image was shown, followed by a contrasting image of the opposite valence. In the double-rating condition, the first image was rated and then the second was as well (SbS). However, in the single-rating condition, the presentation was the same but a rating was only requested for the second image (EoS).

Their results from all three experiments revealed that when the second (contrasting) image was of positive valence, final responses were more positive when an intermediate response had been given (the double-rating condition) than if it had not (the single-rating condition). Similarly, if the second stimulus had negative valence, responses were more negative in the double-rating condition. While intuitive theories like "levels of processing" (Hyde & Jenkins, 1969) may predict that the intermediate evaluation would require participants to engage with the first stimulus, increasing its memorability and impact on the second rating, these results show the opposite: Issuing a response to a preliminary stimulus reduced its impact on responses to a subsequent stimulus. Further experimentation bolstered the reliability of these effects and argued that they are not artifacts of differences in timing between the two conditions or the specific response scale used (White et al., 2015).

To help explain this measurement effect, let us first consider what a naive observer may predict in the task: Because the instructions do not ask participants to combine information from the two stimuli (it is not a cumulative task), we may expect that responses to the second stimulus should be independent of the first stimulus. However, the data show this to not be the case. Consistent with decades of research on priming, contrast effects, and so on, responses to the second stimulus are more like a weighted average between the two, with most of the weight going to the second stimulus. While this lack of independence is now unsurprising, the more novel finding was that the impact of the first stimulus was diminished when participants were asked to respond to it and had a stronger effect when they did not.

The measurement effect shown by White et al. (2014) cannot be explained by models of order effects, since order does not vary across the two judgment conditions (though order effects definitely happen in this experiment: Seeing a positive stimulus followed by a negative is different than the reverse). White et al. further allege that the anchoring and adjustment models of Hogarth and Einhorn (1992) are only able to predict such differences between EoS and SbS tasks for tasks with more than two pieces of evidence and that their models are identical when there are only two, though we note that this is only the case under certain assumptions they made, and other parameterizations of Hogarth and Einhorn's model can predict differences between these conditions. Regardless, the third experiment from White et al. (2014) casts further doubt on an anchoring and adjustment explanation, as they did not find evidence for an anchoring effect from experimentally provided judgments and found no significant correlation between responses to the first and second stimuli in any of their experiments, which would have been expected for these models. There is also little theoretical motivation for why the anchoring effect of the first stimulus should diminish when it is responded to: It seems that being asked to assign a numerical response should form a stronger and more concrete anchor rather than the opposite.

White et al. (2020) constructed a quantum model to capture these measurement effects, which they now refer to as Evaluation Bias. Quantum measurement effects can occur because a participant's belief state is treated as existing in a superposition, which is then collapsed or projected onto a specific dimension when they are asked to provide a response. As White et al. describe it in their paper, the first stimulus establishes the initial cognitive state, the second stimulus rotates the belief vector according to its strength and valence, and then the belief vector is projected onto either the positive or negative affect ray in order to provide a response. The novel measurement effect occurs in the SbS condition because the initial cognitive state established by the first stimulus is projected onto the positive or negative affect ray for the intermediate judgment before the rotation for the second stimulus is applied. White et al. (2015) interpret this as an abstraction process, whereby some information about the first stimulus is lost when participants are required to issue a response about the stimulus.

However, it is unclear how this process would lead to the pattern of data they observed. When an initial stimulus is rated positively, the projection onto the positive affect ray should lose information inconsistent with a positive evaluation. This should increase the effect of the first, positive stimulus on the second, negative stimulus, leading the second response to be less negative in the SbS condition than in the EoS single-rating condition (that does not entail an intermediate projection). The data show the opposite pattern, however. White et al. (2014) say that the more positive belief state resulting from the intermediate projection "…would make more obvious the fact that the second advert is negative, leading to a more negative rating," but they fail to provide an impression of how they would construct a model where the second judgment was based on contrast with the preceding belief state. Rather, they state that in their conception, a given stimulus produces "a fixed shift from the current state toward the ray for negative or positive affect," which does not seem consistent with a contrast effect, where this shift would have to be magnified based on the discrepancy between the stimuli. The mechanism by which quantum projection effects should reduce the impact of previous stimuli following measurement has not been made clear.

The experiments by White et al. (2014) are similar to what have sometimes been referred to as studies of *question order*, in that they concern two responses which participants are supposed to consider separately (not cumulatively). This is the type of order effect that gives rise to the QQ-equality (Wang et al., 2014). Less well studied are whether similar measurement effects occur in situations where participants are explicitly supposed to combine subsequent pieces of evidence in a cumulative process, tasks which have been delineated as concerning *information order*. This is the type of task addressed by Hogarth and Einhorn (1992), and Trueblood and Busemeyer (2011) showed that quantum models naturally produce these order effects as well.

To help illustrate the important difference between cumulative and noncumulative tasks, let us return to the expectations of the naive observer: In a cumulative task where two pieces of evidence are supposed to be combined to yield a final judgment, we may expect the two pieces to be equally weighted. However, in most situations, the data violate this expectation by showing recency effects, where the second piece of evidence has larger weight than the first (Trueblood & Busemeyer, 2011). In this situation, it is not clear whether we should expect to see similar measurement effects as found by White et al. (2014): The contrast effect explanation no longer makes sense when the task instructs participants to combine the first and second stimuli. If a measurement increases the extremity of the first belief state, even if contrast effects then cause the second stimulus to also appear more extreme, these effects should cancel out when the two pieces of information are combined. Furthermore, both the previous quantum models of Trueblood and Busemeyer (2011) and the anchoring and adjustment models of Hogarth and Einhorn (1992) predict that a more positive belief state following a first stimulus would lead to a more positive final state after viewing a negative second stimulus, opposite the pattern seen in the noncumulative task used by White et al. (2014).

Yearsley and Pothos (2016) took the idea of information loss through the projections required to issue intermediate responses and pushed it to the extreme in a demonstration of the quantum Zeno effect, where quantum physics predicts that an unstable particle which is continuously observed will not decay due to the continuous influences of measurement. They designed a task where participants made judgments of guilt for a criminal trial in which they were presented with 12 relatively weak pieces of evidence for guilt. The crucial manipulation was how often they were asked to issue a response: One group of participants only issued an EoS judgment after all 12 arguments were presented, while others responded after every six pieces of evidence, every four, three, two, or after each piece of evidence. The more often they were asked to respond, the more the participants resisted changing their opinion, despite having seen the same collection of evidence. They modeled the results using both a Bayesian and a quantum model in which all the parameters were fixed using the data from the first response each participant provided (before any measurement effects could emerge). The remaining data could therefore be fit in a parameter free manner, demonstrating that the superior fit of the quantum model is due to its inherent structure rather than specific parameter fits. Note that the quantum projection here is reducing the impact of inconsistent information: An initial belief in innocence is preserved in the face of weak contradictory evidence when the participant continues to state that belief after every piece of evidence.

Our understanding of how these measurement effects influence judgment and decision-making is still in its infancy. While White et al. (2014) provided intriguing evidence about how these effects can exist even when participants are not attempting to combine pieces of information, the nature of their noncumulative, affective task makes it challenging to compare with previous work like that of Hogarth and Einhorn (1992) and Trueblood and Busemeyer (2011). Similarly, while Yearsley and Pothos (2016) provided a ground-breaking demonstration of the quantum Zeno effect in behavioral data, their paradigm used the unique situation of 12 pieces of evidence that were meticulously constructed to all be of approximately equal strength and all weakly indicating guilt. Little-to-no research has yet investigated the effect of intermediate measurement when conflicting information is integrated to render a final decision, and the mechanisms by which models can predict the pattern of measurement effects that have been found remain elusive.

In this work, we present two experiments studying measurement effects, both using a cumulative framework. Participants were asked to integrate evidence from two (Experiment 1) or four (Experiment 2) arguments to make a final decision, with the crucial manipulation where half the time they only responded once after seeing all arguments in a group and half the time they also provided intermediate responses after every argument.

2 | EXPERIMENT 1

In this study, we focused on measurement effects and their potential interactions with order effects in a cognitive task with high relevance to daily life: deciding whether to make behavioral changes to be more environmentally conscious. Our experimental materials paralleled those of the jury trial in Trueblood and Busemeyer (2011) but with a methodological twist similar to White et al. (2014), where some judgments were made only at the EoS of arguments, what we will call the single-rating condition, and others were made SbS after each argument, which we call the double-rating condition.

Based on previous work by Trueblood and Busemeyer (2011) and others, we expected to find recency effects such that the second piece of evidence in a pair is weighted more heavily than the first. We were especially interested in examining potential measurement effects, as the quantum projection explanation leads us to anticipate the opposite pattern of data that was shown in the noncumulative tasks of White et al. (2014). Such an effect would be seen through a dependency of the order effect on the measurement condition, such that recency is smaller in the double-rating condition than in the single-rating condition, since responding to the first stimulus should strengthen its impact.

2.1 | Participants

Participants were recruited through Amazon's Mechanical Turk platform. Participants were required to have their MTurk Masters qualification (not related to educational attainment), live in the United States, and have an approval rating greater than 90% with more than 500 approved tasks. Participants were randomly assigned to one of eight conditions which differed in terms of the question order. To recruit 30–50 participants in each of the eight versions, 400 slots were available, and 370 responses were collected within the allotted time window. The average completion time was 7.5 min, and participants were compensated with $1.33 for their time, equivalent to an hourly rate above $10. All participants completed an informed consent document prior to participation.

Of the 370 submitted surveys, 25 were incomplete and 2 were second submissions from the same participant, and these were excluded. A further 19 were excluded for taking less than 2 min to complete the survey and 25 for scoring lower than a 3 out of 4 on the attention check question. This left 299 participants for all further analyses. The eight different versions of the survey (same arguments but in different orders) had between 34 and 40 participants each. All exclusion decisions were made before examining any data.

2.2 | Materials

The survey was implemented through Qualtrics following the general design of Experiment 1 from Trueblood and Busemeyer (2011). Scenarios were designed around how likely participants thought an average American would be to make a change in four different domains: adopting a new behavior around the house, purchasing a “greener” appliance, voting for a new local energy policy, or changing their commuting behavior. For each of these domains, there were strong and weak arguments for and against the decision. Participants were told that arguments were left purposefully vague and that they were not supposed to base their responses on any outside information beyond what was explicitly provided.

An initial pilot test was conducted with a separate group of 50 MTurk participants who were asked to rate the strengths of the 16 arguments (4 for each domain). On a 21-point scale (−10 to 10), strong arguments for were rated an average of 6.9, weak arguments for averaged 4.6, weak arguments against were −3.6, and strong arguments against were −5.9. Responses were reasonably consistent across domains, with no specific argument average differing by more than one point from the overall mean for that argument strength.

2.3 | Procedure

Participants were shown eight different pairs of arguments, with each pair containing one argument for and one against each decision. Four of these argument pairs (one for each domain) belonged to a double-rating (SbS) condition where participants were asked to provide an initial response to the first argument and then another response taking both arguments into account. The other four pairs of arguments (again one in each domain) were in the single-rating (EoS) condition, where participants read both arguments and then issued a single response. For ease of discussion, arguments will be denoted using S/W for strong or weak and F/A meaning for or against.

Arguments were arranged such that each participant saw each argument only once, with the order of the arguments and the domains balanced across eight different groups of participants. The decision conditions were blocked such that half of the participants did all the double-rating questions first and the other half did the single-rating questions first. Participants were asked to estimate how likely an average American would be to make a behavioral change in response to the presented arguments using a 21-point scale, with −10 labeled as not adopting a behavior, 10 as adopting it, and 0 labeled neutral. Participants were instructed to not base their responses on any other information they may have outside of the presented arguments. After the survey, the participants completed an attention check question which asked them which of the eight topic domains were featured in the survey.

This procedure yielded a hybrid design where all of the variables were manipulated within-participants (argument strength, response condition, and topic domain) but not in a full factorial manner. For a given topic domain, there are eight possible argument pairs: two strengths of the for argument, two strengths of the against, and two possible orders of which comes first. For each of the four topic domains, each participant saw one argument pairing in a double-rating format (e.g., SF-WA, meaning strong argument "for" and weak argument "against") and then later a second pair using the remaining unseen arguments for that domain in the opposite order and in the single-rating format (e.g., SA-WF). So while every participant sees every argument, other groups of participants would have seen the other potential parings for a given domain (e.g., WF-WA-double-rating). This design prevents potential confounds with participants seeing the same argument twice but does mean that data must be averaged across participants to compare measurement conditions for a given pair of arguments. The different topic domains were not intended as a manipulation of interest but are merely four different sets of arguments.

2.4 | Results

All data, analysis code, and experimental materials are shared through the Open Science Foundation at osf.io/j3gy8.

Figure 1 shows response data averaged across all participants, the four topic domains, and the different argument strengths. We see the expected recency effects, with stimulus pairs ending with an argument for the behavior change being rated higher than those pairs that end with an argument against (all pairs have one of each). Critically, we see that this effect is moderated by measurement condition and is substantially larger when participants provide an intermediate response to the first argument in a pair (the double-measurement condition). This can be thought of as the first argument having a weaker impact on final judgments in this condition. This effect is in the same direction as the one observed by White et al. (2014).

We used a linear mixed effects model to examine the size and significance of these effects. Our main variables of interest were measurement condition (single or double), recency (whether the final argument was "for" or "against"), and the interaction between them, but we also included fixed effects to account for the strength of the "for" argument and the strength of the "against" argument. Finally, random intercepts were fit for each participant and each topic domain (with no random slopes). The reference levels were for the single-rating condition where weak-for was followed by weak-against. The model had a marginal $R^2 = .24$ and conditional $R^2 = .43$. As hypothesized, there was no significant main effect of measurement condition, $t(497) = 1.10$, $p = .27$, $d = .08$; a small significant recency effect in the single-rating condition $t(497) = 2.31$, $p = .02$, $d = .17$; and a significant interaction between recency and measurement, $t(297) = 2.66$,

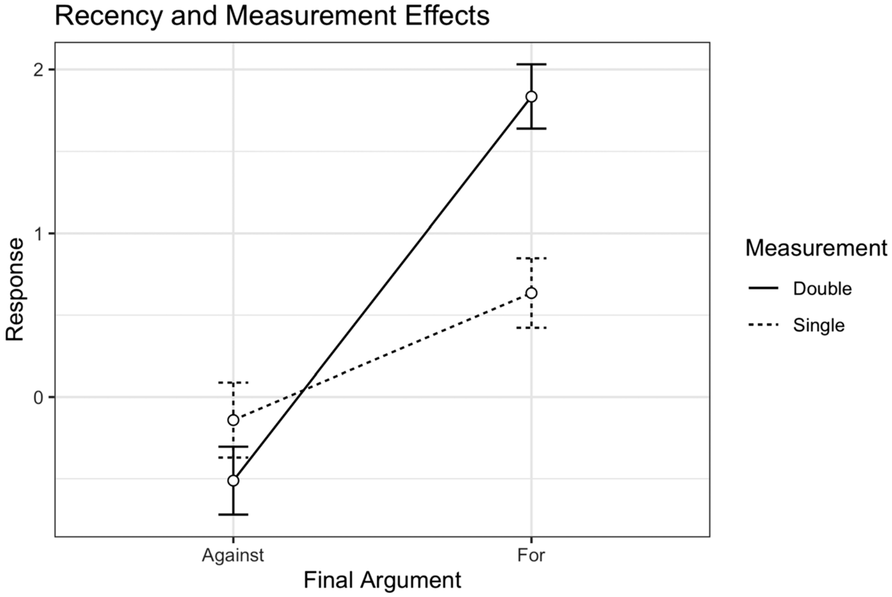

FIGURE 1 Responses averaged across participant, domain, and argument strength. Responses were on a −10 to 10 scale concerning how likely an average American would be to make a "green" decision. We see larger recency effects in the double-measurement condition.

$p = .008$, $d = .51$, resulting in a moderate recency effect for double-ratings. There were also the expected effects of argument strength in both the for ($t(2085) = 14.73$, $p < .001$, $d = .52$) and against ($t(2085) = 26.13$, $p < .001$, $d = .93$) arguments. Estimates of these effects on the original 21-point scale are shown with 95% confidence intervals in Table 1. The primary result of interest is that the recency effect was .78 points for single judgments but jumped to 2.35 points for double judgments ($.78 + 1.57$).

2.5 | Model specification

We fitted four different models to the averaged participant data for the different stimulus combinations to evaluate their ability to fit the results and display appropriate measurement effects. We used the classical adding model (Hogarth & Einhorn, 1992), a quantum model in the style of Trueblood and Busemeyer (2011), and new versions of each with a novel mechanism we designed to produce measurement effects. The measurement effects shown in the data are inconsistent with the idea of a quantum projection that strengthens the effect of the first argument, so we built different conceptualizations designed to reduce the impact of the first argument, as the data show.

The obtained data were scores ranging from -10 to 10 indicating how the participant thought "Joe, an average American" would respond to arguments, specifically whether he would change his behavior or not. The SbS double-rating condition provided data for each of the four arguments as an intermediate response, after reading a single argument, and the eight possible pairs of arguments for the second response (any of the four arguments could be followed by either of the two arguments of opposite valence). The EoS single-rating condition only produced data for the eight pairs, as there were no intermediate judgments in this condition, thus yielding 20 different evaluation conditions in total.

2.5.1 | Adding models

The specific form of the adding model, as defined by Hogarth and Einhorn (1992), depends on whether the evidence is negative or positive and defines the belief state after hearing the k^{tH} piece of evidence as

$$S_k = \begin{cases} S_{k-1} + \alpha * S_{k-1} * s(x_k), & \text{if} s(x_k) \leq 0 \\ S_{k-1} + \beta * (1 - S_{k-1}) * s(x_k), & \text{if} s(x_k) > 0 \end{cases}, \quad (1)$$

where $0 \leq S_{k-1} \leq 1$ is the previous belief state (with S_0 being the prior), $-1 \leq s(x_k) \leq 1$ is the subjective evaluation of the k^{th} piece of evidence, and $0 \leq \alpha, \beta \leq 1$ are constants representing sensitivity to negative and positive evidence, respectively. Following the lead of Trueblood and Busemeyer (2011), we set $\alpha = \beta = 1$ for a more fair comparison with the quantum model, though of course some degree of improved fit would be expected by fitting these additional parameters.

TABLE 1 Estimates and 95% confidence intervals for effects as computed by a linear mixed effects model with random intercepts for each participant and topic domain

Effect	Estimate	95% CI
Intercept (Single, WF-WA)	0.78	(−0.18, 1.74)
Measurement (double)	−0.37	(−1.03, 0.29)
Recency (for)	0.78	(0.12, 1.44)
Strength (SA)	−4.24	(−4.55, −3.92)
Strength (SF)	2.39	(2.07, 2.70)
Measurement (double) x Recency (for)	1.57	(0.41, 2.73)

Note: Numbers are on the −10 to 10 scale used by participants.

Basically, this adding model pushes belief a proportional distance toward either extreme (0 or 1) based on evidence strength. With the strongest possible evidence, $s(x) = \pm 1$ and the belief is changed to 1 or 0 (respectively), while if $s(x) = \pm .5$, then belief is adjusted half of the distance to the relevant extreme. The four parameters to fit to the data are the $s(x_k)$ parameters corresponding to each of the four arguments and are denoted as s_{WF},s_{SF},s_{WA}, and s_{SA}.

For the single-rating condition, where both arguments are read before any response is given, Hogarth and Einhorn suggest that the model should perform only one update step where the two presented arguments are first combined together using a weighted average. We introduced a fifth parameter, $0 \leq w \leq 1$, to account for this weighting (thus allowing this model to show the observed recency effects). The single condition then uses $s(x) = (1 - w) * s(x_1) + w * s(x_2)$ and chooses the appropriate equation above based on the sign of this value.

This method of producing different predictions for the single and double-rating conditions does not parallel the theoretical description of measurement effects from White et al. (2014). Their conception of measurement effects is that rather than changing the way single decision-making works, the double process should be modified such that each evaluation produces a loss of information. This broad framework can be applied to classical as well as quantum models, so we created a new version of the adding model that we call "Add-Forget."

This new model is similar to the above Adding model, but single judgments are made with two updates (the same way double judgments are treated in the adding model) and the weighting parameter is not used. Instead, the fifth parameter for this model, $0 < f < 1$, controls the degree to which previous arguments are "forgotten" when an intermediate response is provided in the double-rating condition. This Add-Forget model is parameterized as above, but after the first argument has been read and the intermediate response S_1 provided, the belief state is updated back toward the initial opinion S_0 before incorporating the second argument, reducing the impact of the first argument:

$$S_1^* = \begin{cases} S_1 - S_0 * s(x_1) * f, & \text{if} s(x_1) \leq 0 \\ S_1 - (1 - S_0) * s(x_1) * f, & \text{if} s(x_1) > 0 \end{cases}, \quad (2)$$

where $f = 1$ would imply full forgetting, and thus, $S_1^* = S_0$, while $f = 0$ implies no forgetting, and $S_1^* = S_1$.

2.5.2 | Quantum models

The base quantum model used in this work is built following the jury trial model from Trueblood and Busemeyer (2011). The full technical description of the model is presented in the Appendix, but here is a conceptual overview of the relevant steps:

1. The initial belief state is represented as a vector in a four-dimensional space.
2. Arguments impact the belief state through a rotation matrix that depends on the valence and salience of the argument.
3. This rotated belief vector is projected to either the positive or negative evidence subspace in accordance with the valence of the argument and is then normalized to maintain unit-length.
4. The intermediate response (for the double-rating condition) can then be predicted by projecting the belief vector into the positive or negative hypothesis subspace according to the valence of the argument.
5. The second argument then rotates the belief vector according to its valence and salience.
6. The belief is now projected to the evidence subspace corresponding to the second argument and normalized.
7. The final response is obtained by projecting to the positive or negative hypothesis subspace according to the valence of the second argument.

We then built a variant of this model along similar lines as the Add-Forget model. In our "Quantum-Forget" model, we changed the way that measurement effects are produced in step 4 (above) for the double-rating condition. As in the quantum model, the belief state is projected to the appropriate hypothesis subspace in order to provide a response to the first argument. However, in contrast to the previous model, this projection does not actually modify the belief state. Instead, the belief state following the first stimulus is weakened by shifting the amount of belief in the positive versus negative hypothesis spaces as dictated by a single new parameter. This weakening happens before the rotation corresponding to the second argument is applied. Although the idea behind this "forgetting" parameter is similar to the Add-Forget model, nonlinearities in the Quantum-Forget model make interpreting the parameter value challenging.

2.5.3 | Fitting procedures

The basic quantum model only uses four parameters, while the other three models each add a fifth parameter: the weighting parameter for the classical adding model and the "forgetting" parameter for the two novel models. All four models assume that participants' initial beliefs start at the same point, which was computed as the average of all intermediate responses (so as to not incorporate measurement effects).

For model fitting, the parameters of each model were grouped into a vector **p**. There were a total of 20 response averages taken from experimental data: from the double-rating condition, we get four intermediate judgments after reading the first argument and eight final judgments after reading the second argument and then an additional eight judgments from the single-rating condition. These 20 responses were grouped into the vector $\mathbf{v}_{\mathbf{data}}$. Each model also predicted 20 responses, and these were grouped into the vector $\mathbf{v}_{\mathbf{predict}}$, whose values depend on the parameters in the model, **p**. Thus, $\mathbf{v}_{\mathbf{predict}}$ can be denoted as $\mathbf{v}_{\mathbf{predict}}(\mathbf{p};\ model)$. The fitting is thus an optimization problem to minimize the cost function $||\mathbf{v}_{\mathbf{data}} - \mathbf{v}_{\mathbf{predict}}||$, where $||\ ||$ is the L2 norm. Both versions of the adding model are linear, and so the fminunc function in Matlab was used to find the minimum. Each parameter of the adding models was searched for in the interval of $(-1, 1)$. The quantum models are highly nonlinear and have many local minima. In order to find the global minimum for the quantum model, we ran the minimization 2000 times with a genetic algorithm. In each run, a random value in the interval of $(-30, 30)$ was assigned as the initial value of each parameter, and the minimization was also performed in $(-30, 30)$. Each run gives a local minimum, and the lowest among them was taken as the global minimum.

2.6 | Model comparison

Participant data and model fits can be seen in Figure 2. The black lines (data) are the same for every facet. The slope of the lines indicate recency effects, and comparing the slope for the solid (double) versus dashed (single) lines shows the measurement effect.

The adding model matches the slopes well but puts the single-measurement responses much higher than the data indicate. Swapping the weighting parameter for our forgetting parameter matches the data more closely, though it slightly overestimates the recency effect in the single-measurement condition. The four-parameter Quantum model shows little differences across the four final response conditions, with almost no recency effects, and the small difference between slopes for the single and double conditions is in the wrong direction. The Quantum-Forget model fares much better but underestimates the recency effect in double-measurement trials.

Summed squared error and Akaike Infomation Criteria (AIC) values for the four models are shown in Table 2. AIC is calculated as $2*k+n*ln(SS\mathbf{e})$, where k is the number of parameters (4 for Quantum, 5 for the others) and n is the number of data points, which is 20 in this case (4 initial and 16 final judgments). The Bayesian Information Criterion (BIC) is calculated similarly but with a larger penalty on the number of parameters: $log(n)*k+n*ln(SSE)$.

AIC and BIC are a more fair comparison between models than raw SSE values since they provide the quantum model an extra allowance for making do with only four parameters compared with the five parameters used in the other models. These values indicate that the original adding and quantum models perform worse than the new versions we developed. The Quantum-Forget model was preferred to the Add-Forget model but either was much better than the other two.

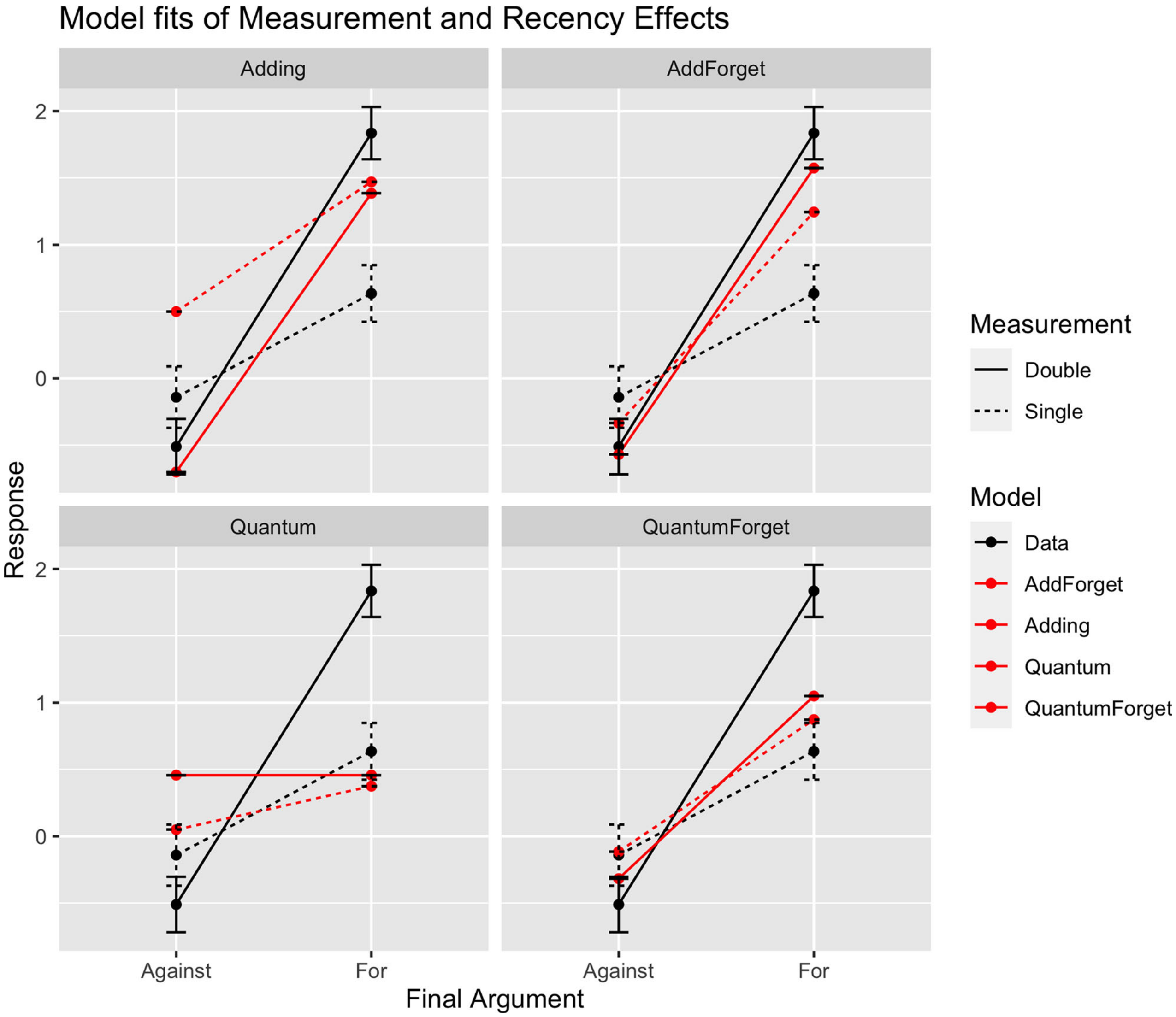

FIGURE 2 Final responses (after seeing two arguments) averaged across participant, topic domain, and argument strength to compare the data with the model fits

TABLE 2 Summed squared error, AIC, and BIC values for the four models of interest

Model	SSE	AIC	BIC
Adding	26.48	75.52	80.51
Add-Forget	9.20	54.39	59.37
Quantum	20.61	68.51	72.49
Quantum-Forget	7.10	49.19	54.17

Note: Lower values indicate a better (or more parsimonious) fit.
Abbreviations: AIC, Akaike Infomation Criteria; BIC, Bayesian Information Criterion.

The relative weight of evidence for one model compared with another can be examined using Bayes factor, which Wagenmakers (2007) showed could be approximated using $\exp((BIC_1 - BIC_2)/2)$. According to this approximation, the Quantum-Forget model had 13.49 as much evidence as the Add-Forget model.

2.7 | Discussion

These results show the same type of measurement effects seen in White et al. (2014): Providing an intermediate response following the first stimulus decreases its impact on the response following the second stimulus. While diminishing the impact of the first stimulus can be seen as desirable in the noncumulative task that they had used (participants were just supposed to respond to the second stimulus), our task explicitly asked participants to combine information across the two arguments for a final judgment. In this case, reducing the weight given to the first stimulus (increasing the recency effect) is likely an undesirable outcome. Finding the same pattern under these opposite task demands could indicate that measurement effects are an inherent feature of decision-making that is difficult to "turn off."

The adding model and quantum models that we used from previous work provided moderate fits to the data, but the quantum model failed to replicate the patterns of recency and measurement effects that we saw in the data. The two modified models we created, where the impact of the first stimulus was reduced after an intermediate judgment was provided, showed much better fits to the data, potentially indicating that this is an important step in the decision-making process.

3 | EXPERIMENT 2

The differences between single-measurement (EoS) and double-measurement (SbS) response conditions should be exaggerated when more arguments are being combined, so we conducted a second experiment using a series of four arguments instead of the pairs used in experiment 1. We used the exact same set of stimuli but put all four arguments for a given topic domain into a question, with participants asked to respond either after each argument (Quadruple-rating) or only after all four (Single).

This design also allowed us to test for dilution effects: When a strong argument or piece of evidence is followed by neutral or weakly consistent evidence, responses sometimes decease in extremity, with a previously strong belief being “diluted” by the additional weaker information. Some of the earliest research into this topic was done by Nisbett et al. (1981), who looked at how probability judgments were influenced by nondiagnostic (irrelevant) information and found that it decreased the influence of the diagnostic information, “watering it down,” so to speak. They interpreted this effect as stemming from the representativeness heuristic: Adding nondiagnostic information reduced the similarity between the target and the outcome, reducing the extremity of the response.

The previously mentioned McKenzie et al. (2002) study used a jury trial paradigm to investigate how evidence is combined, and they found evidence of something akin to an extreme dilution effect (though they did not use that label). They found that when a strong case by the prosecution was followed by a weak defense, belief in guilt often *increased* in reaction to the weak contradictory evidence. They admitted that this effect may not extend to other situations and may depend strongly on the particular adversarial context of a jury trial. There is an assumption that the defense will present the strongest evidence available, so while participants initially may have withheld belief in the prosecution's case (due to the biased source), a flimsy defense can lead to a full embrace.

McKenzie et al. (2002) advocated that this data supported a Minimum Acceptable Strength (MAS) adding model, where arguments weaker than the MAS reference point will have zero or potentially negative effect. However, Trueblood and Busemeyer (2011) reanalyzed their data and showed that a quantum model could also predict the same pattern of data. They also conducted several follow-up experiments based on this jury trial paradigm but using more experimental conditions and more participants. These experiments did not show any cases where belief was adjusted in the opposite direction of the argument, however. No further applications of quantum models to explain the dilution effect appear to have been published, though White et al. (2015) briefly mention that they should be well suited for doing so.

We preregistered this experiment through OSF at https://osf.io/j3gy8/ predicting the same pattern of measurement and recency effects found in experiment 1 in addition to the newly predicted dilution effects.

3.1 | Participants

Participants were recruited in the same manner as experiment 1, with 543 total responses. Of these, 25 were incomplete, 1 was a second submission from the same participant, 42 were submitted in under two minutes (the average duration was 10 min), and 56 had attention check scores less than three out of four. Excluding these left 419 participants for analysis. This experiment had participants grouped by 16 different question orders and had between 23 and 30 usable participants in each group.

3.2 | Procedure

Experiment 2 used the exact same materials as experiment 1, but this time, all four arguments in a given topic domain were presented before final judgments were made. The SbS condition had participants respond after each of the four arguments, and they were explicitly instructed to have each of these judgments be cumulative (e.g., “based on the three arguments you have read so far, please rate your opinion”). In the end of sequence condition, participants only responded after reading all four. Each participant completed only four trials, one for each topic domain, so that no arguments were repeated: two single-rating trials and two quadruple-rating trials, alternating between the two styles, with half of the participants starting with single and the other half with quadruple. The order in which the arguments were presented and the measurement condition used for a particular topic domain was varied across participants, who were randomly assigned into 16 different groups. By having a sequence of four arguments, we could now have trials in which arguments of the same valence (for or against) were blocked together (e.g., WF-SF-SA-WA) or mixed (e.g., WF-SA-SF-WA), and each participant had two topics presented in each fashion. No trials were used in which the first and last arguments were of the same valence (e.g., WF-SA-WA-SF) to try to keep the total number of possible orders tractable.

This design yielded a total of 64 different conditions. In the quadruple-rating condition, there were four possible first arguments, 12 possible ratings following the first pair (each of the four arguments could be followed by any of the other 3), 16 different ratings for sets of three arguments (the four blocked pairs could be followed by the other pair in either order, but the eight mixed pairs can only be followed by a single argument, e.g., SA-WF has to be followed by WA in our design), and finally 16 possible orders of all four. The single-rating condition only provides data for the 16 combinations of all four arguments. Each participant contributed data to 10 of these conditions: four for each of the two quadruple-rating trials and a single data point for each of the two single-rating trials.

3.3 | Results

Figure 3 shows final response data after seeing all four arguments, averaged across all participants, the four topic domains, and different orders of the first three arguments. We again see strong recency effects in the quadruple-measurement condition (responses were higher when the last argument was "for") and a much weaker effect of recency in the single-rating condition.

We again used a linear mixed effects model to examine the size and significance of these effects. Because all four arguments were used every time (in different orders), the model is slightly different than that used for experiment 1. Rather than including variables for the strengths of the for and against argument, we instead have fixed effects for the strength (weak or strong) and valence (for or against) of the argument shown fourth (since recency effects dominate), the measurement condition (single or quad), and all interactions between them. As before, random intercepts were fit for each participant and each topic domain (with no random slopes).

The model had a marginal $R^2 = .11$ and conditional $R^2 = .38$. The reference levels in the model were for a single response where the final argument is weak-against, and this did not differ significantly from the corresponding quad condition, $t(1387) = 0.70$, $p = .48$, $d = .06$. The single condition had a weak but significant recency effect, with responses being higher when the final argument was for rather than against, $t(1427) = 2.20$, $p = .03$, $d = .19$. Single responses did not, however, depend on the strength of the final argument, $t(1504) = 0.12$, $p = .90$, $d = .01$, and there was no significant interaction between the final argument's strength and valence (for or against) for single-ratings, $t(1551) = 0.01$, $p = .99$, $d < .01$. These nonsignificant results are reflected in the almost-flat segments on either side of the single-measurement line in Figure 3, while the recency effect is the central, weakly positively sloped portion.

Significant interactions indicate that the quadruple-rating (SbS) condition showed a moderate effect of the fourth argument's strength $t(1498) = 4.21$, $p < .001$, $d = .52$; a small-to-moderate recency effect for weak arguments, $t(1512) = 1.70$, $p = .09$, $d = .40$; and a large recency effect for strong arguments, $t(1641) = 4.66$, $p < .001$, $d = 1.27$. Estimates of these effects on the original 21-point scale are shown with 95% confidence intervals in Table 3. The primary result of interest is that the recency effect was only .86 for single judgments, regardless of argument strength but increased by .96 to 1.82 for quadruple-ratings ending in a weak argument and by a further 3.99 all the way up to 5.81 for quad-ratings ending in a strong argument.

Tests of the dilution effect were conducted using the responses from the quadruple-rating condition when the first argument was strong and the second argument was weak and of the same valence (either SF-WF or SA-WA). Paired samples t-tests were conducted to compare the first response to the second, within-participants, and topic domains. Participants' initial responses to the strong-for argument were on average 0.92 higher ($sd = 3.54$) than after it was followed with a weak-for argument, $t(110) = 2.76$, $p = .007$, 95% CI$= (.26, 1.60)$, $d = 0.26$. An even stronger difference was found in the against arguments, with initial responses to the strong argument 1.66 lower ($sd = 3.24$) than after it was followed by the consistent weak argument, $t(105) = 5.27$, $p < .001$, 95% CI$= (-2.29, -1.04)$, $d = .51$.

TABLE 3 Estimates and 95% confidence intervals for effects as computed by a linear mixed effects model with random intercepts for each participant and topic domain

Effect	Estimate	95% CI
Intercept (Single, WA)	−1.23	(−2.03, −0.44)
Measurement (Quad)	0.27	(−0.49, 1.03)
Recency (For)	0.86	(0.10, 1.63)
Strength (Strong)	0.05	(−0.75, 0.84)
Recency (For)) x Strength (Strong)	0.01	(−1.12, 1.14)
Measurement (Quad) x Strength (Strong)	−2.41	(−3.53, −1.29)
Measurement (Quad) x Recency (For)	0.96	(−0.14, 2.07)
Measurement (Quad) x Strength (Strong) x Recency (For)	3.99	(2.31, 5.67)

Note: Numbers are on the −10 to 10 scale used by participants.

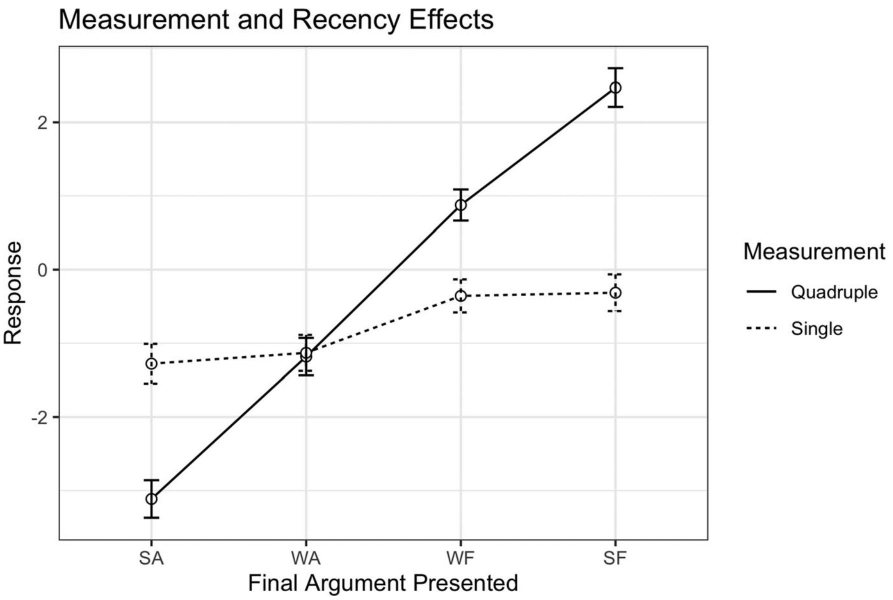

FIGURE 3 Final responses after seeing all four arguments, averaged across participant, domain, and the sequence of the first three arguments

Dilution effects were also found for the final pair of arguments in the quad condition. Participants' responses when the third argument was strong-for were on average 0.89 higher ($sd = 2.82$) than after it was followed with the weak-for argument, $t(110) = 3.33$, $p < .001$, $d = 0.32$, 95% CI= (.36,1.42). Again, a stronger difference was found in the against arguments, with responses when the third argument was strong-against 2.63 lower ($sd = 4.36$) than after hearing the weak-against the fourth argument, $t(103) = 6.16$, $p < .001$, 95% CI= $(-3.48, -1.79)$, $d = .60$.

3.4 | Fitting procedures

We applied the same models and fitting procedures as for experiment 1, continuing to use only five parameters for all models except the basic Quantum model, which only has four. Rather than the 20 different response averages from experiment 1, experiment 2 has 64, as described in the procedure section. The Adding model fits the single-rating condition by creating a weighted average of all four arguments, where the *w* parameter determines the weight of the final argument and all three previous arguments are equally weighted (additional parameters could provide extra flexibility, but such additional flexibility did not justify the additional complexity in model testing). Add-Forget again uses the same formula and SbS updating rule for both rating conditions but has the additional "forgetting" step after each response in quadruple-rating condition (so three times before the final judgments). The basic quantum model distinguishes single and quad through a projection step after each response, while the Quantum-Forget model instead applies a "forgetting" step after each response, similar to Add-Forget.

3.5 | Model comparison

Measurement and recency effects are shown in Figure 4 for the data and each of our four models. As in experiment 1, the Adding model

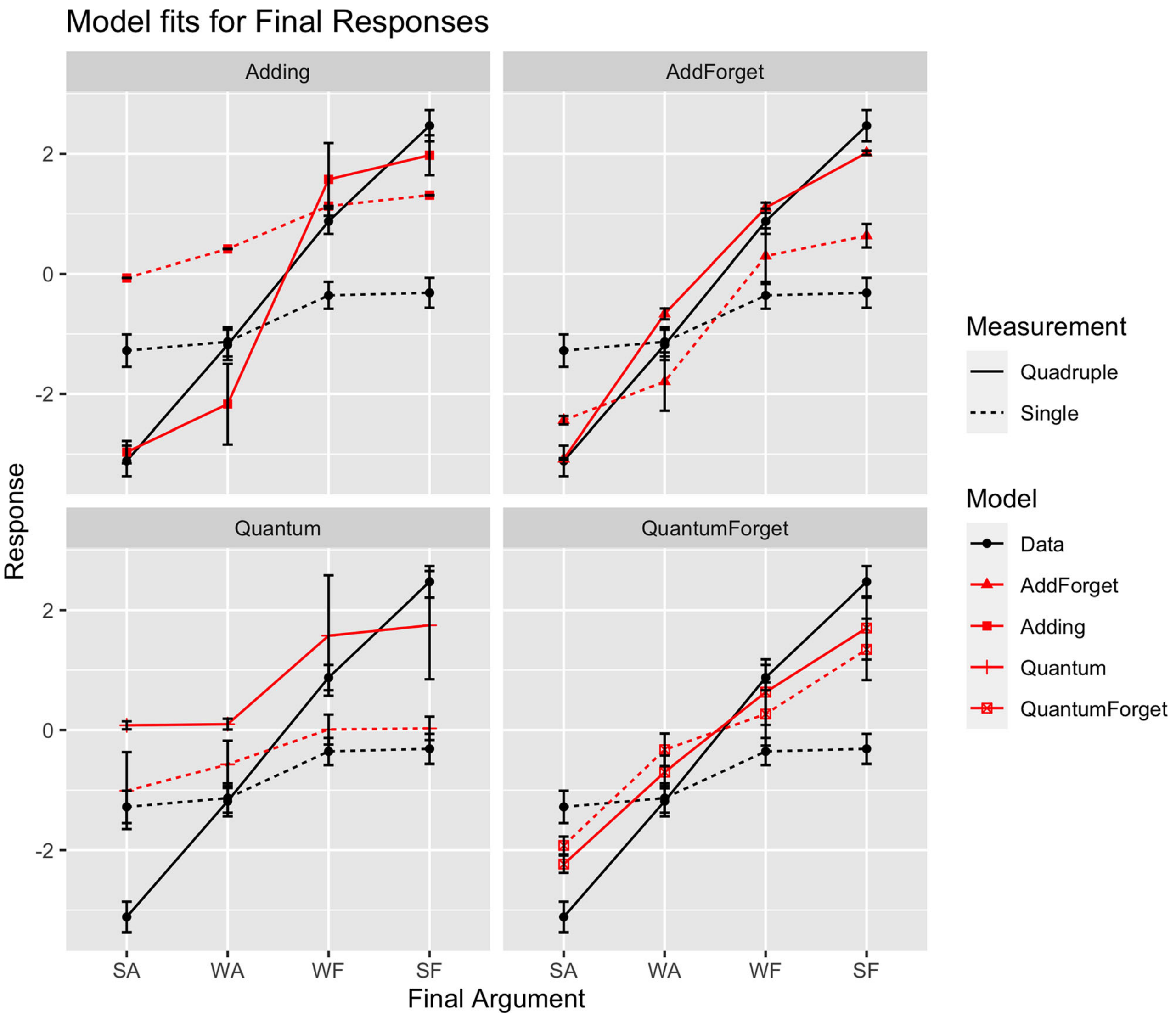

FIGURE 4 Responses averaged across participant, domain, and the sequence of the first three arguments

predicts that responses in the single-response condition would be overall higher than they were but seems to capture other data patterns fairly well, approximately matching the slopes for both conditions, thus capturing recency and measurement effects. The four-parameter Quantum models fit the single-response data reasonably well but was a poor fit for quadruple-response, predicting much weaker effects of recency and expecting quad judgments to be consistently more positive than single. Add-Forget and Quantum-Forget both captured quad data well but predicted larger recency effects for the single condition than were seen in the data, thus underestimating measurement effects.

To illustrate the dilution effects, we present one particular sequence of arguments that produces two such effects: SA-WA-SF-WF. As can be seen in Figure 5, participants' negative views following the strong-against argument lessen when it is followed by the weak-against argument, demonstrating dilution. Similarly, positive responses following the third argument, strong-for, move back toward neutral when the weak-for argument comes afterwards. While the Add-Forget and Quantum-Forget models capture both of these qualitative patterns, the four-parameter Quantum model misses the latter, and the Adding model misses both, instead predicting that the additional consistent information from the weak arguments should strengthen belief.

Fit statistics are shown in Table 4. The four-parameter Quantum model had much higher error than the others, and even after penalizing the others for their higher complexity, it does not come close. This is likely because this model provided a good fit to single-judgment data but poor fit to the quad condition, while the other three showed the reverse pattern. Because the quad condition produces four times as many judgments (and therefore data points) as the single condition, those errors are more costly overall.

The Adding model shows similar overall fit to the Quantum-Forget model, though the latter is slightly preferred. The most significant result, however, is that the Add-Forget model provides a clearly superior fit compared with the other three. Figure 5 shows that this model provides an excellent description of the dilution effects, though Figure 4 does not show notably better performance at capturing measurement and recency effects.

3.6 | Discussion

This second experiment showed the same measurement effects found in experiment 1, with intermittent responses increasing the relative weight of the final argument, producing much stronger recency effects. Similar to the “Quantum Zeno” effect found by Yearsley and Pothos (2016), these repeated measurements appear to interfere with the accumulation of evidence, leading final responses to be based to a greater degree just on the last argument which was presented.

The fact that this paradigm also produced dilution effects offers the tantalizing possibility that it is produced by the same mechanism. If issuing a response decreases the impact of previous stimuli in favor of the most recent, then reading a weak argument after a strong one would be expected to decrease response extremity, as shown in these data. A critical test of this hypothesis that was not available in with our experimental design would be to test if dilution effects diminish or disappear when no response is provided to the first stimulus.

TABLE 4 Summed squared error, AIC, and BIC values for the four models of interest

Model	SSE	AIC	BIC
Adding	123.64	318.31	329.10
Add-Forget	70.62	282.47	293.27
Quantum	304.04	373.90	382.53
Quantum-Forget	113.73	312.96	323.76

Note: Lower values indicate a better (or more parsimonious) fit.
Abbreviations: AIC, Akaike Infomation Criteria; BIC, Bayesian Information Criterion.

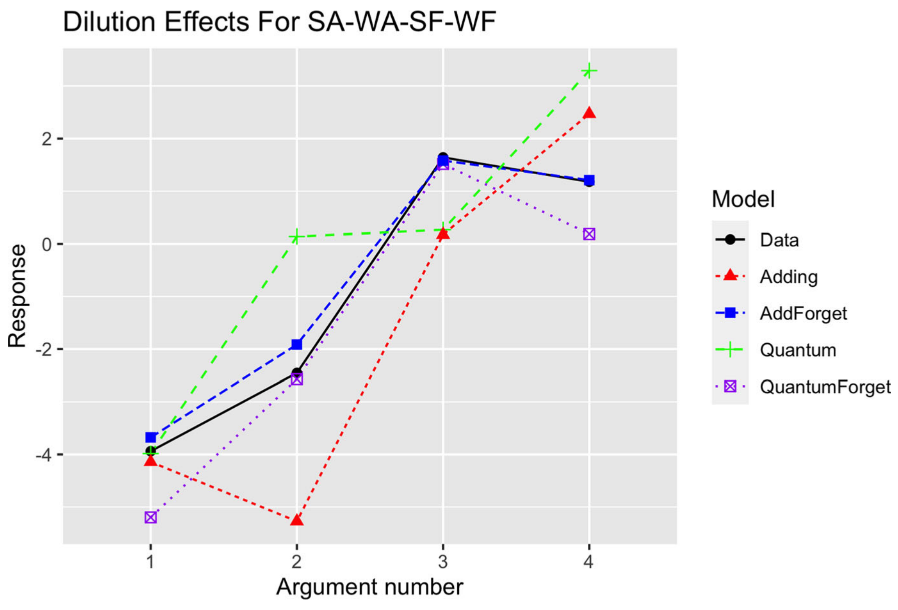

FIGURE 5 Data and model fits for the argument sequence SA-WA-SF-WF to demonstrate dilution effects.

The modeling results showed that the Quantum model from Trueblood and Busemeyer (2011) did not fare well, with a lower quantitative fit than the three others, most notably underestimating the size of recency and measurement effects in the quadruple-rating condition. The Adding model from Hogarth and Einhorn (1992) had a better quantitative fit but was incapable of producing dilution effects. Quantum-Forget provided a similar overall fit as the Adding model but successfully captured the important qualitative patterns in the data. The Add-Forget model was the clear winner for this data, with a near perfect fit of the dilution effects, suggesting that measurement effects could be well modeled by an appropriately constructed classical probability model.

4 | GENERAL DISCUSSION

Simple conceptions of rational decision-making predict that it should make no difference which of a pair of arguments is presented first. However, in agreement with a host of previous investigations, we found recency effects with participants reliably biased in the direction of the final argument they heard. We also found robust evidence of less well-known measurement effects, where final responses depended on whether participants were asked to provide intermediate responses to previous arguments.

Our data showed that recency effects were present for both decision conditions but that these effects were substantially larger in conditions where responses were give after every stimulus (double-/quadruple-rating or SbS) compared with only responding after all stimuli (single-rating or EoS). In our first experiment using sequences of two arguments, the recency effect was only .8 points on a 21-point scale for the single-rating condition but tripled to 2.4 points for double-rating. Experiment 2 used sequences of four arguments and similarly found a recency effect of .9 for single judgments, 1.8 points for quadruple judgments where the final argument was weak, and a whopping 5.8 for quad with a strong final argument. These findings are generally consonant with the data pattern found by White et al. (2014) but extend them in several important ways by showing that similar patterns hold even when using a cumulative task and showing that the size of the effect increases the way we would expect when using larger stimulus sets.

A coherent explanatory mechanism for these measurement effects is still largely lacking. While White and colleagues produced similar measurement effects in their task, they have only used noncumulative tasks and have not postulated whether such effects should be expected to remain similar in a cumulative framework or why. Their explanation for measurement effects is that the process involved in making a response entails a loss of information. When a response is required, a participant's belief is projected down onto the relevant dimension for the response in a nonlinear operation. However, as detailed in Section 1, this mechanism should selectively lose information about the prior stimuli that is *inconsistent* with the response provided, which should increase the impact of this argument instead of lessening it. An important caveat is that the impact of this projection depends on the way in which the model is constructed, and further explorations of such models would be appreciated.

Our second experiment also allowed us to investigate dilution effects within the same task structure. When participants read a strong argument followed by a weak argument of the same valence (both arguments either for or against), their responses following the weak argument were less extreme (more neutral) than they were before it, despite the additional congruent information. This again violates naive expectations, and most existing models of anchoring and adjustment do not produce this pattern. White et al. (2015) claim that quantum models can produce these effects but have not actually demonstrated them doing so.

This raises the question of whether dilution effects can find their explanation in a response-triggered "forgetting" story in a way similar to measurement effects. An important test of this possibility would be to look for dilution effects in an EoS (single-rating) task that should not induce such forgetting. We eagerly await the results of such a test in future experimentation, as such a comparison was not possible with our data set.

It is worth noting an important distinction between the way that White et al. (2020) characterize these measurement effects (or Evaluation Bias, to use their term): Here we refer to the measurement reducing the impact of the previous stimulus, while they refer to it strengthening the impact of the subsequent stimulus. In many situations, these will be indistinguishable, but we have at least two arguments to support our conception. Firstly, if dilution effects can indeed be explained through measurement effects, they are clearly a case of reducing the impact of the first (stronger) argument rather than strengthening the second (weaker but consistent) argument, which would have the opposite result. Secondly, we assert that it makes more sense for measurement to impact a currently held belief rather than having an effect on a future stimulus, for cleaner causality.

These findings also provide reason to reexamine previous investigations of order effects that have used a SbS double-rating procedure, such as Trueblood and Busemeyer (2011) and many others. Since our results show that this procedure enlarges recency effects (presumably by reducing the impact of the first piece of evidence), it is likely that the order effects found in such prior research are larger than they would be with a single-rating procedure and may in some cases overestimate what would be expected in real-world scenarios where intermediate judgments are not mandatory. For example, in a real jury trial, there is no formalized preliminary assessment of the prosecution's case, so this should be treated as an EoS situation.

Although this experiment and others that inspired it have tended to show recency effects, there are other quite similar paradigms where researchers tend to find primacy effects. The body of work on Information Distortion (DeKay, 2015; Kostopoulou et al., 2012) has found that participants stick with an initial opinion formed by early evidence and interpret subsequent evidence in a way that is more consistent with their initial belief. Such paradigms typically use a SbS measurement procedure, and to our knowledge, these researchers have not compared that with single-rating conditions, but it would be interesting to see if measurement effects also strengthen primacy effects where they occur, exacerbating order effects in either direction, or if they continue to diminish earlier evidence, fighting against primacy.

The other main topic requiring further elaboration is what counts as "measurement." Are participants required to issue a numeric response in order for measurement effects to occur? Would such effects be qualitatively different if participants instead provided a free-form linguistic description regarding their impression of the argument? And what if we just ask participants to think about their impression but not provide any response? Does a process of Bayesian belief updating in response to new evidence necessitate such a measurement effect? We cannot always expect them whenever new evidence is incorporated and beliefs are updated, or else we would expect the same results in SbS and EoS conditions, so they seem to only apply when an actual response is required. These questions are important for understanding the existence of measurement effects. It is notable that participants were always asked the same question, to estimate "what Joe would do" following the argument(s), and yet the measurement effect differed following "for" or "against" arguments. This seems to imply that the measurement/projection effects do not depend only on the question asked but to some degree on the valence of the stimulus and/or the participant's reaction to it.

5 | CONCLUSION

In this work, we present robust evidence for the importance of measurement effects. It is counter-intuitive that such effects should consistently reduce the impact of previous stimuli across both cumulative and noncumulative tasks, when participants should desire to forget previous stimuli in the latter but strive not to in the former.

A coherent theoretical explanation for these effects remains elusive, although much important work in this direction has been done by White et al. (2020), who advocate the use of quantum models but show that the classical anchoring and adjustment model from Hogarth and Einhorn (1992) can also capture them. Our work indicates that neither of these models provided a great fit to our data but that augmenting either one with a parameter responsible for "forgetting" previous stimuli each time a response is provided could prove sufficient. Modeling results from experiment 2 show that the classical adding model provided the best fit of any of our models once such a parameter was added, so perhaps quantum probability is unnecessary to capture these effects.

Previous arguments from White et al. (2015) and other quantum advocates have held that although classical models can be augmented with ad hoc parameters to simulate the breadth of behavioral patterns observed, quantum models are to be preferred because they can produce them naturally. This was not the case for this data with the particular instantiation of models we fit, though it should clearly be stated that there are a wide variety of approaches to constructing quantum models and we have only investigated one approach here. According to Trueblood et al. (2017), an alternative quantum model of lower dimensionality may be able to be developed for this application, and it would be expected to show a larger degree of contextual effects. Such characterizations of models which can vary in the amount of assumed incompatibility also show promise in describing individual differences in these effects (Mistry et al., 2018). More investigation is needed, as well as a greater recognition of how measurement effects could be inflating the presence of order effects in many paradigms.

DATA AVAILABILITY STATEMENT

All data, materials, and analysis code are posted on the open science foundation.

ORCID

Devin M. Burns https://orcid.org/0000-0002-2011-9127

REFERENCES

Boyer-Kassem, T., Duchêne, S., & Guerci, E. (2016). Testing quantum-like models of judgment for question order effect. *Mathematical Social Sciences*, *80*, 33–46.

Busemeyer, J.R., Kvam, P.D., & Pleskac, T.J. (2019). Markov versus quantum dynamic models of belief change during evidence monitoring. *Nature Publishing Group*, *9*, 18025.

DeKay, M.L. (2015). Predecisional information distortion and the self-fulfilling prophecy of early preferences in choice. *Current Directions in Psychological Science*, *24*(5), 405–411.

Feldman, J.M., & Lynch, J.G. (1988). Self-generated validity and other effects of measurement on belief, attitude, intention, and behavior. *Journal of Applied Psychology*, *73*(3), 421–435.

Hogarth, R.M., & Einhorn, H.J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55.

Hyde, T.S., & Jenkins, J.J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology*, *82*(3), 472–481.

Kellen, D., Singmann, H., & Batchelder, W.H. (2018). Classic-probability accounts of mirrored (quantum-like) order effects in human judgments. *Decision*, *5*(4), 323–338.

Kostopoulou, O., Russo, J.E., Keenan, G., Delaney, B.C., & Douiri, A. (2012). Information distortion in physicians' diagnostic judgments. *Medical decision making: An international journal of the Society for Medical Decision Making*, *32*(6), 831–839.

Kvam, P.D., Pleskac, T.J., Yu, S., & Busemeyer, J.R. (2015). Interference effects of choice on confidence: Quantum characteristics of evidence accumulation. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(34), 10645–10650.

McKenzie, C. R. M., Lee, S.M., & Chen, K.K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making*, *15*(1), 1–18.

Mistry, P.K., Pothos, E.M., Vandekerckhove, J., & Trueblood, J.S. (2018). A quantum probability account of individual differences in causal reasoning. *Journal of Mathematical Psychology*, *87*, 76–97.

Morwitz, V.G., & Fitzsimons, G.J. (2004). The mere-measurement effect: Why does measuring intentions change actual behavior? *Journal of Consumer Psychology*, *14*(1-2), 64–74.

Morwitz, V.G., Johnson, E., & Schmittlein, D. (1993). Does measuring intent change behavior? *Journal of Consumer Research*, *20*(1), 46–17.

Nisbett, R.E., Zukier, H., & Lemley, R.E. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *13*(2), 248–277.

Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition*, *25*(5), 638–656.

Sharot, T., Velasquez, C.M., & Dolan, R.J. (2010). Do decisions shape preference? *Psychological Science*, *21*(9), 1231–1235.

Trueblood, J.S., & Busemeyer, J.R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, *35*(8), 1518–1552.

Trueblood, J.S., Yearsley, J.M., & Pothos, E.M. (2017). A quantum probability framework for human probabilistic inference. *Journal of Experiment Psychology: General*, *136*(9), 1307.

Wagenmakers, E.J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin and Review*, *14*(5), 779–804.

Wang, Z., & Busemeyer, J.R. (2013). A quantum question order model supported by empirical tests of an a priori and precise prediction. *Topics in Cognitive Science*, *53*(4), n/a–n/a.

Wang, Z., Solloway, T., Shiffrin, R.M., & Busemeyer, J.R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(26), 9431–9436.

White, L.C., Barque-Duran, A., & Pothos, E.M. (2015). An investigation of a quantum probability model for the constructive effect of affective evaluation. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2058), 20150142–15.

White, L.C., Pothos, E.M., & Busemeyer, J.R. (2014). Sometimes it does hurt to ask: The constructive role of articulating impressions. *Cognition*, *133*(1), 48–64.

White, L.C., Pothos, E.M., & Busemeyer, J.R. (2015). Insights from quantum cognitive models for organizational decision making. *Journal of Applied Research in Memory and Cognition*, *4*(3), 229–238.

White, L.C., Pothos, E.M., & Busemeyer, J.R. (2017). A quantum probability model for the constructive influence of affective evaluation, *The palgrave handbook of quantum models in social science*. London: Palgrave Macmillan UK, pp. 267–291.

White, L.C., Pothos, E.M., & Jarrett, M. (2020). The cost of asking: How evaluations bias subsequent judgments. *Decision*, *7*(4), 259–286.

Yearsley, J.M., & Pothos, E.M. (2016). Zeno's paradox in decision-making. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *283*(1828), 20160291.

AUTHOR BIOGRAPHIES

Dr. Devin M. Burns received his Ph.D. in Cognitive Psychology at Indiana University and is now an Associate Professor of Psychological Science at Missouri S&T. He researches Decision Making, Rationality, and Augmented Perception.

M.Sc. Charlotte Hohnemann is a Ph.D. student and researcher in work, organizational, and business psychology at the University of Wuppertal, Germany. She applies the latest findings about cognitive processes, performance, and well-being in the development and implementation of work-related interventions.

How to cite this article: Burns, D. M., & Hohnemann, C. (2022). Measurement effects in decision-making. *Journal of Behavioral Decision Making*, e2311. https://doi.org/10.1002/bdm.2311

APPENDIX A: MODELING DETAILS

For the quantum model, there are two hypotheses and two kinds of evidence. The two hypotheses are $|\mathbf{h}_1\rangle$, Joe will adopt the change, and $|\mathbf{h}_2\rangle$, he will not. The two kinds of evidence are positive evidence $|\mathbf{e}_1\rangle$ (both of the for arguments) and negative evidence $|\mathbf{e}_2\rangle$ (both of the against arguments). The hypotheses and evidence can be represented as two-dimensional basis vectors:

$$|\mathbf{h}_1\rangle = \begin{bmatrix}1\\0\end{bmatrix}, |\mathbf{h}_2\rangle = \begin{bmatrix}0\\1\end{bmatrix}, |\mathbf{e}_1\rangle = \begin{bmatrix}1\\0\end{bmatrix}, |\mathbf{e}_2\rangle = \begin{bmatrix}0\\1\end{bmatrix}, \tag{A1}$$

using the ket notation $|\cdot\rangle$ commonly used in quantum mechanics to represent a column state vector. The belief state is represented as a superposition state:

$$|\boldsymbol{\psi}\rangle = \sum_{i,j\in\{1,2\}} \boldsymbol{\omega}_{ij}|\mathbf{N}_{ij}\rangle. \tag{A2}$$

This is a vector that takes complex values, with $|\mathbf{N}_{ij}\rangle$ being the basis state. As an event, $\mathbf{N}_{ij}$ is the intersection of the participant being exposed to evidence $\mathbf{e}_j$ and adopting the hypothesis $\mathbf{h}_i$, that is, $\mathbf{N}_{ij} = \mathbf{h}_i \cap \mathbf{e}_j$. For example, $\mathbf{N}_{11}$ means that the participant decides that Joe will adopt the change after reading a "for" argument. As a vector, $|\mathbf{N}_{ij}\rangle$ is the Kronecker product of $|\mathbf{h}_i\rangle$ and $|\mathbf{e}_j\rangle$, that is,

$$|\mathbf{N}_{ij}\rangle = |\mathbf{h}_i\rangle \bigotimes |\mathbf{e}_j\rangle. \tag{A3}$$

The coefficient $\boldsymbol{\omega}_{ij}$ is the amplitude of $|\mathbf{N}_{ij}\rangle$. In quantum mechanics, when the superposition state is measured, it will collapse onto one of its basis states. When $|\boldsymbol{\psi}\rangle$ is measured, the probability of getting $|\mathbf{N}_{ij}\rangle$ is $|\boldsymbol{\omega}_{ij}|^2$. Therefore,

$$\sum_{i,j\in\{1,2\}} |\boldsymbol{\omega}_{ij}|^2 = 1. \tag{A4}$$

In this work, the prior (initial belief in Joe's probability to adopt the change) was not directly measured (to prevent additional measurement effects!) and instead is estimated empirically as the average of the four different intermediate responses:

$$\mathbf{P}_{ini} = [\Pr(WF) + \Pr(WA) + \Pr(SF) + \Pr(SA)]/4, \tag{A5}$$

where $\Pr(WF)$ is the average response across participants after seeing the "weak-for" argument, converted from the -10 to 10 scale used into a 0 to 1 scale for computational convenience, and the other three probabilities are defined in similar ways. The initial belief state is then set as

$$|\boldsymbol{\psi}\rangle = \begin{bmatrix} \sqrt{\mathbf{P}_{ini}/2} \\ \sqrt{\mathbf{P}_{ini}/2} \\ \sqrt{(1-\mathbf{P}_{ini})/2} \\ \sqrt{(1-\mathbf{P}_{ini})/2} \end{bmatrix}. \tag{A6}$$

In our quantum cognition model, making a decision is modeled by measuring the superposition belief state of the decision maker. The decision is just the basis state to which the superposition state collapses. The measurement is performed with the measurement operator **M**:

$$|\boldsymbol{\psi}\rangle \rightarrow \frac{\mathbf{M}|\boldsymbol{\psi}\rangle}{\sqrt{\langle\boldsymbol{\psi}|\mathbf{M}^{\dagger}\mathbf{M}|\boldsymbol{\psi}\rangle}}. \tag{A7}$$

M can be represented by a matrix and $\mathbf{M}^{\dagger}$ is the transpose conjugate of **M**. Measurement occurs via a family of projectors that sum to 1. For example, applying the measurement operator regarding the basis of $\mathbf{h}_1 \cap \mathbf{e}_1$ to $|\boldsymbol{\psi}\rangle$ leads to

$$\mathbf{M}_{11}|\boldsymbol{\psi}\rangle = \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_{11} \\ \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} \\ \boldsymbol{\omega}_{22} \end{bmatrix}. \tag{A8}$$

The outcome of the measurement is thus

$$\frac{\mathbf{M}_{11}|\boldsymbol{\psi}\rangle}{\sqrt{\langle\boldsymbol{\psi}|\mathbf{M}_{11}^{\dagger}\mathbf{M}_{11}|\boldsymbol{\psi}\rangle}} = \frac{1}{\sqrt{\boldsymbol{\omega}_{11}^{*}\boldsymbol{\omega}_{11}}} \begin{bmatrix} \boldsymbol{\omega}_{11} \\ 0 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{|\boldsymbol{\omega}_{11}|} \begin{bmatrix} \boldsymbol{\omega}_{11} \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{A9}$$

When $\boldsymbol{\omega}_{11}$ is a positive real number, the subsequent state is just $[1\ 0\ 0\ 0]^T$. Similarly, the result of measurement regarding $|\mathbf{h}_1\rangle$ is

$$\begin{aligned} \frac{\mathbf{M}_{\mathbf{h}_1}|\boldsymbol{\psi}\rangle}{\sqrt{\langle\boldsymbol{\psi}|\mathbf{M}_{\mathbf{h}_1}^{\dagger}\mathbf{M}_{\mathbf{h}_1}|\boldsymbol{\psi}\rangle}} &= \frac{1}{\sqrt{\langle\boldsymbol{\psi}|\mathbf{M}_{\mathbf{h}_1}^{\dagger}\mathbf{M}_{\mathbf{h}_1}|\boldsymbol{\psi}\rangle}} \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 0 & \\ & & & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\omega}_{11} \\ \boldsymbol{\omega}_{12} \\ \boldsymbol{\omega}_{21} \\ \boldsymbol{\omega}_{22} \end{bmatrix} \\ &= \frac{1}{\sqrt{|\boldsymbol{\omega}_{11}|^2 + |\boldsymbol{\omega}_{12}|^2}} \begin{bmatrix} \boldsymbol{\omega}_{11} \\ \boldsymbol{\omega}_{12} \\ 0 \\ 0 \end{bmatrix}. \end{aligned} \tag{A10}$$

In quantum probability theory, some events are treated as *incompatible*, with their joint probability space undefined, relying instead on order dependent conditional probabilities. Recent work (Trueblood et al., 2017) has examined the effects of varying the degree of incompatibility using different quantum models and has found that more incompatibility produces more of the context-dependent effects quantum models are often chosen for. In this work, the events regarding the four types of arguments (SA, SF, WA, WF) are considered to be incompatible. The same belief state should be described in different bases when considering incompatible events. Such different bases can be considered as corresponding to different points of view. Let $|\mathbf{N}_{ij}\rangle$ still be the basis for the event $\mathbf{h}_i \cap \mathbf{e}_j$ and let $|A_{ij}\rangle$ and $|B_{ij}\rangle$ be the bases corresponding to the SA and SF arguments (different bases correspond to WA and WF). Then the same quantum belief state of the participant can be expressed with respect to any one of these three points of view:

$$|\boldsymbol{\psi}\rangle = \sum \boldsymbol{\omega}_{ij}|\mathbf{N}_{ij}\rangle = \sum \boldsymbol{\alpha}_{ij}|A_{ij}\rangle = \sum \boldsymbol{\beta}_{ij}|B_{ij}\rangle. \tag{A11}$$

where $\boldsymbol{\alpha}_{ij}$ and $\boldsymbol{\beta}_{ij}$ are still in the sense of $\mathbf{h}_i \cap \mathbf{e}_j$ just like $\boldsymbol{\omega}_{ij}$.

When incompatible events are concerned, measurement cannot be applied directly. The basis of the belief state must be changed to that which corresponds to that event. The basis is changed by multiplying the state vector with a unitary operator **U** ($\mathbf{U}^{\dagger}\mathbf{U} = \mathbf{U}\mathbf{U}^{\dagger} = I$, where I is the identity matrix). The belief state can be changed to the SA basis $|A_{ij}\rangle$ from the initial basis $|\mathbf{N}_{ij}\rangle$ via $\mathbf{U}_{\mathrm{AN}}$, such that $|\boldsymbol{\alpha}\rangle = \mathbf{U}_{\mathrm{AN}}|\boldsymbol{\omega}\rangle$, and can similarly be changed back to the initial basis via $\mathbf{U}_{\mathrm{AN}}^{\dagger}$. While this change of basis is reversible, the unitary rotation described here refers only to a coordinate change and not to an actual change of the state vector. In contrast, the measurement operator is not unitary, and so these operators do not commute, which is why we see measurement effects.

A.1 | Application of the quantum model

Let us consider the double judgment case where the participant reads the SF argument first, provides a response, and then reads the SA argument and is asked to respond to both. The quantum model describes this situation as follows:

1. Change basis to the SF point of view: Let the belief state in the initial basis be $|\boldsymbol{\omega}\rangle$. When the participant is exposed to the SF argument, the basis of the belief state is changed by the unitary operator $\mathbf{U}_{\mathbf{N}\mapsto SF}$ so that the same belief state in the new basis is denoted by $|\boldsymbol{\alpha}\rangle$:

$$|\boldsymbol{\alpha}\rangle = \mathbf{U}_{\mathbf{N}\mapsto SF}|\boldsymbol{\omega}\rangle. \tag{A12}$$

2. Project to positive evidence subspace: When the participant considers a "for" argument, the belief state is projected to the positive evidence subspace $|\mathbf{e}_1\rangle$ by the measurement operator $\mathbf{M}_{\mathbf{e}_1}$ (because the SF argument contains positive evidence):

$$\mathbf{M}_{\mathbf{e}_1}|\boldsymbol{\alpha}\rangle = \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & 1 & \\ & & & 0 \end{bmatrix} |\boldsymbol{\alpha}\rangle. \tag{A13}$$

3. Normalize: The obtained new belief state is defined as $|\boldsymbol{\alpha}_{\mathbf{e}_1}\rangle$:

$$|\boldsymbol{\alpha}_{\mathbf{e}_1}\rangle = \frac{\mathbf{M}_{\mathbf{e}_1}|\boldsymbol{\alpha}\rangle}{\sqrt{\langle\boldsymbol{\alpha}|\mathbf{M}_{\mathbf{e}_1}^{\dagger}\mathbf{M}_{\mathbf{e}_1}|\boldsymbol{\alpha}\rangle}}. \tag{A14}$$

4. Predict the first judgment in double-rating condition: The rated probability that Joe will adopt the change at this point is $\langle\boldsymbol{\alpha}_{\mathbf{e}_1}|\mathbf{M}_{\mathbf{H}_1}^{\dagger}\mathbf{M}_{\mathbf{H}_1}|\boldsymbol{\alpha}_{\mathbf{e}_1}\rangle$, where $\mathbf{M}_{\mathbf{h}_1}$ is the measurement operator that projects the belief state to the positive hypothesis subspace $|\mathbf{h}_1\rangle$:

$$\mathbf{M}_{\mathbf{h}_1} = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 0 & \\ & & & 0 \end{bmatrix}. \tag{A15}$$

This is where the measurement effect occurs, as this step is not done in the single-rating condition where participants do not respond to the first argument.

5. Change basis to the SA point of view: When the participant is exposed to the SA argument, the basis of the belief state is changed by the unitary operator $\mathbf{U}_{\mathbf{N}\mapsto SA}$. Before that, we need to change back to the initial basis from the SF basis. The same belief state in the new basis is denoted by the following:

$$|\boldsymbol{\beta}\rangle = \mathbf{U}_{\mathbf{N}\mapsto SA}\mathbf{U}^{\dagger}_{\mathbf{N}\mapsto SF}\frac{\mathbf{M}_{\mathbf{h}_1}|\boldsymbol{\alpha}_{\mathbf{e}_1}\rangle}{\sqrt{\langle\boldsymbol{\alpha}_{\mathbf{e}_1}|\mathbf{M}^{\dagger}_{\mathbf{h}_1}\mathbf{M}_{\mathbf{h}_1}|\boldsymbol{\alpha}_{\mathbf{e}_1}\rangle}}. \tag{A16}$$

Note that in the single-rating condition, there is no measurement step following the first stimulus, so this change of basis would just be $|\boldsymbol{\beta}\rangle = \mathbf{U}_{\mathbf{N}\mapsto SA}\mathbf{U}^{\dagger}_{\mathbf{N}\mapsto SF}|\boldsymbol{\alpha}_{\mathbf{e}_1}\rangle$.

6. Project to negative evidence subspace: When the participant considers this second argument, the belief state is projected to the negative evidence subspace $|\mathbf{e}_2\rangle$ by the measurement operator $\mathbf{M}_{\mathbf{e}_2}$ (because the SA argument contains negative evidence):

$$\mathbf{M}_{\mathbf{e}_2}|\boldsymbol{\beta}\rangle = \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & 0 & \\ & & & 1 \end{bmatrix}|\boldsymbol{\beta}\rangle. \tag{A17}$$

7. Normalize: The obtained new belief state is $|\boldsymbol{\beta}_{\mathbf{e}_2}\rangle$ as follows:

$$|\boldsymbol{\beta}_{\mathbf{e}_2}\rangle = \frac{\mathbf{M}_{\mathbf{e}_2}|\boldsymbol{\beta}\rangle}{\sqrt{\langle\boldsymbol{\beta}|\mathbf{M}^{\dagger}_{\mathbf{e}_2}\mathbf{M}_{\mathbf{e}_2}|\boldsymbol{\beta}\rangle}}. \tag{A18}$$

8. Predict the final judgment. The participant's rating that Joe will adopt the change at this point is $\langle\boldsymbol{\beta}_{\mathbf{e}_2}|\mathbf{M}^{\dagger}_{\mathbf{H}_1}\mathbf{M}_{\mathbf{h}_1}|\boldsymbol{\beta}_{\mathbf{e}_2}\rangle$.

Predictions can be made for other combinations of arguments in a similar fashion. In order to predict all probability ratings, the four unitary matrices are needed, namely, $\mathbf{U}_{\mathbf{N}\mapsto SA}$, $\mathbf{U}_{\mathbf{N}\mapsto WA}$, $\mathbf{U}_{\mathbf{N}\mapsto WF}$, and $\mathbf{U}_{\mathbf{N}\mapsto SF}$. They can be considered parameters of the quantum cognition model and should be found by fitting experimental data. However, these matrices are all in the dimension of 4×4 and thus each have 16 elements. In order to reduce the number of parameters, the following property of unitary matrices is utilized: any unitary matrix $\mathbf{U}$ can be constructed from a Hermitian matrix $\mathbf{H}$ as $\mathbf{U} = \mathbf{e}^{-i\theta\mathbf{H}}$.

Note that a Hermitian matrix satisfies $\mathbf{H}^{\dagger} = \mathbf{H}$ and i is the imaginary unit. If all four unitary matrices share a common H, then there are only four scalar parameters to fit, namely, θ_{SA}, θ_{WA}, θ_{WF}, and θ_{SF}. Then the remaining question is which $\mathbf{H}$ to use. In this work, we use the proposal from Trueblood and Busemeyer (2011), where $\mathbf{H}_1$ is chosen to strengthen and weaken evidence amplitudes to the greatest extent possible, while $\mathbf{H}_2$ is chosen to function similarly on hypothesis amplitudes:

$$\begin{aligned}
\mathbf{H} &= \frac{1}{\sqrt{2}}(\mathbf{H}''_1 + \mathbf{H}''_2) \\
&= \frac{1}{\sqrt{2}}\left(\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}\right) \\
&= \frac{1}{\sqrt{2}}\begin{bmatrix} 2 & 1 & 1 & 0 \\ 1 & -2 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{bmatrix}.
\end{aligned} \tag{A19}$$

We then built a variant of this model along similar lines as the Add-Forget model. In our "Quantum-Forget" model, we changed the way that measurement effects are produced in step 4 (above) for the double-rating condition. As in the quantum model, the belief state state is projected to the appropriate hypothesis subspace in order to provide a response to the first argument. However, in contrast to the previous model, this projection does not actually modify the belief state. Instead, the belief state following the first stimulus is weakened by shifting the amount of belief in the positive versus negative hypothesis spaces as dictated by a single parameter, $0 < f < 1$, which is used to define a diagonal matrix which takes the form $\mathrm{diag}([\sqrt{f}, \sqrt{f}, \sqrt{1-f}, \sqrt{1-f}])$ when the first argument supported the positive hypothesis, $\mathbf{H}_1$, and $\mathrm{diag}([\sqrt{1-f}, \sqrt{1-f}, \sqrt{f}, \sqrt{f}])$ if it supported $\mathbf{H}_2$. The belief state following the initial response in the double-rating condition is multiplied by this matrix before the rotation corresponding to the second argument is applied. Although the idea behind this "forgetting" parameter is similar to the Add-Forget model, nonlinearities in the Quantum-Forget model make interpreting the parameter value challenging.