Cognitive reflection and Normality Identities: two new benchmarks for models of probability judgments

Jiaqi Huang (huangajq@iu.edu)[1], **Jerome Busemeyer (jbusemey@indiana.edu)**[1]
& **Zo Ebelt (Zo.Ebelt@city.ac.uk)**[2], **& Emmanuel M. Pothos (emmanuel.pothos.1@city.ac.uk)**[2]
[1] Department of Cognitive Science, Indiana University, 1001 E. 10th Street, Bloomington, IN 47405 USA,
[2] Department of Psychology, City, University of London, 32-38 Whiskin Street, London, EC1R 0JD, United Kingdom

Abstract

We propose two novel benchmarks for assessing models of probability judgments: the impact of Cognitive Reflection Test (CRT) on probability judgment expressions and 16 "normality identities" expected to sum to 1 under classical probability theory. We compared three models on these benchmarks – the Probability Plus Noise Model (PPN), the Bayesian Sampler (BS), and the Quantum Sequential Sampler (QSS) – using the largest dataset to date for probability judgments. Our results reveal that higher CRT scores are associated with fewer probabilistic fallacies and identity violations, a trend most accurately captured by the QSS, although we also identified QSS limitations. Regarding the normality identities, the QSS outperformed the PPN and the BS, which had difficulty with both the average values of the normality identities and their dependence on CRT scores. Additionally, we uncovered a unique "1 crossing" effect for normality identities N8 and N11, an effect PPN and BS cannot capture.

Keywords: probability judgment, probabilistic fallacies, sampling, cognitive reflection

Introduction

Research on probabilistic reasoning has consistently driven theoretical innovation (Oaksford & Chater, 2007; Griffiths et al., 2010; Sanborn et al., 2013), and produced some of the most evocative empirical findings in psychology, including the conjunction fallacy (Tversky & Kahneman, 1983), and the unpacking effect (Tversky & Koehler, 1994). In the last 15 years, the focus of research has shifted from analyzing specific probabilistic fallacies to computational models that offer extensive explanations for a broad range of known fallacies (Costello & Watts, 2014; Zhu et al., 2020; Huang et al., 2023; Huang, 2023). The increasing sophistication of these models necessitates novel approaches for model comparisons; merely evaluating models based on traditional fallacies invariably proves insufficient. In this work, we propose two innovative benchmarks to investigate models of probabilistic fallacies, utilizing a recently compiled dataset of probability judgments in the context of the 2020 US Presidential election, which we will call the 'elections dataset' (Huang et al., 2023). Our analysis indicates that none of the currently established models are completely adequate, though the three models exhibit varying levels of success.

The first benchmark examines how accurately each model predicts the relationship between individual characteristics in reasoning and probabilistic fallacies. Kahneman (2011) popularized the concept of two systems of reasoning: System 1, which is intuitive, prone to errors, and reflexive; and System 2, which is more deliberate, analytical, and reflective. This "two system theory" is widely supported in the literature (Kahneman, 2002; Fernbach & Sloman, 2009; Elqayam & Evans, 2013). Individual differences in the preference for System 1 versus System 2 reasoning in judgment are reasonably well-documented (Kahneman, 2011), with the cognitive reflection test (CRT) (Toplak et al., 2011, 2014) being a widely used measure. CRT involves three arithmetic questions, with a higher CRT score (higher accuracy) suggesting stronger tendency to use System 2 over System 1 reasoning. Despite ongoing debate regarding its construct validity and the criticism that it is overused in online studies (Welsh et al., 2013; Pennycook et al., 2016; Campitelli & Gerrans, 2014), CRT remains one of the most reliable tools for assessing individual differences in "two-system reasoning" (Stagnaro et al., 2018; Bialek & Pennycook, 2018). Prior research has established significant links between CRT scores and various key probabilistic fallacies. For example, Huang et al. (2023) reported a significant anti-correlation between conjunction and disjunction fallacy rates and CRT scores. Similar results for other fallacies, such as question order effects and subadditivity effects, have also been reported in Yearsley & Trueblood (2018) and Trueblood et al. (2017). Motivated by these findings, our study further explores the dependency of previously unexplored probabilistic fallacies on CRT scores, such as violations of binary complementarity (Epping & Busemeyer, 2023) and violations of probability identities (see just below) first reported in Costello & Watts (2014). This exploration presents a significant, untapped modeling opportunity: can current models accurately capture the relationship between CRT scores and the emergence of these fallacies?

Our second innovation involves proposing a novel set of probability identities for evaluating computational models of probability judgments. Initially introduced by Costello & Watts (2014) and expanded in their 2016 work, probability identities are combinations of probabilities that should, according to classical theory, equal zero. While their study highlighted how certain probability combinations can cancel each other out (e.g., $Z1$ shows that $P(A)+P(B)-P(A\cap B) = P(A\cup B)$), Costello and Watts' identities did not involve tests that conformed to specific values that uphold normality (e.g., $P(A)+P(B)-P(A\cap B) = 1-P(\neg A\cap\neg B)$). We therefore extend the set of probability identities to include those that ex-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Name	Normality Identity (= 1)	PPN	QSS
N1	$P(A)+P(\neg A\mid B)P(B)+P(\neg A\mid\neg B)P(\neg B)$	1	$\approx 1+k+$ Z11
N2	$P(A\mid B)P(B)+P(A\mid\neg B)P(\neg B)+P(\neg A\mid B)P(B)+P(\neg A\mid\neg B)P(\neg B)$	1	$\approx 1+k+2*$Z11
N3	$P(A\wedge B)+P(A\wedge\neg B)+P(\neg A\wedge B)+P(\neg A\wedge\neg B)$	$\approx 1+2(d+\Delta d)$	$\approx 1+k+2*$Z5
N4	$P(A\mid B)P(B)+P(A\wedge\neg B)+P(\neg A\wedge B)+P(\neg A\wedge\neg B)$	$\approx$ N3 $-0.5*d$	$\approx$ N3 − Z15
N5	$P(A\mid B)P(B)+P(\neg A\mid B)P(B)+P(A\wedge\neg B)+P(\neg A\wedge\neg B)$	$\approx$ N3 $-d$	$\approx$ N3 − 2*Z15
N6	$P(A\mid B)P(B)+P(B\mid\neg A)P(\neg A)+P(A\wedge\neg B)+P(\neg A\wedge\neg B)$	$\approx$ N5	$\approx$ N5
N7	$P(A\wedge B)+P(A\mid\neg B)P(\neg B)+P(\neg A\mid B)P(B)+P(\neg A\mid\neg B)P(\neg B)$	$\approx$ N5 $-0.5*d$	$\approx$ N5 − Z15
N8	$P(A\vee B)+P(\neg A\mid\neg B)P(\neg B)$	$\approx 1-0.5*d$	$\approx 1+k-$ Z15
N9	$P(A)+P(B)-P(A\wedge B)+P(\neg A\wedge\neg B)$	1	$\approx 1+k+$ Z1
N10	$P(A)+P(B)-P(A\mid B)P(B)+P(\neg A\wedge\neg B)$	$\approx 1+0.5*d$	$\approx$ N9 − Z15
N11	$P(A)+P(B)-P(A\wedge B)+P(\neg A\mid\neg B)P(\neg B)$	$\approx 1-0.5*d$	$\approx$ N9 + Z15
N12	$P(A)+P(B)-P(A\mid B)P(B)+P(\neg A\mid\neg B)P(\neg B)$	1	$\approx 1+k+$ Z1
N13	$P(A\mid B)P(B)+P(A\mid\neg B)P(\neg B)+P(B)-P(A\wedge B)+P(\neg A\wedge\neg B)$	1	$\approx 1+k+$ Z1 + Z11
N14	$P(A\mid B)P(B)+P(A\mid\neg B)P(\neg B)+P(B)-P(A\mid B)P(B)+P(\neg A\wedge\neg B)$	$\approx 1+0.5*d$	$\approx$ N13 + Z15
N15	$P(A\mid B)P(B)+P(A\mid\neg B)P(\neg B)+P(B)-P(A\wedge B)+P(\neg A\mid\neg B)P(\neg B)$	$\approx 1-0.5*d$	$\approx$ N13 - Z15
N16	$P(A\mid B)P(B)+P(A\mid\neg B)P(\neg B)+P(B)-P(A\mid B)P(B)+P(\neg A\mid\neg B)P(\neg B)$	1	$\approx 1+k+$ Z1 + Z11

Table 1: The 16 normality identities that should sum to 1 according to classical probability theory. The PPN column shows the estimated predictions of the probability plus noise model and the QSS column shows that of the Quantum Sequential Sampler. Predictions of Bayesian Sampler would be slightly different regarding the conditionals but mostly similar to that of PPN (Zhu et al., 2020). Note in these two columns, N stands for the predicted values of the normality identities and Z stands for the predicted values of the probability identities in Zhu et al. (2020) of the two models. In the PPN column, $d \in [0,1]$ is the cognitive noise parameter. In the QSS column, k stands for the predicted magnitude of violation of binary complementarity: $k = J(A)_{QSS} + J(\neg A)_{QSS} - 1$. Event A in the table can be any arbitrary event in the probability space (e.g. in the election dataset A can be either Biden wins or Trump wins).

plicitly test adherence to the normalization principle in probability judgments, which we referred to be the 'normality identities' (see Table 1). We will evaluate how well probability models predict these identities. Moreover, we will investigate whether CRT scores can lend further insight about model performance, in terms of the extent of normality identity violations for participants with different CRT scores.

Models of Probability Judgments

We first introduce several recent and influential models of probability judgments. In later sections, we will evaluate the effectiveness of these models in predicting the two novel benchmarks. Note that we only focus on models capable of offering a comprehensive explanation for a wide range of probabilistic fallacies, rather than those limited to addressing only one or two fallacies, e.g. averaging models (Abelson et al., 1987; Yates & Carlson, 1986).

First, we consider the probability plus noise model (**PPN**) (Costello & Watts, 2014, 2016a, 2017). This model asserts that while subjective probabilities adhere to Kolmogorov's probability principles, they are not directly accessible. Rather, they serve as the basis for a noisy sampling process that yields the observed probability judgments. Specifically, the average probability judgment $J(A)_{PPN}$ for arbitrary event A is given by:

$$J(A)_{PPN} = dP(A) + (1-d)(1-P(A)), \tag{1}$$

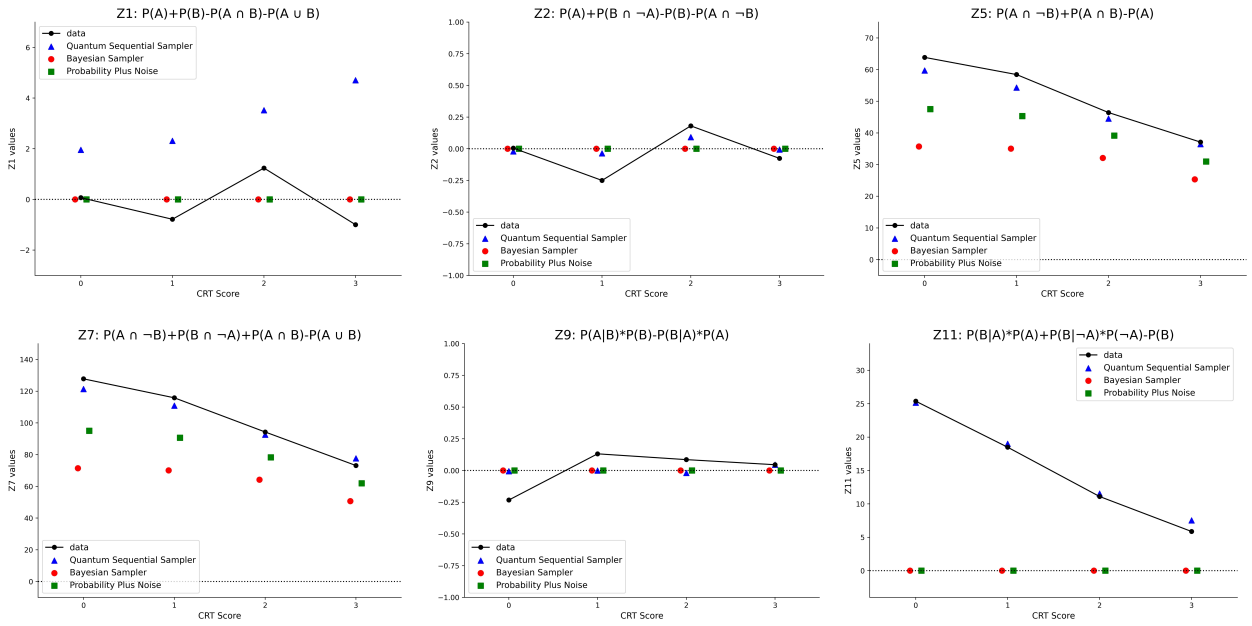
where $d \in [0,1]$ is the chance that people miscount event A as $\neg A$ due to cognitive noise and $P(A)$ denotes the subjective probability of A. To explain phenomena such as conjunction and disjunction fallacies in mean judgments, Costello & Watts (2017) proposed that error propagation amplifies the probability of miscounting to $d + \Delta d$ (with $\Delta d < d$) for such judgments. Regarding conditional events $A|B$, Costello & Watts (2016a) introduced a separate formula.

Another model employing the idea that probability judgments stem from a noisy sampling process is the Bayesian Sampler model (**BS**) (Zhu et al., 2023, 2020). But, in contrast to the PPN, the BS posits that fallacies emerge from a Bayesian prior influencing probability judgments. Zhu et al. (2020) proposed a formula indicating that for arbitrary event A, the averaged judgment $J(A)_{BS}$ under a symmetric beta prior $Beta(\beta,\beta)$ is given by:

$$J(A)_{BS} = \frac{NP(A)+\beta}{N+2\beta}, \tag{2}$$

where N denotes the sample size of the noisy sampling process and $P(A)$ the subjective probability. For conditional events, Zhu et al. (2020) do not introduce a distinct formula like in PPN, rather use the same formula as for the marginals in Equation 2. To address conjunction and disjunction fallacies, Zhu et al. (2020) propose the use of reduced sample size $N' \leq N$ for conjunctions and disjunctions, assuming that such judgments are computationally more expensive. BS was recently modified to produce autocorrelated samples, different from its original design of generating independent ones (Zhu et al., 2023). Nonetheless, this autocorrelated version of BS provides predictions that align with those of the original model concerning the new benchmarks we introduce.

In addition to models that utilize cognitive noise to explain probabilistic fallacies, there exist models that question whether subjective probabilities adhere to Kolmogorov


Z1: P(A)+P(B)-P(A ∩ B)-P(A ∪ B)

Z2: P(A)+P(B ∩ ¬A)-P(B)-P(A ∩ ¬B)

Z5: P(A ∩ ¬B)+P(A ∩ B)-P(A)

Z7: P(A ∩ ¬B)+P(B ∩ ¬A)+P(A ∩ B)-P(A ∪ B)

Z9: P(A|B)*P(B)-P(B|A)*P(A)

Z11: P(B|A)*P(A)+P(B|¬A)*P(¬A)-P(B)

Figure 1: Representative probability identities (Z1, Z2, Z5, Z7, Z9, Z11) in relation to CRT scores. The black lines represent the averaged empirical data over all participants in each CRT group and the markers indicate the averaged predictions of the models. The dashed line illustrates the normative expected values for these identities.

principles. A notable instance is a family of models based on quantum probability theory (Pothos & Busemeyer, 2009; Busemeyer et al., 2011). These models assert that various probabilistic fallacies, which appear to contradict classical principles, actually conform to the axioms of quantum probability. However, a significant issue with quantum probability models is their inability to account for violations of certain probability identities (Costello & Watts, 2016b). This challenge has recently been addressed by the Quantum Sequential Sampler (QSS) model (Huang et al., 2023), which combines quantum probabilities with noise. Unlike the noisy sampling approaches of PPN and BS, QSS employs a Markov process to formulate probability judgments. Notably, this Markov process is characterized as a constant force process, where its drift and diffusion parameters are contingent upon the quantum subjective probabilities. Due to the inherent stochastic behavior of the model near its boundaries, it is impractical to derive definitive predictive formulas. For more details, Huang et al. (2023) provides a comprehensive examination of the model's construction.

Dataset

We summarize several key aspects of the dataset in Huang et al. (2023). First, Huang et al. (2023) recruited 1451 (908 male) participants just before the US Presidential election in 2020. Due to failures in attention tests, incomplete responses, and other technical issues, the final dataset comprised 1162 participants (730 male). This dataset is, to our knowledge, the largest of its kind for probability judgment research.

During the experiment, each participant took part in a single 25-minute session, providing responses to 78 probability judgments. These judgments evaluated the probability of Trump and Biden winning the popular vote in various states, treating Trump as the negation of Biden. The set of 78 questions encompassed all potential probability questions (including all marginal, conditional, conjunction, and disjunction probabilities) concerning the Biden vs. Trump scenario across a triplet of states. The study incorporated two different state triplets for this assessment (Triplet 1: Ohio, Missouri, Michigan; Triplet 2: Georgia, Montana, Nevada). To provide an example, a representative probability question for Triplet 1 could be, "What is the probability that Trump will win Ohio and Biden will win Missouri?". For each probability question, participants indicated their response using a slider of integers from 0 to 100, marked at multiples of 5s. The slider was initially positioned at 50, with the exact numerical values displayed to the slider's left for clarity. Additionally, probability judgments were always blocked by type (e.g., marginals were shown together). At the end of the experiment, participants performed the cognitive reflection test. There were 386 participants with a CRT score of 0, 199 participants with a CRT score of 1, 297 participants with a CRT score of 2, and 280 with a CRT score of 3.

BS and QSS were fitted to this dataset through the maximum likelihood method (detailed in Huang et al. (2023)). Given that PPN employs a binomial likelihood distribution, which can not be applied with the maximum likelihood method, we fit it using the sum of square error approach, following Costello & Watts (2014).

Binary Complementarities

Binary complementarity, a principle stating that the probabilities of complementary events must sum to 1 (e.g., $P(A) + P(\neg A) = 1$), is a cornerstone of support theory (Tversky & Koehler, 1994) and has received extensive backing in literature (Wallsten et al., 1993, 1997). However, instances where binary complementarity is not upheld do occur occasionally (Macchi et al., 1999; Windschitl, 2000; Epping & Busemeyer, 2023). BS and PPN are required to conform to binary complementarity for averaged judgments. More strongly, in these two models, for any event $E \subseteq \neg A$, the condition $P(A) + P(E) \leq 1$ must hold. A notable discovery in Huang et al. (2023) is that binary complementarity does not consistently hold in averaged judgments. This finding is particularly striking when considering that marginal probabilities are presented adjacently within the same block. According to Huang et al. (2023), this discrepancy may arise from the replacement of the complementary event ”not Biden” with ”Trump,” potentially leading to subadditivity effects that exceed 1 (Tversky & Koehler, 1994). Theoretically, only QSS can account for this observed violation of binary complementarity in averaged judgments.

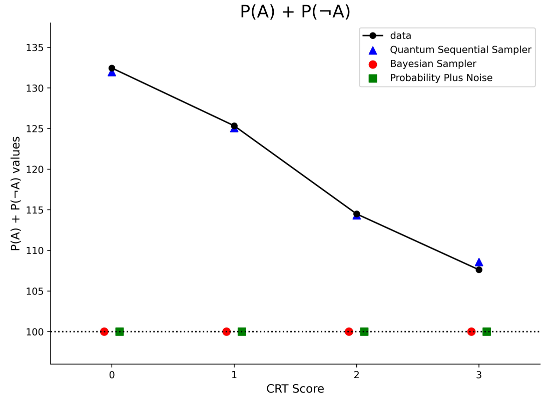


Figure 2: $P(A) + P(\neg A)$ as a function of the CRT score. Both predicted and empirical values of $P(A) + P(\neg A)$ are computed from averaging the magnitudes of all complementary pairs across participants sharing the same CRT score.

Figure 2 displays the empirical and predicted values of $P(A) + P(\neg A)$[1] in relation to the participants' CRT score. Empirically, as expected, $P(A) + P(\neg A)$ trends downwards towards 1 (or 100 in the figure, since probability judgments in Huang et al. (2023) are integers) as CRT scores increase. This is plausibly because individuals inclined towards analytical and mathematical reasoning (higher CRT score) are less likely to err in relation to binary complementarity, compared to those who rely more on intuitive judgment (lower CRT score). Statistically, this downward trend is quantified by a slope of -8.4925 (calculated across all data points within each CRT group, not just their means), with a highly significant p-value $< 10^{-16}$. To further validate the significance of this trend, Tukey's range test was applied to the CRT groups, revealing that in all comparisons, the p-value remained consistently below 10^{-16}. The QSS model closely aligns with this trend with a significant slope (p-value $< 10^{-16}$) of -6.6237. In contrast, other models, due to their theoretical constraints, fail to capture this effect.

Probability Identities

Costello & Watts (2014, 2016) introduced 18 probability identities as benchmarks for evaluating computational models of probabilistic fallacies. Table 1 in Zhu et al. (2020) summarizes these identities. According to Kolmogorov rules, these identities are expected to be zero, but empirically, this may not always be the case. Our current study aims to investigate whether these identities exhibit any correlation with CRT scores. Given space constraints, we concentrate on identities Z1, Z5, Z7, Z9, Z10, and Z15, with their relationships to CRT scores depicted in Figure 1. Plots and analyses for the remaining identities can be found in our Github Repository (available after anonymous review).

Violations of Z5 and Z7 are of significant interest, as PPN predicts that Z7 should be approximately twice as large as Z5 (similarly for Z8 and Z6), a prediction that is empirically confirmed (Costello & Watts, 2014). This finding is replicated in Huang et al. (2023). However, when it comes to explaining the substantial magnitude of violations observed in Huang et al. (2023), both PPN and BS encounter difficulties, resulting in a sum of square error significantly higher than that of QSS for both identities across all CRT groups. Despite this, Figure 1 shows that all three models generally exhibit a trend of decreasing average violations for Z5 and Z7 as CRT scores increase, aligning with the empirical trends. To confirm this empirical decreasing trend for Z5 and Z7, a linear regression model with CRT as the independent variable was fitted to all data across each CRT group for both Z5 and Z7. The resulting slopes are -9.0466 for Z5 and -18.2085 for Z7, with p values $< 10^{-16}$ for both. These trends are further validated by Tukey's range test, which shows significant differences across all comparisons for both identities, with p values consistently below 10^{-16}. The slopes again reveal that Z5 is about half the magnitude of Z7 across all CRT groups. All three models exhibit significant negative slopes for both Z5 and Z7 (p values $< 10^{-16}$) with slopes for Z5 being approximately half that of Z7, effectively reflecting the empirical trend. Similar outcomes are also observed for Z6 and Z8.

The identities Z10 to Z13 are of particular importance as they highlight scenarios where the PPN fails to accommodate deviations from zero, a phenomenon that has been reported in both Zhu et al. (2020) and Huang et al. (2023). While BS theoretically allows for individual-level violations of these identities, when it was applied to the elections dataset, the average prediction yields zero across participants. A clear

[1] Here and after, event A symbolizes any arbitrary event in the probability space, and it can be either marginal, conditional, conjunction, or disjunction.

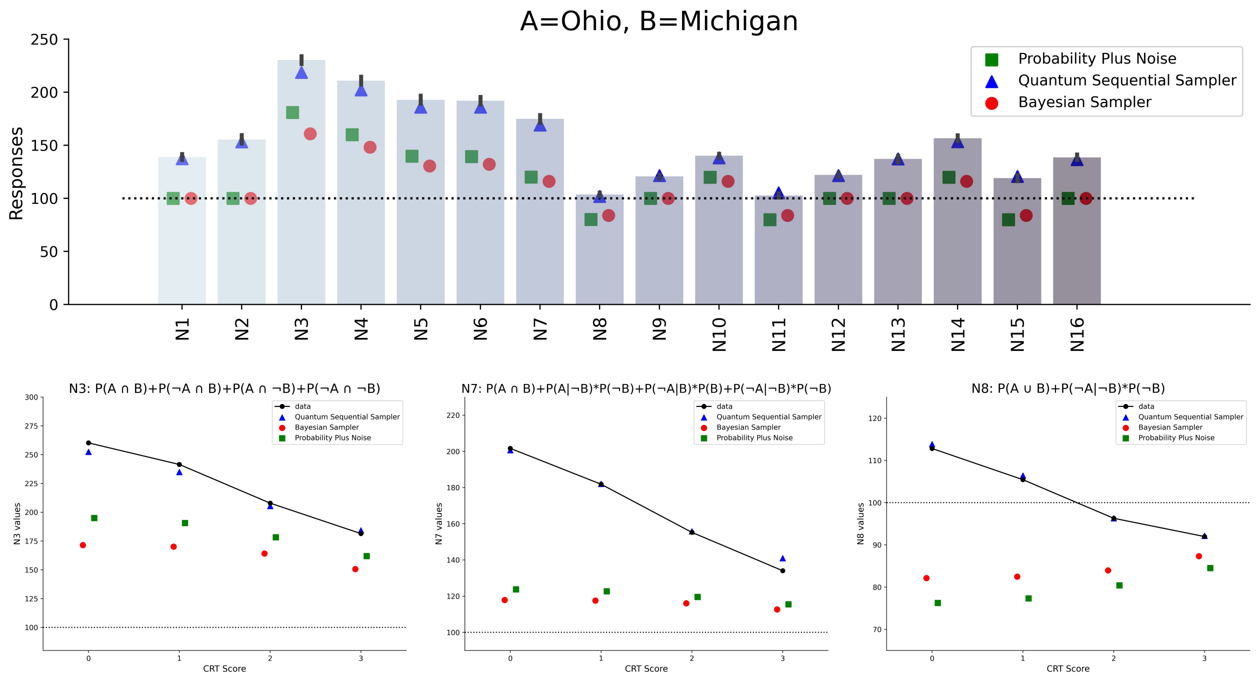


Figure 3: The upper bar plot shows the averaged value of normality identities for empirical data and model predictions over all participants for the Michigan and Ohio pair. The lower plots show representative normality identities (N3, N7, N8) in relation to CRT scores. The bars in the upper plot represents the averaged empirical values. The lower plots use the same setup as those in Figure 1.

trend of averaged empirical deviations from zero in relation to CRT scores is observed for these identities, with more substantial deviations at lower CRT scores (as illustrated by Z10 in Figure 1, with Z11 to Z13 displaying similar trends). Using linear regression with the same setup as previously, we determined that the slopes for Z10 to Z13 (all near -6.6) are statistically significant, each with p values $< 10^{-16}$. Tukey's range test reinforces this result, indicating significant differences across all comparisons. QSS adeptly reflects this trend, showing slopes for Z10 to Z13 (all around -6.1) that are significant, each with p values $< 10^{-16}$. However, as expected, BS and PPN do not exhibit slopes significantly different from 0 for Z10 to Z13, with corresponding p values nearing 1.

There are also identities that, on average, do not exhibit empirical violations and remain consistent across CRT groups. Z2 and Z9 are examples of such identities, as illustrated in Figure 1. Linear regression analysis indicates that neither the models nor the data predict a significant nonzero slope for these identities, with p values all close to 1. The stability of these identities were confirmed by Tukey's range test, which also yields p values near 1 for all comparisons. This consistency is noteworthy, considering that participants made each judgment independently, unaware of the expected values for these identities. The stability of these identities around zero, regardless of CRT score, confirms the invariances of Bayes' rules (Z9) and additivity properties (Z2) reported in Costello & Watts (2018). It further underscores that such invariances are not solely the result of deliberate mathematical thinking: even participants prone to fast and intuitive responses, as indicated by a CRT score of 0, adhere to these identities.

Finally, while QSS surpasses PPN and BS in most identities, it underperforms in one specific case: Z1. As depicted in Figure 1, both PPN and BS accurately reflect the empirical trend that averaged Z1 remains stable across CRT groups around zero. In contrast, QSS exhibits a statistically significant positive trend for Z1 as a function of CRT group, with a slope of 0.91352 and a p value $< 10^{-16}$. This shortcoming in QSS's performance can likely be attributed to its 'more likely first' assumption, where in conjunctions the more probable predicate is processed first. Due to this assumption, the interference terms from the conjunction and disjunction do not cancel out each other, resulting in an overestimation of Z1. Our finding about Z1 suggests that there is room for refining the probabilistic calculus within the QSS model.

Normality Identities

Building upon the probability identities introduced in Costello & Watts (2014, 2016a), we develop a new set of 16 identities, termed "normality identities." These identities, as per classical probability theory, should always sum to 1. Table 1 summarizes these normality identities, along with the corresponding approximate predictions from PPN and QSS. Normality identities can be derived from binary complementarity violations (denoted as v) and probability identities, but violations of either do not automatically imply a violation of normality identities. For instance, when Z15 equals v, N8 can still meet classical expectations of being 1, despite non-zero values of Z15 and v, similar to noise cancellations in Costello and Watts' identities. On the other hand, violations of normality identities also don't necessarily indicate violations in both binary complementarity and probability identities: for instance, N1 may deviate from 1 with either v or Z11 being zero (but not both). Moreover, even without binary complementarity violations, normality identities can help elucidate the connection between different probability identities, like N4 highlighting the connection between Z5 and Z11. Nonetheless, the focus of the current paper is to evaluate model predictions regarding these identities and their correlation with CRT scores, rather than to extensively analyze the interplay between various identities, a task reserved for future research.

The upper panel in Figure 3 presents the averaged empirical values of the normality identities alongside the models' average predictions over all participants. It is evident that the means of all 16 normality identities differ from 1 (or a rating of 100) in the election dataset. One-sample t tests confirm this deviation from a rating of 100 for all normality identities, with all p values $< 10^{-16}$. A closer examination of the figure also reveals that QSS consistently outperforms both PPN and BS across all 16 normality identities. This superior performance is partly attributed to the theoretical limitations of BS and PPN, which are constrained to not exceed 1 for particular identities: for N9 and N12, both models are theoretically expected to predict empirical means of 1.

A more pronounced issue arises for PPN and BS concerning the correlation between normality identities and CRT scores. Using linear regression as in the previous section, we observe that all normality identities exhibit a significant negative slope against CRT scores, with all p values $< 10^{-16}$. However, for many identities, including those where PPN and BS can theoretically predict deviations from 1, the two models either fail to show a significant negative slope or present one much less steep than that of the empirical data (e.g. N3, N7, N8 in Figure 3). In the case of N3, although PPN and BS predict significant negative slopes, their magnitudes (-10.77326 and -6.47037, respectively) are substantially lower than the empirical data's slope of -26.60998. Conversely, QSS's prediction of -23.16711 is closely aligned with this empirical trend. Similarly, for N7, while the data shows a slope of -22.82394, PPN and BS predict much smoother slopes (-2.693315 and -1.617592, respectively), in contrast to QSS's estimate of -20.51246, which more accurately reflects the empirical tendency.

More critically, for N8 (and similarly for N11), PPN and BS predict a significant positive slope, in contrast to the significant negative slope observed in both the empirical data and the QSS predictions. The reason for this issue lies in the prediction from these models that N8 and N11 should be approximately $1 - Z15$. Given that Z15 is empirically shown to have a positive value, these models must predict that N8 and N11 are mostly below 1. As CRT scores increase, N8 and N11 are predicted to approach their normative expectation of 1, leading PPN and BS to predict a positive slope. This qualitative mismatch between actual trends of N8 and N11 with CRT scores and the predictions by PPN and BS models underscores potential areas for their improvement.

An intriguing aspect of N8 and N11 are their "1 crossing" effect: the mean values of N8 and N11 continue to decrease below a rating of 100 as CRT score increases (as illustrated in Figure 3). This decreasing trend that persists below 100, robustly supported by regression analysis, stands in stark contrast to the patterns observed for all other identities, including those in Costello & Watts (2014, 2016b), where higher CRT scores typically correlate with values aligning more closely with normative expectations. An important implication of the "1 crossing" effect is that probabilistic fallacies are likely influenced by factors beyond just cognitive noise, as lower cognitive noise is generally associated with more normative responses. Why QSS can capture this trend and the underlying mechanisms inducing this effect remain open questions for future exploration.

General Discussion

We've introduced two new benchmarks for evaluating probability judgment models: the impact of CRT scores on probability judgment expressions, and normality identities, which complement existing probability identities from Costello and Watts (2014, 2016). Assessing three models – Probability Plus Noise (PPN), Bayesian Sampler (BS), and Quantum Sequential Sampler (QSS) – we found a decrease in violations with higher CRT scores, best captured by QSS. In terms of normality identities, QSS also outperforms PPN and BS, which fail to accurately capture the way these identities correlate with CRT scores. This is especially notable for N8 and N11, where PPN and BS demonstrates a correlation opposite to that of the empirical data.

We briefly mention three future directions to pursue. First, although CRT score is a valuable measure of individual differences, its relationship to cognitive noise might be only partially diagnostic. A more diagnostic indicator of cognitive noise is cognitive load, which quantifies the mental resources required to process information in working memory. In upcoming experiments, one can investigate how cognitive load correlates with probabilistic fallacies and probability judgment expressions in a manner similar to our analyses for CRT scores. Secondly, the notable "1 crossing" effect seen in N8 and N11 merits further investigation. This phenomenon suggests that there are underlying processes identified by CRT that encompass more than just cognitive noise and straightforward mathematical reasoning. Finally, the two novel benchmarks proposed can also be applied to other dataset besides the current one, including those with repeated measurements.

Acknowledgment

EMP was supported by European Office of Aerospace Research and Development (EOARD) grant FA8655-23-1-7220. JRB was supported by the Air Force Office of Scientific Research under award number FA9550-20-1-0027.

References

Abelson, R. P., Leddo, J., & Gross, P. H. (1987). The strength of conjunctive explanations. *Personality and Social Psychology Bulletin*, *13*(2), 141–155.

Bialek, M., & Pennycook, G. (2018). The cognitive reflection test is robust to multiple exposures. *Behavior research methods*, *50*, 1953–1959.

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological review*, *118*(2), 193.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? a mathematical modeling approach. *Memory & cognition*, *42*, 434–447.

Costello, F., & Watts, P. (2014). Surprisingly rational: probability theory plus noise explains biases in judgment. *Psychological review*, *121*(3), 463.

Costello, F., & Watts, P. (2016a). People's conditional probability judgments follow probability theory (plus noise). *Cognitive psychology*, *89*, 106–133.

Costello, F., & Watts, P. (2016b). A test of two models of probability judgment: quantum versus noisy probability. In *Cogsci.*

Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, *30*(2), 304–321.

Costello, F., & Watts, P. (2018). Invariants in probabilistic reasoning. *Cognitive psychology*, *100*, 1–16.

Elqayam, S., & Evans, J. S. B. (2013). Rationality in the new paradigm: strict versus soft bayesian approaches. In *New paradigm psychology of reasoning* (pp. 216–234). Routledge.

Epping, G. P., & Busemeyer, J. R. (2023). Using diverging predictions from classical and quantum models to dissociate between categorization systems. *Journal of Mathematical Psychology*, *112*, 102738.

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 678.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Huang, J. (2023). Models of human probability judgment errors.

Huang, J., Busemeyer, J. R., Ebelt, Z., & Pothos, E. (2023). Quantum sequential sampler: a dynamical model for human probability reasoning and judgments. *Annual Meeting of Cognitive Science Society 45 (45).*

Kahneman, D. (2002). Maps of bounded rationality: A perspective on intuitive judgement and choice.

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review*, *106*(1), 210.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality: The probabilistic approach to human reasoning*. Oxford University Press.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior research methods*, *48*, 341–348.

Pothos, E. M., & Busemeyer, J. R. (2009). A quantum probability explanation for violations of 'rational'decision theory. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1665), 2171–2178.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the cognitive reflection test is stable across time. *Judgment and Decision making*, *13*(3), 260–267.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & cognition*, *39*(7), 1275–1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & reasoning*, *20*(2), 147–168.

Trueblood, J. S., Yearsley, J. M., & Pothos, E. M. (2017). A quantum probability framework for human probabilistic inference. *Journal of Experimental Psychology: General*, *146*(9), 1307.

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

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological review*, *101*(4), 547.

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and combining subjective probability estimates. *Journal of Behavioral Decision Making*, *10*(3), 243–268.

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society*, *31*(2), 135–138.

Welsh, M., Burns, N., & Delfabbro, P. (2013). The cognitive reflection test: How much more than numerical ability? In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).

Windschitl, P. D. (2000). The binary additivity of subjective probability does not indicate the binary complementarity of perceived certainty. *Organizational Behavior and Human Decision Processes*, *81*(2), 195–225.

Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: Evidence for multiple judgment procedures, including "signed summation". *Organizational Behavior and Human Decision Processes*, *37*(2), 230–253.

Yearsley, J. M., & Trueblood, J. S. (2018). A quantum theory account of order effects and conjunction fallacies in political judgments. *Psychonomic bulletin & review*, *25*, 1517–1525.

Zhu, J.-Q., Sanborn, A. N., & Chater, N. (2020). The bayesian sampler: Generic bayesian inference causes incoherence in human probability judgments. *Psychological review*, *127*(5), 719.

Zhu, J.-Q., Sundh, J., Spicer, J., Chater, N., & Sanborn, A. N. (2023). The autocorrelated bayesian sampler: A rational process for probability judgments, estimates, confidence intervals, choices, confidence judgments, and response times. *Psychological Review*.