MIAMI UNIVERSITY

The Graduate School

Certificate for Approving the Dissertation

We hereby approve the Dissertation

of

Christopher Ryan Fisher

Doctor of Philosophy

Director
Christopher R. Wolfe

Reader
Joseph G. Johnson

Reader
Robin D. Thomas

Reader
Andrew Reffett

ARE PEOPLE NAÏVE PROBABILITY THEORISTS? AN EXAMINATION OF THE PROBABILITY THEORY + VARIATION MODEL

by Christopher Ryan Fisher

Four experiments tested the Probability Theory + Variation model of probability judgment. The model posits that judgments follow the rules of probability theory. Errors occur because otherwise normative judgments are perturbed with noise. Experiment 1 found some evidence for the model's account of noise and errors. However, no support was found for a prediction derived from the variance sum law and the integration rules of the model. Experiment 2 found some support that noise is associated with more errors in conditional probability judgment and judgments adhered stochastically to Bayes' theorem. Experiment 3 reformulated the model as a simple process model in which judgments are formed through the dynamic accumulation of exemplars. Noise was increased through a response deadline, but only resulted in less semantic coherence for conditional probabilities. In Experiment 4, three interventions based on the model and variants the wisdoms of crowds effects were largely ineffective in reducing errors.

ARE PEOPLE NAÏVE PROBABILITY THEORISTS? AN EXAMINATION OF THE PROBABILITY THEORY + VARIATION MODEL

A Dissertation

Submitted to the Faculty of

Miami University in partial

fulfillment of the requirements

for the degree of

Doctor of Philosophy

Department of Psychology

by

Christopher Ryan Fisher

Miami University

Oxford, Ohio

2014

Dissertation Director: Christopher R. Wolfe

Table of Contents

List of Tables

List of Appendices

Introduction

The ability to judge and reason with probabilities is an integral component to decision making and has far-reaching consequences in many practical domains. For example, scientists must acquire a basic understanding of probability theory in order to conduct and evaluate research. In medical decision making, physicians and patients must have a basic understanding of probability theory in order to understand risk and interpret diagnostic tests. Although probability judgment is an important skill, forty years of research has revealed a multitude of biases in probability judgment. Performance in judgment can be evaluated according to two basic criteria: correspondence and coherence (Hammond, 2000). Correspondence refers to the empirical accuracy of judgments. One common finding regarding correspondence is under-confidence—that is, judgments are less extreme compared to their empirical probabilities (e.g Erev, Wallsten, & Budescu, 1994). For example, an event with a true probability of 20% might be judged as 30% whereas an event with a true probability of 80% might be judged as 70%. By contrast, coherence refers to the internal consistency of judgments. A set of judgments is considered to be coherent if it conforms to the rules prescribed by probability theory. A multitude of systematic deviations from probability theory have been observed in the literature (e.g. Tversky & Kahneman, 1983). Perhaps the most well known violation of probability theory is the conjunction fallacy, which occurs when a subset is judged as more probable than the superordinate set in which it is contained.

In the classic demonstration of the conjunction fallacy, participants read the following personality sketch that describes a fictitious woman named Linda in terms of a feminist stereotype:

> Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations. (Tversky & Kahneman, 1983).

After reading the personality sketch, participants are typically instructed to judge or rank the probability that Linda is (1) a bank teller and (2) a bank teller and a feminist. Approximately 85% of participants stated that Linda was more likely to be a feminist bank teller than a bank teller, thereby committing the conjunction fallacy. What makes the conjunction fallacy interesting is its simplicity and robustness. Both novices and individuals trained in statistics commit the fallacy at approximately equal rates (Tversky & Kahneman, 1983). Many of the judgment phenomena reviewed herein share these properties, suggesting they have the potential to reveal the fundamental nature of underlying cognitive processes. The goal of the present paper is to test a model that shows potential to provide a comprehensive account of probability judgment.

Overview

Several theories have been proposed to explain probability judgment. One common limitation among these theories is that they only account for a subset of phenomena individually. My primary goal in the present paper was to test a model that has the potential to provide a more comprehensive account of probability judgment and generate novel predictions. The model is called the probability theory + variation model (PTV; Costello, 2009). According to the PTV model, probability judgments adhere to the rules of probability theory, but are perturbed with noise (i.e. random variability). The noise in the judgments can produce many systematic errors found in the literature when they are combined according to the rules of probability theory. This marks a departure from many of the theories reviewed below, which propose non-normative mechanisms to account for non-normative judgments.

The remainder of the paper will be organized as follows. The sections Errors in Joint Probability and Errors in Conditional Probability provide an overview of the phenomena that have been uncovered in probability judgment. These sections provide the necessary background information to evaluate the limitations of other theories in the subsequent section titled Theories and Models. The Theories and Models section includes a formal description of the PTV model and the phenomena for which it can account. The remaining sections of the paper detail four experiments designed to test the PTV model. Experiment 1 compared the PTV model to the configural weighted average model in joint probability judgment (CWA; Nilsson, Winman, Juslin, & Hansson, 2009). These models offer differing accounts of the relationship between noise and judgment errors. According to the PTV model, errors should increase as noise increases. By contrast, the CWA model generally predicts fewer errors with more noise. An additional critical property of the PTV model was derived from the variance sum law and tested empirically. Experiment 2 tested predictions of the PTV model in conditional probability judgment, a domain in which the model has not been previously tested. In particular, Experiment 2 examined the relationship between noise and errors and tested whether judgments adhere stochastically to Bayes' Theorem. Experiment 3 provided the initial groundwork for instantiating the PTV model as a cognitive process model. According to this simple model, judgments are formed from exemplars that are sampled dynamically from memory until a precision threshold is met. This model predicts that the variance in judgments will decrease over time as more exemplars are accrued. A response deadline was instated to test whether judgments made more quickly are more variable. Experiment 4 investigated three interventions based on the PTV model and variants of the wisdom of crowds effects. The interventions employed various judgment and averaging methods to hone in on participants true judgments and thereby improve coherence. The paper concludes with a discussion of limitations, alternative formulations of the PTV model and future directions.

Errors in Joint Probability Judgment

Conjunction and Disjunction Fallacies

Numerous studies show that the conjunction and disjunction fallacies are affected by many of the same factors (Costello, 2009b; Fisk, 2002, Crisp & Feeney, 2009; Wolfe, Fisher & Reyna, 2013). For this reason, the conjunction and disjunction fallacies will be presented concurrently. A conjunction fallacy occurs when

$$\max\big(\mathrm{P(A)}, \mathrm{P(B)}\big) > P(\mathrm{A} \cap \mathrm{B}) > min(P(\mathrm{A}), \mathrm{P(B)}) \tag{1}$$

In other words, the conjunction fallacy occurs when the conjunction is larger than is logically possible. The conjunction fallacy is distinguished from the double conjunction fallacy, which occurs when

$$\mathrm{P(A \cap B)} > max(P(\mathrm{A}), \mathrm{P(B)}) \tag{2}$$

Thus, the double conjunction fallacy occurs when the conjunction is larger than both of the components. A disjunction fallacy occurs when

$$\min\big(\mathrm{P(A)}, \mathrm{P(B)}\big) < P(\mathrm{A} \cup \mathrm{B}) < max(P(\mathrm{A}), \mathrm{P(B)}) \tag{3}$$

indicating that the disjunction is smaller than is logically possible. The disjunction fallacy is distinguished from the double disjunction fallacy in which

$$\mathrm{P(A \cup B)} < min(P(\mathrm{A}), \mathrm{P(B)}) \tag{4}$$

Thus, the double disjunction fallacy occurs when the disjunction is smaller than both component probabilities.

One robust finding is that the fallacy rates depend on the component probability estimates (Fisk, 2002; Nilsson et al., 2009). The fallacy rates are highest when one component is low (e.g. bank teller) and the other one is high (e.g. feminist). By contrast, the fallacy rates are lower when both component probabilities are high (e.g. feminist; vegan) or both components are low (e.g. bank teller; stamp collector). Another robust finding is that the fallacy rates increase as the conditional probability between the components increases (Crisp & Feeney, 2009).

Minimum Conjunction Error

Unlike the conjunction fallacy, the minimum conjunction error occurs when a conjunction is judged to be too low rather than too high (Wolfe & Reyna, 2010; Fisher & Wolfe, 2011). As an example, suppose P(A) = P(B) = .60. According to probability theory, the sum of all disjoint events must equal 1. Because P(A) + P(B) = 1.20, probability theory requires that the minimum conjunction must be .20. In general, a minimum conjunction error occurs when:

$$\max(0, \mathrm{P(A)} + \mathrm{P(B)} - 1) > P(A \cap B) \tag{5}$$

Maximum Disjunction Error

Unlike the disjunction fallacy, a maximum disjunction error occurs when a disjunction is too high rather than too low (Wolfe & Reyna, 2010; Fisher & Wolfe, 2011). Consider the addition law:

$$\mathrm{P(A \cup B)} = \mathrm{P(A)} + \ \mathrm{P(B)} - \mathrm{P(A \cap B)} \tag{6}$$

According to the addition law in Equation 6, the disjunction cannot be larger than the sum of the components: P(A) + P(B). In general, the maximum disjunction error occurs when:

$$\min(1, \mathrm{P(A)} + \mathrm{P(B)}) < P(A \cup B) \tag{7}$$

Semantic Coherence

A set of judgments may adhere to the rules of probability theory without being consistent with the semantic content of a problem. As an example, suppose a political analyst makes the following prediction about an election outcome: a 75% chance that Smith will win, a 25% chance that Davis will win and a 20% chance that both Smith and Davis will win. Although the conjunction fallacy was not committed in the example, it does not accord with the fact that only one candidate can win. Thus, the probability that both candidates win is zero because election outcomes are mutually exclusive. A set of judgments that maps onto the qualitative relationship between two sets is considered to be semantically coherent (Wolfe & Reyna, 2010; Wolfe, Fisher & Reyna, 2013). There are five qualitative relationships between two events, A and B: identical (e.g. H_2O and water), mutually exclusive (e.g. bee and wasp), subset (e.g. cat and mammal), independent (e.g. heads on a coin flip and rain) and overlapping (e.g. feminist and bank teller). A general finding is that people have higher semantic coherence for identical sets, independent sets, and mutually exclusive sets compared to subsets and overlapping sets (Wolfe & Reyna, 2010; Wolfe, Fisher, & Reyna, 2013).

Stochasticity

People tend to provide different judgments to the same question when asked multiple times (Nilsson et. al., 2009). Thus, probability judgment is generally stochastic rather than deterministic. As detailed in subsequent sections, the noise in judgments may be the source of systematic errors in judgment.

Stochastic Adherence to the Addition Law

The addition law can be rewritten as:

$$\mathrm{P(A)} + \ \mathrm{P(B)} - \mathrm{P(A \cap B)} - \mathrm{P(A \cup B)} = 0 \tag{8}$$

Costello and Watts (2013) found that judgments adhere stochastically to the addition law. In other words, the sum of the judgments is distributed around zero across problems.

Subadditivity

When superordinate set is partitioned into mutually exclusive and exhaustive subsets, the probability of the sum of the subsets must equal the probability of the superordinate set. More formally, let a_1, a_2 … a_n be mutually exclusive and exhaustive events in set A. Additivity requires

$$P(A) = \sum_{i=1}^{n} P(a_i) \tag{9}$$

When judged events are decomposed or "unpacked" into subsets, the sum of the judgments often exceeds the judgment for the superordinate set, a phenomenon known as subadditivity (Tversky & Koehler, 1994; Bearden, Wallsten & Fox, 2007). A common finding is that the degree of subaddivity increases as the superordinate set is decomposed into more subsets (e.g. Dougherty, & Hunter, 2003). Another key finding is that judgments exhibit binary complementarity, a phenomenon in which complementary judgments sum to 1 on average. As an example, suppose the average judgment across participants for event A is .60. Binary complementarity would be satisfied if the average judgment for the complementary event ~A is .40.

Order effects

One study found that the conjunction fallacy rate is higher when the conjunction is judged before the component probabilities (Stolarz-Fantino, Fantino, Zizzo, & Wen, 2003). However, no studies have examined order effects for the other types of fallacies. Experiment 1 examined whether order effects occur in the other fallacies and errors.

Errors in Conditional Probability Judgment

Base rate neglect

Learning contingencies in one's environment requires integrating new information with existing information. When uncertainty is involved, Bayes' theorem provides a rational basis for updating information. A common finding is that people tend to underweight base rates relative to individuating information, a phenomenon known as base rate neglect. In a classic demonstration of base rate neglect, participants read a description of a person named Jack, who resembled an engineer (Kahneman & Tversky, 1973). Participants judged the probability Jack is an engineer when the base rate was consistent with the description (high) or inconsistent with the description (low). According to Bayes' theorem, the judgments should vary with the base rates. However, the judgments were not sensitive to the stated base rates, indicating base rate neglect.

Conversion Error

A conversion error is a special case base rate neglect in which P(A|B) and P(B|A) are erroneously judged to be equal:

$$\mathrm{P(A|B) = P(B|A), P(A) \neq P(B)} \tag{10}$$

In a typical Bayesian inference problem, participants are provided with the hit rate and false alarm rate of a diagnostic test and the base rate of a disease. They are instructed to judge the posterior probability that a person has a disease given a positive test result. Consider the following:

> The probability of breast cancer is 1% for a woman at age forty who participates in routine screening [base-rate]. If a woman has breast cancer, the probability is 80% that she will get a positive mammography [hit-rate]. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography [false-alarm rate]. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? _%. (Gigerenzer & Hoffrage 1995, p. 685)

Although the solution is approximately 8%, a typical response approximates the hit rate of the test, suggesting the commission of the conversion error (e.g. Barbey & Sloman, 2007).

Conditional Reversal

Base rate neglect and the conversion errors occur when the judged posterior probability is too high compared to the Bayesian solution. A conditional reversal is an even more extreme case in which the conditional probabilities reverse their logical rank ordering (Fisher & Wolfe, 2011). For example, if P(A) > P(B) then P(A|B) > P(B|A). Thus, a conditional reversal occurs when P(A) > P(B) and P(A|B) < P(B|A).

$$\mathrm{P(A)} > P(\mathrm{B}), \mathrm{P(A|B)} < P(B|A) \tag{11}$$

Minimum Conditional Error

A minimum conditional error occurs when a conditional probability judgment is smaller than logically possible (Fisher & Wolfe, 2011). The minimum conditional error is related to the minimum conjunction error through the definition of a conditional probability:

$$\mathrm{P(A|B)} = \frac{\mathrm{P(A \cap B)}}{\mathrm{P(B)}} \tag{11}$$

The minimum conditional error can be found by dividing both sides of Equation 5 by P(B) and substituting equation 11 on the right hand side:

$$\frac{\max(0, \mathrm{P(A)} + \mathrm{P(B)} - 1)}{\mathrm{P(B)}} > P(A|B) \quad (12)$$

A minimum conditional error for P(B|A) can be derived in a similar manner by dividing both sides of Equation 12 by P(A).

Models and Theory

Representativeness Heuristic

The original explanation for the conjunction fallacy and base rate neglect put forth by Tversky and Kahneman (1983) attributed the errors to the use of the representativeness heuristic (RH). According to the RH, the similarity of a target (e.g. Linda) to a category (e.g. feminist) forms the basis of probability judgments. Judgments based on representativeness can produce the conjunction fallacy because they are not bound by the class inclusion rules of probability theory. For example, Linda is more representative of feminist bank teller than bank teller because the personality sketch describes Linda in terms of a feminist stereotype. As a result, feminist bank teller is generally rated as more probable than bank teller. In support of the RH, mean probability judgment and representativeness ratings correlated at .95 or higher for five problems (Tversky & Kahneman, 1983).

As intuitively compelling as this explanation may seem, the RH received little support in subsequent studies. One simple hypothesis is that the conjunction fallacy should decrease substantially when the RH is not applicable. Contrary to the RH, removing the personality sketch produces only some (Stolarz-Fantino, Fantino, & Kulik, 1996) or no decrease in the conjunction fallacy (Stolarz-Fantino et al., 2003). Along similar lines, Gavanski & Roskos-Ewoldsen (1991) created mixed and probability combination conditions in which the RH is not applicable. In the mixed condition, participants read two personality sketches and judged the conjunctive probability of one event from each description (e.g. Linda is a feminist and Jason is an artist) as well as the component events. For mixed problems, the RH is applicable to the component events but not their conjunction. In the probability combination condition, the problems pertained to an unfamiliar topic (e.g. fictitious creatures from a fictitious planet) and thus provided no basis for using the RH. Furthermore, the component probabilities in the probability combination condition where yoked to judgments made in the standard condition. In both cases, the rates of the conjunction fallacy were similar to those observed in the standard condition in which the RH was applicable. Collectively, these results suggest the conjunction fallacy is due to the manner in which component probabilities are integrated into conjunctive probabilities rather than representativeness.

Gigerenzer (1996) criticized the heuristics and biases approach more generally on several interrelated grounds—namely, the vague specification of the heuristics, the high degree of post-hoc flexibility and the lack of process model. Gigerenzer (1996) argued that because the heuristics are vaguely defined, they amount to re-descriptions of the phenomena they purport to explain. Consequentially, they provide little or no insight into the psychological processes and can be evoked post-hoc to explain any empirical finding. For example, base rate neglect could be explained with the RH while conservatism—the opposite of base rate neglect—could be explained by a different heuristic, such as anchoring. Consequentially, the RH offers little explanation of several of the empirical phenomena listed above, such as order effects, stochasticity, semantic coherence and minimum overlap errors.

Linguistic Misinterpretation

Some theorists have proposed that errors such as the conjunction fallacy are not necessarily errors. Instead, so-called errors stem from a misinterpretation of the word probability and logical operators. For example, ‘And’ has several meanings in natural language that differ from the its meaning as a logical operator. Consider the statement “Friends and family came to my party.” ‘And’ refers to the union of friends and family rather than their intersection (family members who are also friends). ‘And’ can also denote temporal succession, as in the statement “Bob went home and ate”. Alternatively, bank teller may not be interpreted inclusively (feminist and bank teller or not feminist and bank teller) due to its redundancy with feminist bank teller. As a result, bank teller may be interpreted as not feminist and bank teller, in which case, $\mathrm{P}(\mathrm{A} \cap \mathrm{B}) > P(\mathrm{B})$ is not fallacious. The word probability may also be interpreted in multiple ways, such as “reasonable”, “believable” and “plausible”, none of which are bound by the rules of probability theory (Hertwig & Gigerenzer, 1999).

Converging evidence does not support the notion that the conjunction fallacy is due entirely or even largely to linguistic misinterpretations. In the original study, Tversky and Kahneman (1983) expressed the event ‘bank teller’ as a disjunction: ‘bank teller whether or not she is a feminist’ to emphasize its inclusive meaning. Nonetheless, the conjunction fallacy rate remained high (57%). Along the same lines, similar rates of the conjunction fallacy were observed when the option ‘B and not A’ were included in a different set of problems (Tentori, Bonini, & Osherson, 2004; Wedell, & Moro, 2008). By including ‘B and not A’, ‘B’ should retain its inclusive meaning according to Gricean maxims.

Betting paradigms provide a method of circumventing the ambiguity inherent in the word probability. Tversky and Kahneman (1983) anticipated this in their original study. Using real stakes, participants betted on one of the following sequences of outcomes from multiple rolls of a colored die: (1) GRGRRR or (2) RGRRR. Notice that sequence 2 is a subset of sequence 1, formed my removing the first outcome, G. Even though the word ‘probability’ and the word ‘and’ were not used, the conjunction fallacy remained at high levels (62%). Similar findings in betting paradigms were observed in Bonini, Tentori, & Osherson (2004). Taken together, these results suggest that the conjunction fallacy is a real phenomenon.

“Natural” Frequencies

One controversial claim is that people have evolved a cognitive algorithm to process frequencies rather than probabilities (Barbey & Sloman, 2007; Gigerenzer, & Hoffrage, 1995). According to the natural frequency perspective, people reason better statistically with frequencies because events are encountered sequentially in the environment, a process termed natural sampling (Gigerenzer, & Hoffrage, 1995). Probabilities are formed through normalizing, a process that eliminates base rate information and must be explicitly re-incorporated through Bayes’ theorem (Hoffrage, Gigerenzer, Krauss, & Martignon, 2002). Unlike probabilities, natural frequencies maintain sample size information and base rate information in its statistical structure, thereby simplifying computations. Facilitation in Bayesian inference is observed when the information is presented in terms of natural frequencies:

> 10 out of every 1,000 women at age forty who participate in routine screening have breast cancer [base-rate]. 8 out of every 10 women with breast cancer will get a positive mammography [hit-rate]. 95 out of every 990 women without breast cancer will also get a positive mammography [false-alarm rate]. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? __ out of __ . (Gigerenzer & Hoffrage 1995, p. 688)

Others have argued that the observed facilitation is not uniquely attributed to statistical format (Barbey & Sloman, 2007). Rather it is due to transparency in the hierarchical set structure of the problem. For example, '8 out of 10 women' are nested within '10 out 1000 women'. In support of this argument, similar facilitation in Bayesian inference problems was observed when the hierarchical set structure was represented with Euler diagrams instead of natural frequencies (Barbey & Sloman, 2007; Wolfe, Fisher & Reyna, 2012). Similar facilitation was observed with a roulette wheel diagram designed to expose hierarchical set structure (Yamagishi, 2003). Moreover, no facilitation was observed when natural frequencies are presented as defective partitions (e.g. changing the problem to 895 out of 990 women without breast cancer will get a negative mammography; Barbey & Sloman, 2007).

Reductions in the conjunction fallacy have been less consistent with natural frequencies. In their original study, Tversky & Kahneman (1983) found a reduction in the conjunction fallacy when using natural frequencies, as did Hertwig & Gigerenzer (1999) and Costello (2008). On the other hand, Wedell, & Moro (2008) did not find an effect of statistical format. Nonetheless, a similar argument could be made that exposing the hierarchical set structure is the locus of the reduction rather than the natural frequencies per se.

Putting these controversies aside, the natural frequency perspective offers no account of the phenomena listed above. It simply asserts that a cognitive algorithm evolved for the processing of natural frequencies and thus facilitation in statistical judgment should be observed when problems are presented in the same format. Another remaining problem is that approximately 50% or more still fail to provide the Bayesian solution with natural frequencies (Barbey & Sloman, 2007; Gigerenzer & Hoffrage, 1995). In short, the natural frequency perspective does not provide a comprehensive account of probability judgment.

Fuzzy Trace Theory

Fuzzy Trace Theory (FTT) explains probability judgment largely in terms of memory representation (Wolfe & Reyna, 2010; Reyna, & Brainerd, 1995). According to FTT, multiple representations are stored in memory, ranging from verbatim to gist. A verbatim representation is highly detailed, while a gist representation distills information into its underlying meaning. For example, the probability of rain may be represented on the following verbatim to gist continuum, ranging from 92% chance of rain, 90% chance of rain, it will probably rain to "It will rain. I better bring an umbrella." In this example, numeric detail was lost as the representation became more gist-like. What remained in its most gist-like representation was the underlying meaning and relevance to a person. One tenet of FTT is that people prefer to use gist-like representations whenever possible. Reliance on gist representations can lead to errors, a phenomenon known as denominator neglect (Wolfe & Reyna, 2010). Denominator neglect can be described in terms of a 2x2 table. Ignoring the marginal rows (i.e. denominators) produces the conjunction fallacy and conversion error because class inclusion relationships are ignored.

FTT has been successful in explaining several phenomena related to semantic coherence. As predicted by FTT, semantic coherence increases when the denominators are identical or not relevant. For example, semantic coherence is lowest for overlapping sets and subsets because they require a full representation of the hierarchical set structure (Wolfe & Reyna, 2010; Wolfe, Fisher & Reyna, 2012). By contrast, semantic coherence on identical sets is generally high because a simplified gist representation is sufficient. Semantic coherence on difficult overlapping set problems improves when the events are independent, in which case, the denominators are simplified: $P(A) = P(A|B)$. Moreover, gist representations (i.e. judgments consistent with identical sets) are used by default in the absence of semantic information (Wolfe & Reyna, 2010). This supports the assumption that people use more gist-like representations whenever possible. However, FTT in its current form lacks the specificity to account for several phenomena. For example, it does not explain why fallacy rates depend on the component and conditional probabilities, the stochasticity of judgments and order effects.

Averaging Models

Probability theory requires the multiplicative combination of probabilities, such as in the conjunction rule. One proposal is that probabilities are combined in an additive rather than multiplicative manner. Various averaging models have been proposed, ranging from a simple mean (Fantino, Kulik, Stolarz-Fantino & Wright, 1997) to a geometric mean (Abelson, Leddo & Gross, 1987). Additive integration allows averaging models to account for the conjunction and disjunction fallacies. However, averaging models have many shortcomings. For example, the ordering implied by averaging functions implies the conjunction fallacy will occur in nearly all cases:

$$\mathrm{P(B)} \leq \mathrm{f(A \cap B)} \leq \mathrm{P(A)} \tag{13}$$

where $f(\cdot)$ is an averaging function. In actuality, the fallacy rates are quite variable and depend on the component and conditional probabilities among other factors (e.g. Crisp & Feeney, 2009; Fisk, 2002). One problem with the geometric model in particular is that the parameters lack a natural psychological interpretation. Even if the geometric model fits the data well, it does not provide any insight without a psychological interpretation. Finally, averaging models cannot account for order effects, double conjunction and double disjunction fallacies, minimum conjunction errors, and maximum disjunction errors because they are deterministic.

Configural Weighted Average Model

As previously mentioned, a major shortcoming of the averaging models is that they predict a conjunction fallacy under nearly all parameterizations and cannot produce double conjunction or double disjunction fallacies (among others). The configural weighted average (CWA) model is a stochastic model that circumvents this problem through the explicit modeling of noise inherent in probability judgments (Nilsson et al., 2009). Formal notation will be introduced here in order to describe the CWA model and the PTV in model next section. Let $k \in \{A, B, A \cap B, A \cup B, A|B, B|A\}$ and let $S(k) \in [0, 1]$ be a random variable representing the reported subjective probability of event k. The subjective probability can be decomposed into a true probability component and an additive error component:

$$S(k) = P(k) + e_k \tag{14}$$

where P(k) is the true judgment such that $P(k) \in [0, 1]$ and e_k is an error term that can assume positive or negative values. The expectation of the subjective probability equals the true probability judgment:

$$E[S(k)] = P(k) \tag{15}$$

Aside from these constraints imposed on the subjective probabilities and their components, no particular claims are made regarding the functional form. According to the CWA model, cues are attended to sequentially, producing an independent and additive adjustment. Joint probabilities can be modeled as a weighted average of component probabilities. Assuming S(A) > S(B), weights are configurally applied to conjunctive and disjunctive judgments as follows:

$$S(A \cap B) = (1 - w)S(A) + wS(B) \tag{16}$$

$$S(A \cup B) = wS(A) + (1 - w)S(B) \tag{17}$$

Nilsson et al. (2009) argued that additive integration is less taxing cognitively than multiplicative integration and is more robust (i.e. accurate) when judgments are perturbed with noise. In their study, the parameters were fixed at w = .80 to provide an approximation to the multiplicative integration used by probability theory. Using independent events, they found the CWA model provided an accurate fit to the data at both the aggregate and individual level.

As a result of modeling noise, the CWA model can account for the fact that fallacy rates vary as a function of component probabilities. For example if one event is likely and the other event is unlikely, the resulting conjunctive probability will be high with respect to the unlikely event. As a result, noise is unlikely to produce a normative judgment and conjunction fallacies will be prevalent. When both events are likely or unlikely, the CWA model correctly predicts decreased conjunction fallacies. In this case, noise is likely to produce a normative judgment by chance. A similar argument can be made for disjunctive probabilities. Unlike other models reviewed to this point, the CWA model accounts for the low rate of double conjunction and disjunction fallacies. As will be detailed below, the CWA model predicts stochastic adherence to the addition law.

Although the CWA model is a marked improvement over its predecessors, it has difficulty accounting for some phenomena. For example, it does not provide an account for subadditivity and the increased fallacy rate associated with a causal relationship between the component probabilities. In the latter case, the weighting parameters could be adjusted. However, it is not clear whether the weighting parameter would have a clear psychological interpretation.

Probability Theory + Variation Model

At an abstract, computational level (Marr & Vision, 1982), the Probability Theory + Variation (PTV) model posits that the mind computes probabilities in a manner consistent with probability theory (Costello, 2009a). Errors in judgment result from the perturbation of noise in the underlying cognitive processes. It is important to note that the PTV model does not assume people *explicitly* follow the rules of probability theory. If that were the case, no errors in judgment should be observed. Instead, the model assumes the cognitive processes are consistent with the rules of probability theory. Considering the numerous violations of probability theory that have been observed, it may seem odd to propose that the mind reasons in accordance to the rules of probability theory. However, there are several reasons why the PTV model is a good candidate for a more comprehensive model. First, as explained below, the PTV model can account for several key findings in probability judgment. Second, the PTV model makes strong, novel predictions that are tested in the experiments reported below. Third, the PTV model can be reformulated as a cognitive process model to make predictions about the time course of the judgment process (see Experiment 3). Finally, other successful models have incorporated noise into a normative framework to provide more accurate accounts of judgment and decision making. For example, Erev et al. (1994) proposed a stochastic model of calibration to account for the seemingly paradoxical finding in which judgments of simple events (e.g. the probability of rain) exhibit underconfidence or overconfidence on similar tasks. For example, underconfidence occurs when judgments are less extreme than their objective probabilities. The stochastic calibration model proposes a two-stage judgment process in which a noisy covert (i.e. internal) confidence judgment is mapped onto an overt probability judgment. The process of mapping an unbounded confidence judgment onto a bounded probability scale causes judgments to regress toward .50. Regression produces underconfidence or overconfidence, depending on the task. Decision Field Theory is another example of a stochastic model that is instantiated in a normative framework (DFT; Roe, Busemeyer & Townsend, 2001). DFT is built upon the classic weighted utility model in which attributes for each option are weighted according to importance and summed into an overall value called a valence. Unlike the classic weighted utility model, the weighting of the attributes is governed by a stochastic attention switching process. Preference for each option accumulates stochastically until a decision threshold is met, at which point the winning option is selected. By incorporating a stochastic (i.e. noisy) attention process, DFT is able to successfully account for speed-accuracy tradeoffs, stochastic choice and classic preference reversals (with additional components). Along similar lines, the PTV model can account for several established phenomena in probability judgment by incorporating noise with the rules of probability theory.

Subjective probabilities can be represented in the PTV model using Equations 14 and 15. Unlike the CWA model, however, joint probabilities are integrated according to the rules of probability. Predictions are derived from the PTV model through algebraic manipulation of the rules of probability theory. As an example, consider the conjunction fallacy represented in terms of the PTV model:

$$[\mathrm{P(A)} + \mathrm{e_A}]\left[\mathrm{P(B|A)} + \mathrm{e_{B|A}}\right] > [\mathrm{P(B)} + \mathrm{e_B}] \tag{18}$$

assuming P(A) > P(B). An important critical property of the PTV model is that error rates will increase as noise increases (i.e. the error terms). Thus, in the absence of noise, judgments should adhere perfectly to the rules of probability theory. By contrast, the CWA model predicts conjunction and disjunction fallacies will decrease as noise increases. As previously noted, the fallacy rates vary as a function of component and conditional probabilities. The PTV model account for these effects in the following manner. As P(A) increases, the chance that random noise will lead to a conjunction fallacy increases. In other words, the left hand side becomes larger and approaches the logical upper boundary in which $\mathrm{P(A \cap B)} = \mathrm{P(B)}$. Conversely, as P(B) decreases, the chance of a conjunction fallacy increases. A similar account can be made for the influence of the conditional probability. As P(B|A) increases, the right hand side will increase relative to the left hand side. As the right hand side increases, noise is more and more likely to produce a conjunction fallacy. Because the fallacies are influenced by the same factors, this reasoning can be extended from the conjunction fallacy to the other fallacies (Costello, 2009a; Costello, 2009b). Supporting this idea, Costello (2009b) found that the rate of conjunction and disjunction fallacies is highly correlated.

The PTV model predicts that judgments will adhere stochastically to the addition law, which can be seen by substituting subjective probabilities into Equation 8 and rearranging. Costello & Watts (2013) found empirical support for stochastic adherence to the addition law. Judgments were distributed around zero when they were combined according to Equation 8. More recently, a version of the PTV model has been proposed to account for subadditivity (Costello & Watts, 2013). The model able to account for two key findings: binary complementarity and increased subadditivity as the superordinate set is partitioned into more subsets.

The previously reviewed models share one shortcoming: they can only account for a subset of the findings individually. For example, averaging models can account for the conjunction and disjunction fallacies, but not the double conjunction and disjunction fallacies. FTT can account for semantic coherence and the four fallacies, but does not have the specificity to account for other phenomena, such as the addition law and the influence of component and conditional probabilities on fallacy rates. The PTV model provides an account of these key findings and makes new predictions as well.

Experiment 1

Experiment 1 was designed to achieve three goals. The first goal was to make critical comparisons between the PTV model and the CWA model regarding the relationship between noise and errors. As explained in more detail below, the PTV model predicts that errors should increase as noises increases whereas the CWA model makes the opposite prediction in some cases. A test-retest approach was adopted to address this question. Participants made judgments for each problem twice so that intra-judgment variance could be estimated. The second goal of Experiment 1 was to test whether the variance sum law holds for the integration rules of the PTV model. As explained further below, the integration rules of the PTV model imply more noise in disjunctive probabilities than conjunctive probabilities. The third goal was to replicate and extend previous findings, such as stochastic adherence to the additive law. The predictions for Experiment 1 are summarized below in Table 1.

Table 1. Summary of model predictions for Experiment 1.

Prediction	PTV	CWA
Correlation Conjunction and Disjunction Fallacy	+	+
Correlation Between Noise and Errors		
Conjunction Fallacy	+	-
Disjunction Fallacy	+	-
Double Conjunction Fallacy	+	+
Double Disjunction Fallacy	+	+
Minimum Conjunction Error	+	+
Maximum Disjunction Error	+	+
Semantic Coherence	-	+
Order Effects: Joint Probabilities First vs. Last		
Noise in Conjunctions	>	>
Noise in Disjunctions	>	>
Conjunction Fallacy	>	<
Disjunction Fallacy	>	<
Double Conjunction Fallacy	>	>
Double Disjunction Fallacy	>	>
Minimum Conjunction Error	>	>
Maximum Disjunction Error	>	>
Semantic Coherence	<	>
Addition Law		
Expectation	~0	~0
Correlation Noise and Absolute Deviation	+	+
Noise in Conjunction vs. Disjunction	≤	NP

Note: divergent predictions are italicized. NP: no prediction.

Relationship Between Fallacies

As previously noted, the conjunction and disjunction fallacies are a function of component and conditional probabilities (e.g Fisk, 2002; Costello, 2009a; Costello, 2009b). One prediction that follows from this finding is that the rate of conjunction and disjunction fallacies should be correlated. Both models make this prediction, which has been supported previously (e.g. Costello, 2009b).

Noise and Errors

In some cases, the PTV model and CWA model make divergent predictions regarding the relationship between noise and errors (see Table 1). The expected rank order of judgments for each model is provided below to explicate the derivations of the predictions. Assuming S(A) > S(B), the PTV model predicts the following rank order:

$$\mathrm{S}(\mathrm{A} \cup \mathrm{B}) \geq \mathrm{S}(\mathrm{A}) > S(B) \geq S(\mathrm{A} \cap \mathrm{B}) \tag{19}$$

By contrast, the CWA model predicts a different rank order:

$$\mathrm{S}(\mathrm{A}) > S(\mathrm{A} \cup \mathrm{B}) > S(\mathrm{A} \cap \mathrm{B}) > S(B) \tag{20}$$

According to both models, deviations from the predicted rank orders are due to noise. The PTV model makes a consistent prediction regarding the relationship between noise and errors: increasing noise will increase errors. By contrast, the CWA model makes different predictions depending on the specific error under investigation.

Whereas the PTV model predicts that conjunction fallacies will increase with noise, the CWA model predicts that conjunction fallacies will decrease with noise. Inspection of the expected rank order of the CWA model reveals that in the absence of noise, a conjunction fallacy is predicted: $\mathrm{S}(\mathrm{A} \cap \mathrm{B}) > S(\mathrm{B})$. Thus, increasing noise will produce more normative responses by chance. In support of the CWA model, Nilsson et al. (2009) found that averaging multiple judgments from the same person on the same problem increased rather than decreased the conjunction and disjunction fallacy. However, others have argued that this averaging procedure does not reduce noise, but instead reduces the proportional difference (Costello and Watts, 2013). In light of this potential problem, I used a straightforward, alternative analysis to evaluate whether increased noise is associated with increased errors. This alternative analysis evaluates the relationship between a person's error rates and the amount of noise in his or her judgments.

The same predictions can be derived for the disjunction fallacy, which occurs when the disjunction is judged as less probable than the larger component probability, S(A). As before, the PTV model predicts that the disjunction fallacy will increase with noise, whereas the CWA model predicts that the disjunction fallacy will decrease with noise.

Next, we turn to the double conjunction and double disjunction fallacies. A double conjunction fallacy occurs when the conjunction is judged as more probable than the high probability component, S(A). A double disjunction fallacy occurs when the disjunction is judged as less probable than the low probability component, S(B). As before, the PTV model predicts that the double conjunction and double disjunction fallacies will increase with noise. The CWA model also predicts that double conjunction and double disjunction fallacies should increase with noise. To see why this is the case, consider the double conjunction fallacy as an example. The CWA model implies $\mathrm{S}(\mathrm{A}) > S(A \cap B)$ in the absence of noise. However, in the presence of noise, the expected rank order may reverse for any given judgment to produce a double conjunction fallacy.

As the minimum conjunction error represents a departure from probability theory, the PTV model predicts it should increase with noise. Similarly, the CWA model predicts the minimum conjunction error will increase with noise because in the absence of noise the model predicts conjunctions that are too large rather than too small. As the maximum disjunction error represents a departure from probability theory, the PTV model predicts it should increase with noise. Similarly, the CWA model predicts the maximum disjunction error will increase with noise because in the absence of noise the model predicts disjunctions that are too small rather too large.

The models make divergent predictions for semantic coherence. Because the rules for semantic coherence are derived from probability theory, the PTV model predicts that increased noise should be associated with lower semantic coherence (for details see Wolfe & Reyna, 2010). By contrast, the CWA model predicts the opposite relationship.

Stochastic Adherence to the Addition Law

The PTV and CWA model both predict that judgments should adhere stochastically to the addition law of probability theory. To see why this is the case for the CWA model, substitute Equations 16 and 17 into Equation 8 and simplify. Although noise will cause individual sets of judgments will deviate from 0, they should be distributed accordingly:

$$\mathrm{S(A)} + \ \mathrm{S(B)} - S(\mathrm{A} \cap \mathrm{B}) - S(\mathrm{A} \cup \mathrm{B}) \sim \mathrm{F}(0, \sigma^2) \qquad (21)$$

where F has a mean of zero. Support for this prediction was found in Costello and Watts (2013). For both models, an untested corollary of this prediction is that the noise in individual sets of judgments should be correlated with the absolute deviation from zero in Equation 21.

Noise in Joint Probabilities

One implication of the integration rules of the PTV model and the variance sum law is that there should be more noise in disjunctive probabilities compared to conjunction probabilities. According to the integration rules of the PTV model, conjunctive and disjunctive probabilities are a function of component and conditional probabilities. According to the variance sum law, the variance of the sum of independent random variables is equal to the sum of the individual variances:

$$\mathrm{Var}[\mathrm{X}_1 + \mathrm{X}_2 + \cdots + \mathrm{X}_\mathrm{N}] = \mathrm{Var}[\mathrm{X}_1] + \mathrm{Var}[\mathrm{X}_2] + \cdots + \mathrm{Var}[\mathrm{X}_\mathrm{N}] \qquad (22)$$

Applying the variance sum law to conjunctions and disjunctions results in:

$$\mathrm{Var}[\mathrm{S}(\mathrm{A} \cap \mathrm{B})] = \mathrm{Var}[\mathrm{S(A)S(B|A)}] \qquad (23)$$

$$\mathrm{Var}[\mathrm{S}(\mathrm{A} \cup \mathrm{B})] = \mathrm{Var}[\mathrm{S(A)}] + \mathrm{Var}[\mathrm{S(B)}] + \mathrm{Var}[\mathrm{S(A)S(B|A)}] \qquad (24)$$

In comparing Equation 23 to Equation 24, it becomes clear that the variance in the conjunction is a subset of the variance in the disjunction. It stands to reason that more noise should be observed in disjunctions, unless $\mathrm{Var}[\mathrm{S(A)}] + \mathrm{Var}[\mathrm{S(B)}] = 0$. The CWA model does not make a simple prediction about the relative noise in conjunctions and disjunctions.

Order Effects

The conjunction fallacy has been found to be higher when the conjunction is rated first followed by its components (Stolarz-Fantino, Fantino, Zizzo, & Wen, 2003). According to the PTV model, attention modulates the amount of noise in the component and conditional probabilities (Costello, 2009a). When the component probabilities are judged first, they are maintained in attention when the conjunction is subsequently judged. As a result, the components become relatively fixed, which, in turn, decreases the chance of a conjunction fallacy. In other words, attention decreases the error terms in Equation 18, thereby making the conjunction fallacy less likely to occur. By contrast, less attention is given to the component and conditional probabilities when the conjunction is judged first. In this case, the judgments are more prone to random variation, which increases the chance of a conjunction fallacy. Because the conjunctions and disjunctions are a function of noisy inputs, this prediction extends to the other errors and semantic coherence. Although Nilsson et al. (2009) did not provide an account of order effects, the attentional mechanism proposed by Costello (2009a) appears to be consistent with the CWA model. For this reason, I extend the attentional mechanism to the CWA model to permit the comparison of the models. To the extent that an order effect is observed, the predictions for the PTV model and CWA model mirror those for the relationship between noise and errors (see Table 1). The models make divergent predictions for conjunction fallacies, disjunction fallacies and semantic coherence for the reasons detailed in Errors and Noise.

Experiment 1

Participants

Participants were 61 introductory psychology students at Miami University, who participated for partial course credit. Consistent with previous studies at this University, participants were disproportionately white and female.

Materials

A pilot study was conducted to develop the problems used in Experiment 1 and Experiment 2. In the interest of brevity, the description of the pilot study is merged with Experiment 1 because both used the same procedures. The problems used in the pilot study were designed to systematically vary low and high component probability combinations, conditional support and set types. Most problems were adopted from published studies and modified as necessary (for example, Crisp & Feeney, 2009; Wolfe & Reyna, 2010; Wolfe, Fisher & Reyna, 2013; Wolfe & Fisher, 2013), while the remaining problems were developed specifically for this study. Each problem featured a short scenario followed by questions for $\mathrm{P(A)}, \mathrm{P(B)}, \mathrm{P(A \cap B)}$ and $\mathrm{P(A \cup B)}$. As an example, consider the following problem: “Steve is 50 years old and has a sedentary lifestyle. He is a movie buff. When he comes home from his job as a computer programmer, he likes to watch movies from his movie collection and eat his favorite ice cream: double fudge, chocolate chip with sprinkles.” The conjunction and disjunction were formed from the two component events: (A) Steve is obese and (B) Steve can do 50 push-ups. This problem was designed to have one high (A) and one low probability event (B), negative conditional support, and depict overlapping sets. The final 34 problems were selected from a larger set of problems developed in the pilot study based on the variety in the component probabilities, conditional support, set type and having sufficient variability between judgments at time 1 and time 2. In a few cases, minor adjustments were made before using the piloted materials in the main experiments. In addition, I included a total of 22 filler probability judgment problems. Finally, an argumentation filler task consisted of a subset of 25 simple arguments adopted from (Wolfe & Britt, 2008).

Procedures

Participants completed the study individually on computers in groups ranging from one to five. A typical completion time ranged from 45-55 minutes. Participants completed two blocks of judgments consisting of 34 target problems and 11 filler problems presented in randomized order. The 34 target problems were presented in both blocks of judgments, but a different set of 11 filler items was used in each block. Each problem consisted of a short description followed by questions for $P(A), P(B), P(A \cap B)$ and $P(A \cup B)$, which were presented individually. For each judgment, participants entered a number between 0 and 100. The judgments were blocked by component probabilities and joint probabilities, with items randomized within each block. To test the order effects, judgment blocks of component and joint probabilities were counterbalanced across participants. After completing the first block of judgments, participants completed the filler argumentation task to interfere with memory. After reading each argument, participants rated the argument in terms of personal agreement and argument strength on a Likert scale. Lastly, participants completed the second block of judgments.

Results

Two participants were excluded due to a computer error. An additional five participants were excluded because they failed to complete experiment within the allotted time, resulting in a total of 54 participants. Before turning to the primary analyses, I examined whether differences between time 1 and time 2 judgments were systematic or simply represented noise. The mean of each judgment type on each problem was computed across participants for time 1 and time 2, resulting in 136 averaged judgments at each time point (34 problems X 4 judgment types). The difference in mean judgments (mean = -.005, SD = .039) was not statistically significant, t(135) = -1.63, p = .11. The small observed difference provides evidence that difference in judgments were noise rather than systematic. According to the PTV and CWA model, the rate of conjunction and disjunction fallacies should be correlated because they are influenced by the same factors. Consistent with Costello (2009b), the correlation between the rate of conjunction and disjunction fallacies was r(32) = .86, p <.001 across problems.

Noise and Errors

One critical property of the PTV model is that error rates should increase as noise in judgments increases. To test this critical property, I computed the correlations between each participant’s error rate and his or her average judgment noise. An error rate for each participant was computed as the number of errors committed within the 68 problems (34 problems X 2 replications). For each participant, average judgment noise was computed as the mean absolute deviation between judgments at time 1 and time 2 across all problems, judgment types and replications. Thus, there was one error rate and one mean absolute deviation value for each participant. The results are summarized below in Table 2. Consistent with both models, there was a correlation between double conjunction and disjunction fallacies and overall judgment noise. However, the negative correlation between semantic coherence and overall judgment noise uniquely supports the PTV model. The remaining correlations failed to reach statistical significance.

Table 2. Correlations between noise and error rates in Experiment 1.

Error	r	p-value	Mean Rate	Standard Deviation Rate
Conjunction Fallacy	.02	.89	.14	.08
Disjunction Fallacy	.06	.68	.12	.07
Double Conjunction Fallacy	.40	.003	.08	.06
Double Disjunction Fallacy	.48	<.001	.08	.08
Minimum Conjunction Error	-.06	.68	.17	.13
Maximum Disjunction Error	.20	.14	.12	.11
Semantic Coherence	-.28	.04	.11	.07
Sum of Errors	.34	.01	.54	.23

N = 54

Order Effects

The PTV model and the extension of the CWA model predict that judging joint probabilities before component probabilities will increase noise in the joint probabilities. The mean absolute difference for the conjunctions and disjunctions were computed across all problems for each participant, resulting in a mean absolute deviation for conjunctions and one mean absolute deviation for disjunctions per participant. Table 3 shows that judgment order increased noise for conjunctions in the predicted direction. The increase for disjunctions was in the predicted direction, but not statistically significant. The most diagnostic result is the increase in conjunction and disjunction fallacies when the joint probabilities are judged first. This finding is consistent with the PTV model, but not with the CWA model. Except for the minimum conjunction error, the differences in error and semantic coherence rates were non-significant but in the direction predicted by the PTV model.

The Addition Law

Both models predict stochastic adherence to the addition law. To test whether the addition law holds, F = $\mathrm{S(A)} + \mathrm{S(B)} - \mathrm{S(A \cap B)} - \mathrm{S(A \cup B)}$ was computed for each participant's judgments on each problem (see Equation 21). Due to the high number of observations (54 participants X 34 problems X 2 replications = sets 3,672), the predictions for the addition law were more precisely evaluated with confidence intervals. There was a small but systematic deviation from the predicted mean of 0, mean = .025, SD = .23, 95% CI [.015, .034]. Although the mean shows a systematic deviation from 0, it is a small deviation in comparison to the full range of possible values [-2 2]. Similarly, the deviation is small in terms of a standardized effect size (d = .09). One corollary of the addition law is that the variability in the distribution of F should be associated with the amount of noise in the individual judgments. Two values were computed for each problem completed by each participant. The first value was composite noise, which represented the noise in the individual judgments $\mathrm{S(A)} + \mathrm{S(B)} -$

$S(A \cap B) - S(A \cup B)$ between time 1 and time 2. Composite noise was formed by summing the absolute deviations between time 1 and time 2 judgments for each problem completed by each participant. As an example, the composite noise for a subject on a given problem would be $|S(A)_1 - S(A)_2| + |S(B)_1 - S(B)_2| + |S(A \cap B)_1 - S(A \cap B)_2| + |S(A \cup B)_1 - S(A \cup B)_2|$, where the subscripts refer to time 1 and time 2. Second, the mean of the absolute deviations of F at time 1 and time 2 represented how much judgments deviated from the addition law. As an example, the mean absolute deviation for a participant on a given problem was computed as $\frac{|F_1|+|F_2|}{2}$. Separate correlations between composite noise and mean absolute deviations were computed for each participant. As predicted by the model, the average correlation across participants was in the predicted direction, M = .32, SD = .19, 95% CI[.26, .37].

Noise in Joint Probabilities

The PTV model predicts that the noise should be greater in disjunctions than conjunctions because disjunctions are a function of more random variables. To test this prediction, the variances for the conjunctions and disjunctions were computed for each set of judgments (54 participants X 34 problems = 1836 sets). Contrary to the PTV model, the mean difference was essentially zero, mean difference = .001, 95% CI[-.005, .002]. The expected difference in conjunctive and disjunctive variances should approximate the sum of the unique variance terms in Equation 24: Var[S(A)] + Var[S(B)]. The sum of the mean variances was .052, which is much larger than the observed difference.

Model Fit

The PTV model was fit to aggregated data using the basic procedures described in Costello (2009a) (see Appendix 1 for details). In brief, the error and semantic coherence rates were estimated for each problem through simulation of the subjective probabilities. The mean and standard deviations of the aggregated judgments were allowed to vary within $\pm$ .02 of their observed values to adjust for sampling error while sufficiently constraining the model. The corresponding mean and standard deviations for conditional probability judgments were taken from Experiment 2. The results are summarized in Table 4. The mean absolute difference between predicted and observed rates was .07 (SD = .09), which was larger in comparison to the results reported in Costello (2009a). The correlation between predicted and observed rates was r = .47. The model performed relatively well on independent and overlapping problems in comparison to identical, mutually exclusive and subset problems. There was a general tendency to over-estimate the rate of maximum disjunction errors and under-estimate semantic coherence, which was particularly pronounced for identical problems.

Table 4. Quantitative predictions for the PTV model Experiment 1.

	CF		DF		DCF		DDF		MCE		MDE		SC		MAD
Identical	.01	(.08)	.01	(.08)	.09	(.17)	.08	(.17)	.16	(.14)	.03	(.11)	.41	(.14)	.11
Independent	.27	(.25)	.20	(.25)	.05	(.04)	.08	(.04)	.04	(.04)	.05	(.17)	.01	(.00)	.05
Mutually Exclusive	.02	(.11)	.04	(.11)	.06	(.06)	.06	(.06)	.10	(.26)	.06	(.42)	.14	(.02)	.12
Overlapping	.20	(.19)	.17	(.19)	.11	(.08)	.08	(.08)	.04	(.05)	.06	(.16)	.05	(.02)	.04
Subset	.08	(.12)	.09	(.12)	.08	(.12)	.08	(.12)	.07	(.09)	.04	(.16)	.12	(.00)	.08
MAD	.05		.05		.05		.05		.06		.15		.09		.07

Predicted rates are in parentheses. CF = Conjunction Fallacy; DF = Disjunction Fallacy; DCF = Double Conjunction Fallacy; DDF = Double Disjunction Fallacy; MCE = Minimum Conjunction Error; MDE= Maximum Disjunction Error; SC = Semantic Coherence; MAD = Mean Absolute Difference.

Discussion

Experiment 1 was designed to achieve two primary goals. One goal was to make critical comparisons between the PTV model and the CWA model. The second goal was to test a critical property of the PTV model derived from the variance sum law. The PTV model and CWA model offer differing accounts of the relationship between noise and errors. According to the PTV model, judgments are generated from the rules of probability theory but errors result from noise in judgments. By contrast, the CWA model assumes that conjunctive and disjunctive probabilities are formed through a weighted average of noisy component probabilities. In stark contrast to the PTV model, the integration rules of the CWA model imply that the conjunction and disjunction fallacy should decrease as noise increases. In cases where the models made divergent predictions, the PTV model generally received better support. Consistent with the PTV model, semantic coherence decreased as noise increased. The PTV model provided a better account of the order effects compared to the CWA model. As predicted by the PTV model, conjunction and disjunction fallacies increased when order effects produced more noise in the conjunctions. Several of the results were consistent with both models. The double conjunction and double disjunction fallacies were associated with higher levels of judgment noise. In addition, judgments adhered to the addition law as predicted by both models.

The second goal of Experiment 1 was to test a critical property of the PTV model derived from the variance sum law and the integration rules of the PTV model. The critical property states that the variance in disjunctions should be greater than the variance in conjunctions, except in a special case in which the variance of the component probabilities are both zero. In Experiment 1, the variances for conjunctions and disjunctions were virtually equivalent and the variance of the component probabilities were much larger than zero. Jointly, these results provide strong evidence against the PTV model.

The quantitative tests revealed mixed support for the PTV model. While the PTV model performed relatively well on overlapping and independent set problems, it had more difficulty accounting for identical, mutually exclusive and subset problems. In addition, the PTV model generally had trouble accounting for the maximum disjunction error and semantic coherence. The difficulty accounting for semantic coherence was particularly pronounced for problems featuring identical sets. One possible reason for this failure is that identical sets are rare in judgment space, but are observed empirically at high rates (Wolfe & Reyna, 2010). For example, assuming judgments are multiples of .05, there are 21^4 permutations of P(A), P(B), $\mathrm{P}(\mathrm{A} \cap \mathrm{B})$ and $\mathrm{P}(\mathrm{A} \cup \mathrm{B})$. Only 20 of the permutations constitute identical sets. For this reason, it is difficult for the PTV model to produce semantic coherence for identical sets.

In summary, Experiment 1 found mixed support for the PTV model. Compared to the CWA model, the PTV model provided a better account of order effects and the relationship between noise and errors. However, the PTV model was not fully supported in absolute terms as several of the effects were small and not statistically significant. The PTV model failed a critical test derived from the variance sum law in which the variance in disjunctions was predicted to be greater than the variance in conjunctions. Contrary to the PTV model, the variances were essentially equivalent. In addition, the PTV model provided a poor quantitative account of several aspects of judgment, including identical sets, mutually exclusive sets and the maximum disjunction error.

Experiment 2

The primary goal of Experiment 2 was to extend tests of the PTV model to conditional probability judgment, using the same procedures as Experiment 1. For this reason, many of the predictions parallel those described for joint probability judgment. In particular, Experiment 2 tests the critical property of the PTV model that noise should be associated with more errors in conditional probabilities. A second goal was to examine whether judgments adhere stochastically to Bayes' theorem using a test analogous to the addition law test in Experiment 1. Each of the predictions is described in detail below.

Noise and Errors

As with joint probability judgment, I tested whether increased noise is associated with increased errors. The PTV predicts that judgments should adhere stochastically to Bayes' theorem, which is defined as

$$\mathrm{P}(\mathrm{A}|\mathrm{B}) = \frac{\mathrm{P}(\mathrm{B}|\mathrm{A})\mathrm{P}(\mathrm{A})}{\mathrm{P}(\mathrm{B})} \tag{25}$$

A test of Bayes' theorem was formed by multiplying both sides of Equation 25 by P(B), subtracting the right hand side from the left hand side and expressing each judgment as a subjective probability:

$$S(B)S(A|B) - S(A)S(B|A) \sim G(0, \sigma^2) \qquad (26)$$

Thus, the PTV model predicts a distribution with a mean of zero if judgments adhere stochastically to Baye's theorem. In addition, the PTV model predicts that absolute deviations from zero should be correlated with the amount of judgment noise.

Methods

Participants

Participants were 62 introductory psychology students at Miami University, who completed the experiment for partial course credit.

Materials and Procedures

Experiment 2 used the same materials and procedures as Experiment 1, with one exception. In Experiment 2, participants judged the probability of events A, B, A|B, B|A instead of $A, B, A \cap B$ and $A \cup B$.

Results

Three participants were excluded due to a computer error. Four additional participants were excluded because they failed to complete the experiment within the allotted time, resulting in a total of 55 participants. The difference in mean judgments was compared for time 1 and time 2 across problems and judgments. A difference (mean = -.009, SD = .035) was detected between time 1 and time 2, t(135) = -3.10, p = .002. However, given the small magnitude of the difference, most of the variability in judgments appears to be noise rather than systematic.

Noise and Errors

According to the PTV model, increased noise should be associated with increased errors and decreased semantic coherence. To test this critical property, I computed the correlations between each participant's error rate and his or her average judgment noise. An error rate for each participant was as the rate of errors committed within the 68 problems (34 problems X 2 replications). For each participant, average judgment noise was computed as the mean absolute difference between judgments at time 1 and time 2 across all problems, judgment types and replications. As shown in Table 5, the prediction held for conditional reversals and semantic coherence. However, the remaining correlations failed to reach statistical significance.

Table 5. Correlations between noise and error rates for Experiment 2.

Error	r	p-value	Mean Rate	Standard Deviation Rate
Minimum Conditional Error A|B	.18	.19	.26	.14
Minimum Conditional Error B|A	.15	.27	.28	.16
Conditional Reversal	.31	.02	.08	.07
Conversion Error	.17	.22	.14	.06
Semantic Coherence	-.43	.001	.24	.09
Sum of Errors	.11	.41	.41	.19

N = 55

Stochastic Adherence to Bayes' Theorem

Bayes' theorem was tested in similar manner to the addition law. G = S(A)S(B|A) – S(B)S(A|B) was computed for each set of judgments (55 participants X 34 problems X 2 replications = 3,850 sets) to test whether Bayes' theorem holds stochastically. Although there was a systematic deviation, the difference was very close to zero, -.008, SD = .12, 95% CI[-.012, -.003], as predicted by the PTV model. Next, I tested whether the noise in judgments is related to the variability in the distribution. As in Experiment 1, two values were computed each for problem completed by each participant. The first value was composite noise, which represented the noise in the individual judgments $S(\mathrm{A}) + S(\mathrm{B}) - S(\mathrm{A} \cap \mathrm{B}) - S(\mathrm{A} \cup \mathrm{B})$ between time 1 and time 2. Composite noise was formed by summing the absolute deviations between time 1 and time 2 judgments for each problem completed by each participant. As an example, the composite noise for a subject on a given problem would be $|\mathrm{S(A)}_1 - \mathrm{S(A)}_2| + |\mathrm{S(B)}_1 - \mathrm{S(B)}_2| + |\mathrm{S(A|B)}_1 - \mathrm{S(A|B)}_2| + |\mathrm{S(B|A)}_1 - \mathrm{S(B|A)}_2|$, where the subscripts refer to time 1 and time 2. Second, the mean absolute deviation of G at time 1 and time 2 represented how much judgments deviated from Bayes' theorem. As an example, the mean absolute deviation for a participant on a given problem was computed as $\frac{|\mathrm{G}_1| + |\mathrm{G}_2|}{2}$. Separate correlations between composite noise and mean absolute deviations were computed for each participant. Consistent with the model, the mean correlation across participants was M = .27, SD = .17, 95% CI[.22, .31].

Model Fit

The model was fit to aggregated data for each of the 34 problems following a similar procedure used in Experiment 1 (see Appendix 1). The observed and predicted rates can be found in Table 6. Compared to Experiment 1, the absolute deviations were somewhat larger, mean = .09 (SD = .12) and the correlation between predicted and observed rates was somewhat lower, r = .38. Compared to overlapping and subset problems, the model performed relatively poorly on identical, independent and mutually exclusive problems. Moreover, the model had difficulty accounting for semantic coherence, particularly for identical and independent sets, where the discrepancy was .65 and .30, respectively. Finally, the PTV model performed poorly on conversion errors. In general, the model predicts conditional reversals should be higher than conversion errors. However, the opposite trend was observed.

Table 6. Quantitative Predictions for Experiment 2.

	MCE A|B		MCE B|A		CR		CE		SC		MAD
Identical	.04	(.07)	.03	(.11)	.03	(.02)	.06	(.09)	.65	(.00)	.16
Independent	.17	(.21)	.17	(.22)	.11	(.16)	.23	(.05)	.30	(.00)	.12
Mutually Exclusive	.14	(.24)	.15	(.24)	.03	(.14)	.08	(.02)	.52	(.38)	10
Overlapping	.08	(.12)	.08	(.11)	.12	(.16)	.20	(.07)	.06	(.02)	.06
Subset	.02	(.15)	.05	(.12)	.06	(.07)	.05	(.02)	.00	(.01)	.06
MAD	.08		.06		.05		.10		.17		.09

Predicted rates are in parentheses. MCE A|B: minimum conditional error for P(A|B); MCE B|A: minimum conditional error for P(B|A); CR: conditional reversal; CE: conversion error; SC: semantic coherence. MAD: Mean Absolute Difference.

Discussion

Experiment 2 extended tests of the PTV model to conditional probability judgment and was conducted to achieve two primary goals. One goal was to test the PTV model's account of noise and errors in the realm of conditional probability judgment. The second goal was to test the novel prediction that judgments adhere stochastically to Bayes' Theorem. By and large, the results of Experiment 2 mirrored those from Experiment 1. Some support was found for the PTV model's critical property in which error rates increase as noise increases. As predicted, increased noise was associated with more conditional reversals and less semantic coherence. The correlations between noise and the other errors were in the predicted direction but less conclusive because they were non-significant.

In Experiment 1, there was some support for the notion that judgments adhere stochastically to the addition law. Experiment 2 tested an analogue prediction that conditional judgments combine in a manner consistent with Bayes' theorem. Two lines of evidence suggest that judgments adhere stochastically to Bayes' theorem. First, when the judgments were combined according to Equation 26, they were approximately distributed around zero. Second, the variance in the resulting distribution was correlated with noise in judgments. These results are consistent with the notion that noisy judgments were generated from a stochastic process that follows the rules of probability theory.

However, two lines of evidence were at odds with the PTV model. First, the PTV model had difficulty accounting for conversion errors (i.e. confusing P(A|B) with P(B|A). In general, the model predicts that conditional reversals should be more likely than conversion errors, except for cases in which judgments are close the boundary of 1 (e.g. identical sets). However, conversion errors were generally higher than conditional reversals. Except for identical sets, the model tended to under-predict conversion errors and over-predict conditional reversals. The rate of conversion errors is more consistent with the concept of denominator neglect (e.g. Reyna & Brainerd, 2008). Second, replicating Experiment 1, the model greatly under-predicted the rate of semantic coherence for identical sets. In particular, the observed rate was .65 whereas the predicted rate was 0. Along similar lines, the PTV model also greatly under-predicted semantic coherence for independent sets. One reason for this failure is that semantic coherence for identical and independent sets are rare in the judgment space, making it unlikely that semantic coherence can be achieved through a purely random process. By contrast, the PTV model can account for semantic coherence on subsets and overlapping sets because they are relatively more common in the judgment space and lower semantic coherence rates are found empirically (see Wolfe, Fisher, & Reyna, 2013). Taken together, these results suggest that judgment noise alone is insufficient to account for the high rates of semantic coherence.

In summary, the mixed pattern of results in Experiment 2 resembles that of Experiment 1. There was some but limited support for the critical property that errors are associated with noise. Paralleling the results for the addition law test in Experiment 1, there was strong evidence that judgments adhere stochastically to Bayes' theorem. However, two lines of evidence do not support the PTV model. First, contrary to the PTV model, the rate of conversion errors exceed that of conditional reversals. In general, one would expect relatively few cases in which the conditionals are rated as equal. Second, the PTV model failed to provide a quantitative account of semantic coherence for independent and identical sets.

Experiment 3

Up to this point, the PTV model has been described and tested at a computational level. Noise in the judgments has been treated as a purely statistical phenomenon without specifying an underlying cognitive process. The primary goal of Experiment 3 was to provide the initial groundwork for casting the PTV model as a cognitive process model. Developing a process model would impose cognitively plausible constraints and allow predictions to be derived about the time course of the judgment process. One possibility is that judgments are based on a dynamic memory retrieval process in which judgments converge on a stable value over time. Exemplars for a given event might be sampled from memory until a desired level of precision is achieved, at which point a judgment is submitted. Several findings in the literature provide indirect but converging support for this possibility. For example, MINERVA-DM explains various judgment phenomena using memory encoding and retrieval processes (Dougherty, Gettys, & Ogden, 1999). MINERVA-DM built upon the framework of MINERVA2 (Hintzman, 1984), an instance-based memory model. According to MINERVA-DM, exemplars (also called traces) are encoded in memory as degraded copies and later retrieved on the basis of similarity to the target event being judged (Dougherty, Gettys, & Ogden, 1999). One limitation of MINERVA-DM is that it does not model the time course of encoding and retrieval processes. A similar model—the Generalized Context Model—has been instantiated in a random walk framework to model dynamic exemplar retrieval processes in perceptual categorization (Nosofsky, & Palmeri, 1997). The model was successful in jointly accounting for speed-accuracy tradeoffs, similarity effects, response time and choice distributions. A similar process may underlie probability judgment in the PTV model.

As an initial starting point, Experiment 3 considers a simple dynamic memory retrieval model of probability judgment. Let $\mathrm{A} = [\mathrm{a}_1, \mathrm{a}_2 \ ... \mathrm{a}_\mathrm{N}]$ be a binary vector of exemplars for event A, in which $\mathrm{a}_\mathrm{t} = 1$ if the event occurred and $\mathrm{a}_\mathrm{t} = 0$ if the event did not occur. According to the model, an exemplar is sampled from memory at every time point, t, until a precision threshold or an externally imposed time limit is researched. Upon termination of the retrieval phase, the N exemplars are normalized to form a probability judgment:

$$\mathrm{S(A)} = \frac{1}{\mathrm{N}}\sum_{\mathrm{t}=1}^{\mathrm{N}} \mathrm{a_t} \tag{27}$$

The precision threshold represents the precision or confidence required for a given judgment. A stringent threshold can be set when a high degree of precision is required. Alternatively, a less stringent threshold can be set when a quick judgment must be made or precision is not important. The threshold is formalized as the desired variance in the judgment estimate, C. At any given point in time, t, the confidence of the probability judgment is defined as the variance estimator:

$$\mathrm{C_t} = \mathrm{Var[S(A)]_t} = \frac{\mathrm{P(A)(1-P(A))}}{\mathrm{t}} \tag{28}$$

A judgment is made when $\mathrm{C_t} \leq$ C. Although this model is admittedly simplistic, it enables important qualitative predictions regarding speed-accuracy tradeoffs and is proposed as a basic starting point. According to the model, noise should decrease as more exemplars are dynamically accumulated. The purpose of Experiment 3 was to test the qualitative prediction that noise decreases as a function of time. A response deadline was instated to determine whether judgments made more quickly were more variable.

Pilot Studies

Two pilot studies were conducted in order to select a response deadline that was sufficiently challenging to increase noise in the judgments without producing haphazard responding. Little is known about the reaction times for probability judgments. To address this issue, reaction times where recorded in Pilot Study 1 while self-paced judgments were made. Two response deadlines were tested in Pilot Study 2 based on the median reaction times in Pilot Study 1.

Pilot Study 1

Participants

Participants were 26 introductory psychology students at Miami University, who completed the experiment for partial course credit.

Materials and Procedures

The probability judgment task consisted of ten target problems that were selected from Experiments 1 and 2 and previous studies (e.g. Tversky & Kahneman, 1983). Two criteria were used in selecting the target problems. The problems were selected on the basis of intermediate error rates to avoid floor or ceiling effects. Second, problems were chosen that had simple events (e.g. bank teller) to reduce variability in reaction time due to reading latency. One problem depicted independent sets, one depicted mutually exclusive sets and the remaining eight problems depicted overlapping sets. Each problem featured a short scenario followed by questions for $P(A), P(B), P(A \cap B), P(A \cup B), P(A|B)$ and $P(B|A)$. Each judgment was presented on a separate screen in a randomized order. In addition to the target questions, four catch problems were included to differentiate between purposeful and haphazard responding (Wolfe & Fisher, 2013). The catch questions have objectively correct answers that can be used as a quality metric. For example, consider the following: "Richard is an avid skier and spends 90% of his vacations skiing. Today he has plane tickets to Aspen, Colorado and has been looking forward to this weekend trip for months. Unfortunately, Richard had a bad accident and both of his legs are broken. What is the probability that Richard will go skiing this weekend?" The correct answer to this question is 0. Before completing the experiments, participants were given brief instructions explaining the types of judgments they would be making (e.g. component, joint and conditional). Participants completed two blocks of the same problems at a self-selected pace. The problems were randomized in each judgment block. The judgment blocks were separated by a filler argumentation task to reduce participants' ability to remember their initial judgments (Wolfe & Britt, 2008).

Results

The median reaction times were approximately 4 seconds, 5 seconds and 6 seconds for component, joint and conditional probability judgments, respectively. These reaction times formed the basis for the response deadlines used in Pilot Study 2.

Pilot Study 2

Participants

Participants were 53 introductory psychology students at Miami University, who completed the experiment for partial course credit.

Materials and Procedures

Pilot Study 2 used the same materials and procedures used in Pilot Study 1, with two exceptions. First, participants began with two practice problems to familiarize themselves with the judgment task. Second, on each trial, a countdown clock was displayed below the response entry box with the phrase "X seconds remaining" in black text. The countdown clock descended to zero in increments of 1 second. When the countdown clock reached 0, the text turned red. If the judgment was not submitted within the allotted time, participants received feedback on the subsequent screen encouraging them to respond faster. Different deadlines were used for each judgment type to adjust for reading time. The first condition used the median reaction times observed in Pilot Study 1 for component (4 seconds), joint (5 seconds) and conditional probability judgments (6 seconds). The second condition added one second to each of the response deadlines (e.g. 5, 6 and 7 seconds).

Results

To determine whether participants were able to respond within the deadline, a success rate was computed for each participant. Participants with the fast deadline (M = .98; SD= .04) and the slower deadline (M = .94, SD = .08) responded within the allotted time at high rates. To examine the quality of the judgments further, the rate of correct catch questions was compared to the no-deadline condition in Pilot Study 1. Compared to the no-deadline condition in Pilot Study 1 (M = .65, SD = .23), the proportion of correct responses in the fast deadline condition (M = .54, SD = .26). However this difference did not achieve statistical significance at a conventional level of .05, t(51) = 1.68, p = .10. Similarly, the slower deadline condition (M = .63, SD = .18) was not statistically different from the no-deadline condition, t(50) = -.46, p = .64. Finally, the proportion of correct responses in the fast and slower deadline conditions did not differ statistically t(51) = -1.41, p = .16.

Next I examined whether the response deadline successfully increased the noise in judgments. For each participant, the mean absolute deviation between time 1 and time 2 was computed across problems and judgment types. Thus, 120 judgments contributed to each participant's mean absolute deviation (10 problem X 6 judgment types X 2 replications). Judgments in the fast deadline condition were more variable (M = .18, SD = .06) than the no deadline condition (M = .13, SD =.04), t(51) = 2.81, p = .007. Similarly, judgments in the slower deadline condition were more variable (M = .17, SD = .05) than the no deadline condition, t(50) = 2.54, p = .01. The two deadline conditions were not statistically different, t(51) = .47, p = .64.

Taken together, these results suggest that the response deadline successfully increased variability in judgments without an appreciable decrease in quality. Although the fast and slow deadlines did not differ at a statistically significant level, the slow deadline was selected for Experiment 3 because the quality was better according to the descriptive statistics.

Main Experiment

Method

Participants

Participants were 60 introductory psychology students at Miami University, who completed the experiment for partial course credit. Data from two participants were excluded because judgments from one block were not properly recorded, resulting in a total of 58 participants.

Materials and Procedures

Participants were randomly assigned to either the deadline or no deadline condition. In the deadline condition, the response deadline was 5, 6, and 7 seconds for component, joint and conditional probability judgments. Otherwise, the materials and procedures in Experiment 3 were identical to those in Pilot Study 2.

Results

Manipulation Check

A success rate for responding within the allotted time was computed for each participant in the deadline condition. Participants in the deadline condition submitted their judgments within the allotted time in the majority of cases (M = .96, SD = .04). Consequentially, reaction times were quicker in the response deadline condition (M = 3.70 seconds; SD = .75 seconds) compared to the no response deadline condition (M = 5.83 seconds; SD = 1.61 seconds), t(56) = 6.43, p < .001. Importantly, the proportion of correct responses for each participant on the catch questions was similar in the no deadline condition (M = .65; SD = .20) and the deadline condition (M = .61; SD = .20), indicating that the quality of judgments was not reduced by the response deadline, t(56) = -.79, p = .43. Next I examined whether the response deadline successfully increased the noise in judgments. For each participant, the mean absolute deviation between time 1 and time 2 was computed across problems and judgment types. Thus, 120 judgments contributed to each participant's mean absolute deviation (10 problem X 6 judgment types X 2 replications). Instating a response deadline lead to more noise in judgment in the deadline condition (M = .20; SD = .07) compared to the no deadline condition (M = .14; SD = .06), t(56) = 3.55, p < .001, d = .95.

Primary Analyses

As in Experiments 1 and 2, mean error rates were computed for each subject. The descriptive and inferential statistics for Experiment 3 are summarized in Table 7. With the exception of semantic coherence for conditional probabilities, no differences were detected between the deadline and no deadline condition. Semantic coherence was reduced when a response deadline was instated. This particular result is consistent with the PTV model.

Table 7. Comparison of mean error rates for the deadline and no deadline conditions.

	No Deadline		Deadline		T-value	P-value
Conditional						
CE	.20	(.12)	.16	(.10)	-1.29	.20
CR	.17	(.11)	.21	(.14)	1.07	.29
MCE A|B	.20	(.11)	.22	(.15)	.55	.58
MCE B|A	.14	(.10)	.16	(.11)	.70	.49
SC	.10	(.08)	.05	(.06)	-2.80	.007
Joint						
CF	.18	(.11)	.16	(.11)	-.77	.48
DF	.17	(.17)	.13	(.11)	-1.07	.29
DCF	.09	(.09)	.12	(.09)	1.43	.16
DDF	.06	(.07)	.11	(.14)	1.77	.08
MCE	.09	(.07)	.09	(.08)	.09	.93
MDE	.06	(.06)	.07	(.07)	.80	.43
SC	.03	(.05)	.03	(.05)	-.12	.90

Standard Deviations are in parentheses. CE: conversion error; CR = conditional reversal; MCE A|B: minimum conditional error for A|B; MCE B|A: minimum conditional error for B|A; SC: semantic coherence; CF = Conjunction Fallacy; DF = Disjunction Fallacy; DCF = Double Conjunction Fallacy; DDF = Double Disjunction Fallacy; MCE = Minimum Conjunction Error; MDE= Maximum Disjunction Error.

Discussion

In Experiment 3, the PTV model was reformulated as a simple, dynamic process model. The model retained the core integration rules of the PTV model, but proposed an exemplar accumulation process through which noise in judgments decreased as a function of time. A response deadline was instated to examine whether quicker judgments were more variable and consequently more error-prone. Instating a response deadline increased the noise in judgments Importantly, the response deadline increased noise without decreasing the quality of the judgments. In particular, participants were able to respond within the deadline in the overwhelming majority of cases and the accuracy of the catch questions did not decrease appreciably. This provides compelling evidence that participants were able to read the problem text and submit a purposeful judgment within the allotted time. By all accounts, the experimental manipulation appeared to be successful.

Despite the success of the experimental manipulation, there was little evidence that increasing noise produces a consistent increase in errors. Among the 12 errors that were examined, the only detectable difference was the decrease in semantic coherence for conditional probability judgment. A proponent of the PTV model may inquire whether the manipulation was strong enough to produce the effect. One way to address this question is to evaluate the standardized effect size. According to general conventions, a standardized effect size of d = .95 is large (Cohen, 1988). On average, the absolute differences differed by .06 between the deadline and no deadline conditions, representing a 43% increase in noise. Therefore, it is difficult to argue that the increase in noise was trivial. Converging evidence across Experiments 1, 2 and 3, suggests that noise plays at least some role in error rates but perhaps not the dominant role stipulated by the PTV model.

Experiment 4

Experiment 4 investigated whether judgments can be improved through the principles of the PTV model and variations of the wisdom of crowds effect. The wisdom of crowds effect refers to the increased accuracy resulting from averaging judgments across multiple judges (e.g Surowiecki, 2004). Error can be decomposed into noise (random variability) and bias. The wisdom of crowds effect improves judgments by canceling or averaging out noise. Bias can be reduced when judges have opposing biases (Herzog & Hertwig, 2009). The wisdom of crowds effect has been used to improve correspondence across a variety of judgment tasks (e.g Herzog & Hertwig, 2009; Surowiecki, 2004).

Experiment 4 examined whether the wisdom of crowds effect can be simulated within the same judge using coherence as a metric of judgment quality instead of correspondence. Vul and Pashler (2008) conceptualized judgments as samples from an internal distribution. Although judgments will vary from time to time, gains in accuracy can be achieved by averaging multiple judgments from the same person. Supporting this notion, Herzog and Hertwig (2009) showed that the average of only two judgments can produce a robust improvement in accuracy. Averaging two judgments will improve accuracy as long as both judgments bracket (assume values above and below) the true judgment and the absolute deviation of the second judgment is no more than three times the absolute deviation of the initial judgment. As an example, assume the true judgment is .20 and the first judgment is .15, yielding an absolute deviation of .05. The second judgment can be as high as .35 before the average performs worse than the initial judgment. When judging the proportion of the world's airports that are located in the United States, the average judgment from each judge was more accurate than his or her individual judgments (Vul & Pashler, 2008). This reasoning is consistent with a basic tenet of the PTV model—namely, that true judgments are perturbed with noise and can be conceptualized as arising from an internal distribution.

Bias can be eliminated in the wisdom of crowds effect when different judges have opposing biases. On this basis, Herzog & Hertwig (2009) reasoned that the bias reducing properties of the wisdom of crowd's effect can be simulated within the same judge through a dialectical bootstrapping process. In dialectical bootstrapping, participants consider reasons their initial judgment could be wrong before providing a second, corrective judgment. This debiasing process is designed to increase the chance that the judgments bracket the true judgment. They found that dialectical bootstrapping improved accuracy compared to the average of non-dialectical judgments. However, the wisdom of crowds method (averaging across judges) still showed an advantage over the dialectical bootstrapping method. Müller-Trede (2011) replicated these results but distinguished between potential improvement and actual improvement. Potential improvement refers to improvement observed when the experimenter averages the judgments (as in Hertzog & Hertwig, 2009). By contrast, actual improvement refers to the improvement resulting from the strategies participant's use to resolve conflicts between their first and second judgments. Müller-Trede (2011) found that potential gains were greater than actual gains, a finding that suggests the majority of participants failed to capitalize on the error-reduction capability of averaging. In fact, many of the third judgments were outside of the bracket.

The purpose of Experiment 4 was to compare two interventions designed to reduce errors based on the principles of the PTV model and variants of the wisdom of crowds effect. In the dialectical bootstrapping intervention, an initial judgment was made followed by two corrective judgments. As argued in Hertzog and Hertwig (2008), the dialectical bootstrapping process increases the chance of bracketing, which, in turn, reduces error in judgments. One shortcoming of this method is that bracketing will not occur if, by sheer chance, the first two judgments are both above or below the true value. I developed what I have termed the Goldilocks intervention to overcome this shortcoming through systematic over and under-estimation of the first and second judgments. The logic of this intervention is based on the tale of Goldilocks and the Three Bears. One judgment was too high and the other judgment was too low. Biasing the initial judgments should increase the chance of bracketing and, by extension, the chance potential for improvement. To the extent that the biased judgments are perceived as extreme, participants may be prompted to provide a compromise judgment—a judgment that is "just right". This should lead to an actual improvement in addition to a potential improvement. Each of the interventions was compared to a repeated control condition in which judgments were simply repeated without any instruction. In this repeated judgment condition, errors were evaluated in each block separately and averaged to form a composite index of performance. The judgments in the repeated judgments condition were also subjected to an alternative analysis in which the judgments were averaged across blocks before being evaluated for errors. The purpose of averaging the judgments in the repeated judgment condition was to determine whether dialectical bootstrapping intervention and the Goldilock's intervention performed better than the wisdom of crowd's effect based on repeated judgments from individual participants. This analysis was termed 'averaged judgments'. The highest rate of errors was expected in the repeated judgment condition because Thus, the conditions can be rank ordered in terms of predicted efficacy: repeated judgments, averaged judgments, dialectical bootstrapping and Goldilocks.

Method

Participants

Participants were 69 introductory psychology students at Miami University, who completed the experiment for partial course credit.

Materials and Procedures

Due to the increased number of judgments, a subset of 16 target problems from Experiment 1 was selected for Experiment 4. One problem represented subsets, another problem represented mutually exclusive sets and the remaining problems represented overlapping sets. The problems were presented three times, once in each block. Within each block, the problems were presented individually in the same random order. Each problem featured a scenario followed by four judgments for $\mathrm{S(A)}, \mathrm{S(B)}, S(\mathrm{A} \cap \mathrm{B})$ and $\mathrm{S(A \cup B)}$.

Participants completed the experiment individually at computers in groups of one to five. In the repeated judgments condition, participants completed three blocks of judgments with no instructions aside from responding on a 0 to 100 scale. In the dialectical bootstrapping condition, participants provided their initial judgments in block 1 with no instruction. In block two, participants reviewed their initial judgments according to the instructions provided in Herzog & Hertwig (2009): "First, assume that your first estimate is off the mark. Second, think about a few reasons why that could be. Which assumptions and considerations could have been wrong? Third, what do these new considerations imply? Was the first estimate rather too high or too low? Fourth, based on this new perspective, make a second, alternative estimate below." In the third block, participants reviewed their first and second judgments and received the following instructions based on Müller-Trede (2011): "For the last time, we would like to present you some of the questions which you have answered during this experiment. On the basis of your previous responses, we would like to ask you for a third answer. Consider reasons why your first and second estimates may have been off the mark. Which assumptions may have been wrong for each estimate? Was the first judgment too high or too low? Did you over or under adjust your second estimate? Based on this new perspective, please make a third and final judgment." In the third block, participants were presented with their first two judgments and instructed to provide their best judgment based on the first two judgments. In the Goldilocks condition, participants systematically under or over-estimated their first and second judgment. In the first and second blocks, participants were instructed to "Consider a judgment that is too low (high) but still in the ballpark. Provide that judgment below." The ordering of low and high judgments was counter-balanced across participants. On the third and final block of judgments, participants reviewed their first and second judgments and were instructed to "Please consider your first and second judgments and make a third and final judgment, which you think is your best."

Results

Manipulation Check

Before proceeding with the primary analyses, the Goldilocks condition was inspected for order effects and evidence that the manipulation was successful. There was no evidence of order effects for the Goldilocks condition for any of the judgment errors (p's > .28). For this reason, the two counterbalanced groups were combined into a single Goldilocks condition for the remaining analyses. Two analyses were conducted to determine whether the under-and overestimation instructions were successful in the Goldilocks condition. For each participant, an average judgment was computed across problems for $S(A), S(B), S(A \cap B)$ and $S(A \cup B)$ in the underestimation and overestimation blocks. As shown below in Table 8, the manipulation was successful. Judgments were higher when participants were instructed to overestimate compared to when they were instructed to underestimate.

Table 8. Comparison of over-and underestimated judgments in Goldilock's condition.

Judgment	Overestimate Mean		Underestimate Mean		T-value	P-value
S(A)	.65	(.10)	.46	(.13)	10.24	<.001
S(B)	.47	(.11)	.33	(.11)	9.03	<.001
S(A ∩ B)	.50	(.12)	.34	(.11)	9.46	<.001
S(A ∪ B)	.57	(.12)	.41	(.13)	7.80	<.001

Standard deviations are in parentheses. Df = 23.

As a more stringent manipulation check, the judgments were compared to the judgments in the repeated judgments condition. The judgments were computed in the same manner for the repeated judgments condition, except the judgments were averaged across all three blocks for each participant. As shown in Table 9, the judgments were lower in the underestimation compared to the averaged judgments condition.

Table 9. Comparison of averaged judgments and underestimated judgments in the Goldilocks condition.

Judgment	Underestimation		Averaged Judgments		T-value	P-value
S(A)	.46	(.13)	.61	(.06)	4.77	<.001
S(B)	.33	(.11)	.40	(.07)	2.54	.02
S(A ∩ B)	.34	(.11)	.41	(.08)	2.33	.02
S(A ∪ B)	.41	(.13)	.52	(.07)	3.87	<.001

Standard deviations are in parentheses. Df = 44

A corresponding analysis was performed for the overestimation instructions. With two exceptions, Table 10 shows that judgments were higher in the overestimation condition compared to the averaged judgments conditions. Judgments for S(A) and S(A ∪ B) failed to reach statistical significance but were in the desired direction.

Table 10. Comparison of averaged judgments and overestimated judgments in the Goldilocks condition.

Judgment	Overestimation		Averaged Judgments		T-value	P-value
S(A)	.65	(.10)	.61	(.06)	-1.97	.06
S(B)	.47	(.11)	.40	(.07)	-2.43	.02
S(A ∩ B)	.50	(.12)	.41	(.08)	-2.98	<.005
S(A ∪ B)	.57	(.12)	.52	(.07)	-1.54	.13

Standard deviations are in parentheses. Df = 44.

The following analyses address whether more bracketing was observed in the dialectical bootstrapping and Goldilocks interventions. Since there is no objective value in which bracketing can be defined, bracketing was classified when the third judgment was between the first two judgments. For each participant, the proportion of bracketing judgments was computed as the number of bracketing judgments relative to all problems judgments (16 problems X 4 judgment types = 64 judgments per participant). As predicted, the Goldilocks intervention increased bracketing (M = .48; SD = .28) compared to making repeated judgments (M = .12; SD = .09), t(44) = 5.70, p <.001. There was also more bracketing in the Goldilocks condition compared to the dialectical bootstrapping intervention (M = .26; SD = .26), t(45) = 2.76, p = .01. As predicted, the most bracketing was observed in the Goldilocks condition. Presumably an intermediate judgment between the high and low extremes was seen as a compromise that was “just right”. Taken together, these analyses suggest the manipulation was mostly successful.

Primary Analyses

In the repeated judgments condition, an average error rate for each participant was computed across problems and blocks. For example, suppose a participant provided the following judgments in blocks 1,2, and 3 for a given problem. Block 1: $\mathrm{S(A)} = .30\,, S(\mathrm{B}) = .40, \mathrm{S(A \cap B)} = .35$; Block 2: $\mathrm{S(A)} = .35\,, \mathrm{S(B)} = .50, \mathrm{S(A \cap B)} = .25$; Block 3: $\mathrm{S(A)} = .40\,, S(\mathrm{B}) = .45, S(\mathrm{A \cap B}) = .30$. Therefore, the error rate for the conjunction fallacy is 1/3 for this particular problem. The preceding process was repeated across problems and averaged to form one error rate per participant. By contrast, error rates in the averaging analysis were computed on the judgments averaged across blocks. Building upon the previous example, the average judgments are $\overline{\mathrm{S(A)}} = .35\,, \overline{\mathrm{S(B)}} = .45$ and $\overline{\mathrm{S(A \cap B)}} = .30$. The error rate for the conjunction fallacy is 0 for the averaged judgments. This process was repeated across problems and averaged into one error rate per participant. In the dialectical bootstrapping and Goldilocks conditions, the error rates were computed on the final set of judgments in block 3 to measure actual improvements in performance. To measure the potential improvement in performance for the dialectical bootstrapping and Goldilocks conditions, the first and second set of judgments were averaged.

Planned contrasts were used to evaluate the actual and potential improvement of the interventions. The interventions were coded according to the following linear contrast to test the predicted rank order of effectiveness: averaged blocks (-3), averaged judgments (-1) averaged judgments, dialectical bootstrapping (1) and Goldilocks (3). Beginning with the analysis for actual improvement, Table 11 shows that the results are uniformly non-significant. Visual inspection of the trends reveals that the interventions were either ineffective or tended to increase errors.

Table 11. Mean error rates and linear contrasts for actual improvement observed in Experiment 4.

Condition	CF		DF		DCF		DDF		MCE		MDE		SC	
Repeated Judgments	.16	(.08)	.14	(.07)	.19	(.05)	.23	(.06)	.08	(.08)	.08	(.05)	.03	(.03)
Averaged Judgments	.27	(.12)	.25	(.13)	.20	(.08)	.26	(.08)	.08	(.08)	.08	(.06)	.01	(.02)
DB	.20	(.13)	.20	(.11)	.21	(.09)	.20	(.08)	.06	(.06)	.08	(.06)	.01	(.03)
GL	.21	(.11)	.22	(.10)	.20	(.09)	.25	(.06)	.06	(.06)	.07	(.07)	.01	(.02)
Contrast F(1,87)	.67		3.01		.26		.03		2.42		.67		3.09	
P-value	.42		.09		.61		.87		.12		.41		.08	

CF = Conjunction Fallacy; DF = Disjunction Fallacy; DCF = Double Conjunction Fallacy; DDF = Double Disjunction Fallacy; MCE = Minimum Conjunction Error; MDE= Maximum Disjunction Error; SC = Semantic Coherence.

Corresponding analyses for the potential improvement can be found below in Table 12. With one exception, the results did not accord with the predicted rank ordering. The rank ordering for the minimum conjunction error was close to the predicted rank order. However, the results for the disjunction fallacy were close to the opposite rank order and the remaining comparisons failed to reach statistical significance.

Table 12. Mean error rates and linear contrasts for potential improvement observed in Experiment 4.

Condition	CF		DF		DCF		DDF		MCE		MDE		SC	
Repeated Judgments	.16	(.08)	.14	(.07)	.19	(.05)	.23	(.06)	.08	(.08)	.08	(.05)	.03	(.03)
Averaged Judgments	.27	(.12)	.25	(.13)	.20	(.08)	.26	(.08)	.08	(.08)	.08	(.06)	.01	(.02)
DB	.22	(.13)	.19	(.13)	.17	(.09)	.23	(.10)	.06	(.05)	.07	(.05)	.02	(.03)
GL	.22	(.12)	.23	(.11)	.21	(.09)	.25	(.08)	.05	(.06)	.06	(.06)	.01	(.03)
Contrast F(1,87)	1.38		3.78		.20		.41		5.96		1.47		.96	
P-value	.24		.06		.66		.52		.02		.23		.33	

F = Conjunction Fallacy; DF = Disjunction Fallacy; DCF = Double Conjunction Fallacy; DDF = Double isjunction Fallacy; MCE = Minimum Conjunction Error; MDE= Maximum Disjunction Error; SC = emantic Coherence.

Discussion

Experiment 4 compared three interventions designed to reduce errors in probability judgment. Each intervention was based on the assumption of the PTV model that errors are the product of noise in judgments and employed various strategies to improve judgments through averaging. Judgments in the averaged judgments intervention were simply averaged in attempt to provide a better approximation of the true judgments. A slightly different approach was employed in the dialectical bootstrapping intervention. After making an initial judgment, participants in the dialectical bootstrapping intervention considered various reasons their initial judgments may have been wrong and provided a corrective judgment. This dialectical process was designed to increase bracketing of the true value so that the average judgment would better approximate the true value. Along similar lines, the Goldilocks intervention attempted to increase bracketing through systematic under-and overestimation of the initial judgments. According to the Goldilocks intervention, the over- and underestimated judgments would be perceived as extreme and thus prompt participants to make a compromise judgment that was “just right.” Contrary to predictions, the interventions were largely ineffective in reducing errors. This was true even for potential improvement in which the experimenter averaged the first two judgments in the dialectical bootstrapping and Goldilocks interventions. The one exception was the observed reduction for minimum conjunction errors. However, there is no theoretical reason that a reduction should be observed for the minimum conjunction but not the other errors.

One would expect averaging to improve judgments if the underlying process was consistent with the rules of probability theory. However, there was little evidence that this was the case and some, albeit weak, evidence that averaging may have increased errors in some cases. As discussed in further detail in the General Discussion, the interventions are predicated on assumptions about the functional form of the judgment distributions that may not hold. One possibility is that the true judgment might be better characterized as the median rather than the mean of a distribution, in which case judgments will regress towards .50 when averaged. Thus, if this were the case, averaging would introduce some degree of bias. Future studies may consider investigating whether the median can improve coherence more than the mean.

In summary, the PTV model assumes that errors in judgment are due to noise in a cognitive system that otherwise follows probability theory. The interventions tested in Experiment 4 were designed to provide more accurate estimates of the true judgments in order to improve coherence. Contrary to the model, the interventions were overwhelmingly ineffective. A proponent of the PTV model may argue that an inappropriate functional form of the model was assumed. Bias could have been introduced through regression when the judgments are averaged. A competing explanation is that the more fundamental assumptions of the PTV model are wrong—namely, the assumption that judgments follow probability theory. Given the other shortcomings of the model observed in Experiments 1, 2 and 3, the latter explanation appears to be more plausible.

General Discussion

Over the course of four decades, researchers have amassed an abundance of evidence indicating that human judgment departures from probability theory. For this reason, the notion that people do not judge probabilities according to the rules of probability theory has become a truism. Most theories account for systematic errors with assumptions that are inherently non-normative. Although this is a natural starting point, a comprehensive account of probability judgment from this approach has yet to be realized. As an alternative approach, the PTV model begins with the assumption that systematic errors emerge from the perturbation of noise in a cognitive system that otherwise operates according to probability theory. Although the noise is random, it exerts a systematic effect that is capable of generating many of the errors documented in the literature. The PTV model has several attractive features that made it a worthy candidate for a comprehensive theory. First, the model can account for several key findings in the literature regarding the conjunction and disjunction fallacies, subadditivity and stochastic adherence to the addition law. Second, as shown in the present article, the PTV model has been the source of several novel predictions regarding the relationship between noise and errors. Third, the PTV model has the potential to be further developed into a cognitive process model.

Despite those attractive features of the PTV model, several shortcomings were revealed across all four experiments. There was mixed support for the PTV model's critical property that noise produces errors in judgment. Although the PTV model provided a better account the relationship between noise and errors compared to the CWA model in Experiment 1, in absolute terms the PTV model was not consistently supported. Support was found for the PTV model in only three of the seven errors in Experiment 1, two of the five errors in Experiment 2 and one of the 12 errors in Experiment 3. Experiments 1 and 3 provided the opportunity to demonstrate the causal role of noise in producing errors. In Experiment 1, the order in which judgments were made was manipulated to increase noise. The manipulation was successful in increasing noise in conjunctive probabilities, but a corresponding increase in errors was only observed for the conjunction and disjunction fallacies. The remaining five errors did not show a statistically significant increase in errors due to order effects. In Experiment 3, a response deadline was instated to experimentally increase the noise in judgments. However, there was little evidence that experimentally increasing noise leads to an increase in errors. This pattern of results suggests that noise is implicated in judgment errors, but not in the consistent manner that would be expected on the basis of the PTV model.

A common theme that emerged in Experiments 1 and 2 was the failure of the PTV model to provide a quantitative account of semantic coherence in certain problem types. For example, the PTV model has difficulty accounting for semantic coherence in identical sets because it requires a high degree of precision to be achieved. Empirically, semantic coherence is high for identical sets but is rare in judgment space (Wolfe, Fisher & Reyna, 2013). For this reason, semantic coherence is unlikely to be produced from a random process.

The PTV model also failed an important qualitative test derived from the variance sum law. The integration rules of the PTV model imply that disjunctions should have more noise compared to conjunctions. Contrary to this prediction, the variances were nearly identical. Another finding that was at odds with the PTV model was the higher rate of conversion errors relative to conditional reversals. Under most conditions, one would expect noise to produce more cases in which the normative rank ordering of the conditional probabilities is reversed than cases in which they are judged to be equal.

According to the PTV model, judgments might be improved by averaging multiple judgments from the same person. Three interventions were developed to improve judgment using various judgment and averaging procedures. However, the interventions were largely unsuccessful in improving coherence. One explanation is that the failure of averaging was due to a wrong assumption about the functional form of the PTV model. Regression may have introduced bias when the judgments were averaged. Another explanation is that the judgments were simply not generated from the rules of probability theory. The distinction between these two explanations is important. The first explanation suggests that coherence can be improved with a different averaging method, such as using the median of multiple judgments. By contrast, the other explanation suggests that judgments can be improved though some other process, such as making the hierarchical set relationships more transparent because people are not following probability theory. Previous studies have found that drawing attention to the set representation of the problems was effective in improving coherence (Wolfe & Reyna, 2010; Wolfe, Fisher & Reyna, 2013).

One question that warrants attention is how should the evidence that is consistent with the model be evaluated? It is often the case that results that are inconsistent with a model are more diagnostic than results that are consistent with a model. A common problem in model testing is that results that are consistent with a model under investigation may be consistent with other models that were not considered (Lewandowsky & Farrell, 2010). Stochastic adherence to the addition law is one case in point. In the absence of competing models, this result may be misinterpreted as very strong support for the PTV model because the addition law is derived the axioms of probability theory. It may seem unlikely that a non-normative model can account for the addition law. However, the CWA model was able to account for this superficially normative pattern of judgments even though it uses non-normative integration rules. Although I am unaware of a model that can account for the stochastic adherence to Bayes' theorem, it is possible that one may exist or be proposed in the future.

In summary, some aspects of the model were inconsistently supported, whereas other aspects did not receive any support. When considered in their entirety, the results cast considerable doubt on the PTV model. Before drawing strong conclusions about the PTV model, it is important to consider the assumptions on which the model was based and the limitations inherent in the experiments. Each of these topics is discussed in turn.

Assumptions of the Model

This section explores the ability of two alternative formulations of the PTV to account for the discrepant findings. One alternative formulation of the model involves relaxing the assumption that $E[S(A)] = P(A)$. Relaxing this assumption would change several of the fundamental predictions of the model without altering the core assumptions that judgments adhere to probability theory and errors are produced by noise. For example, Costello and Watts (2013) recently proposed a variant of the PTV model with a very simple recall mechanism. In this model, events are coded as 1's and 0's in a manner similar to that of the dynamic model introduced in Experiment 3. Formally, let A be a vector that encodes the occurrence of event A as 1 and the non-occurrence of event A as 0.

$$A = [a_1, a_2, \ldots a_n] \tag{29}$$

Let r be the probability that an event is correctly recalled and (1-r) be the probability it is incorrectly recalled. An event could be falsely recalled because $a_i = 1$ is misread as $a_i = 0$ or $a_i = 0$ is misread as $a_i = 1$. The expected value is then:

$$E[S(A)] = P(A)r + \big(1 - P(A)\big)(1 - r) \neq P(A) \tag{30}$$

for r < 1. And

$$P(A) = \frac{1}{N}\sum_{i=1}^{N} a_i \tag{31}$$

As in the stochastic calibration model (Erev, et al., 1994), subjective probabilities produced from this simple recall process regress towards .50. One implication of regression is that the expectation and variance would depend on the rules used to compute the probabilities because additivity and multiplicativity would no longer apply, except when P(A) = .50. For example, $E[S(A \cap B)] \neq E[S(A)]E[S(B|A)]$. Judgments will be subadditive and submultiplicative when P(A) < .50 and superadditive and supermultiplicative when P(A) > .50. Depending on the rules used to compute the probabilities, this variant of the PTV model could potentially account for the equality of variance in conjunctions and disjunctions. A qualitative prediction for the variance sum law cannot be derived if the judgments are computed from $S(A \cap B)$ and $S(A \cup B)$. By contrast, the variance sum law will hold if the conjunctions and disjunctions are computed indirectly from the component and conditional probabilities. Regression might also account for the failure of the interventions to improve coherence. Regression will introduce some degree of bias when multiple judgments are averaged. Consider the Goldilocks condition for illustration. Assume the true judgment, such as .10, is near the lower boundary of zero. Under this condition, judgments can be overestimated by a larger margin (e.g. .25) than they can be underestimated (e.g. 0). As a result, the true value, on average, will tend to be overestimated. This formulation of the model suggests that the median would be better suited for improving coherence.

Although this variant of the PTV model could accommodate some of the discrepant findings, it is unduly flexible and unprincipled. Unless additional constraints are imposed for the computation of probabilities, the model may provide a post hoc account of virtually any result at the expense of prediction and falsifiability. One could arbitrarily posit different rules for computing probabilities to accommodate a discrepant result. The model could be constrained through a principle based on computational simplicity. Such a principle would require probabilities to be computed with the simplest formula (i.e. the formula with fewest terms). For example, event A would be computed as $S(\mathrm{A})$ as opposed to $S(\mathrm{A} \cap \mathrm{B}) + S(\mathrm{A} \cap \sim\mathrm{B})$. Adopting the principle of computational simplicity would circumvent problems with flexibility and allow novel predictions to be derived. Although this variant of the PTV model can accommodate the violation of the variance sum law and the failure of the interventions, it is incapable of rectifying several discrepant findings, including the inconsistent relationship between noise and errors and the higher rates of conversion errors compared to conditional reversals. For these reasons, positing a different functional form would only constitute a partial solution.

Alternatively, the PTV model could account for some of the discrepant findings by relaxing the assumption of independence. The prediction that disjunctions have more noise than conjunctions was predicated on the assumption of independence in the variance sum law. When independence does not hold, covariance terms are added to Equation 22. A sufficiently positive correlation between $\mathrm{S(A \cap B)}$ and $\mathrm{S(B)}$ or $\mathrm{S(A \cap B)}$ and $\mathrm{S(A)}$ might produce equal variance in conjunctions and disjunctions. Relaxing independence may also allow the model to account for semantic coherence in identical sets. Semantic coherence for identical sets requires $\mathrm{P(A)} = \mathrm{P(B)} = \mathrm{P(A \cap B)} = \mathrm{P(A \cup B)}$. If the judgments are highly correlated, they could vary from time to time while allowing the rule to be satisfied. A similar argument could be proposed for the conversion error. If S(A|B) becomes correlated with S(B|A), the rate of conversion errors may increase. However, relaxing the assumption of independence is ill advised for several reasons. One reason is that it makes the model exceedingly complex, mathematically intractable and consequentially difficult to falsify. A second reason is that there appears to be no theoretical justification for relaxing the assumption of independence. Without sound justification, the non-independence could be evoked arbitrarily to account for virtually any discrepant findings.

A critic could argue that independence holds, but artifactual correlations were introduced through the experimental tasks. For example, participants may have remembered judgments between blocks or within problems. However, several methodological safeguards were implemented to minimize the impact of artifactual correlations. A filler task was included to interfere with memory between judgment blocks. Additionally, judgments within problems were randomized and presented individually to further burden memory. For these reasons, artifactual correlations are not the most plausible explanation for the shortcomings of the PTV model.

Limitations

The limitations inherent in any set of experiments should be considered when interpreting the results. In this section, I discuss statistical and methodological limitations and propose solutions to some of the limitations in the subsequent section, Future Directions. In several cases, the relationship between noise and errors was not statistically significant. The lack of statistical significance complicates the interpretation of the results. A critic could argue that the null findings are due to lack of statistical power. In Experiment 1, for example, the power to detect a medium effect size of d = .50 and r = .30 is .43 and .61, respectively (two-tailed). Although this is a valid criticism for any given experiment, converging evidence across the experiments suggests that relationship between noise and errors is not as consistent as predicted by the PTV model. Therefore, lack of statistical power does not appear to be a tenable explanation for the statistically non-significant results when considered across all of the experiments.

Several caveats must be considered when interpreting the quantitative predictions of any model. Although quantitative model fits can provide refined insights into a model, the results may depend on auxiliary assumptions that are not necessarily central to the core assumptions of the model (i.e. functional form) or data aggregation methods that are chosen for practical reasons. Following the procedures detailed in Costello (2009a), the quantitative tests of the PTV model in Experiment 1 and 2 were performed on group level data. One limitation is that group level data may not be representative of some of the individual level data. Unfortunately, fitting the model at the individual level was not feasible because tractable analytical solutions are not currently available and may not exist. As an alternative approach, the quantitative predictions were approximated using computationally extensive simulation based methods that preclude an individual-level analysis.

A proponent of the PTV model may argue that the poor fit of the model was due to inappropriate data aggregation and should not count as evidence against the model. A counterargument could be made on the basis of the high degree of quantitative fit found in Costello (2009a) using group level data. This finding suggests an alternative explanation: the inclusion of semantic coherence and other errors may have simply created a more stringent test that the model failed to pass. Given the higher bar set by semantic coherence, one could further argue it is unlikely that noise could produce the high semantic coherence rates, even at the individual level. Ultimately, this is an issue that can only be resolved once analytical solutions become available or substantial increases in computing power are realized. In the meantime, some degree of caution should be exercised in interpreting the results of the quantitative tests.

Several interrelated limitations stem from the scenario-based problems commonly used to study probability judgment. These limitations make it difficult to quantify the noise in judgments. For example, in scenario-based problems, participants read a scenario before making probability judgments. Because the scenarios are salient, participants may realize the purpose of the experiment while providing judgments in the second block. As a result, participants may employ strategies to remember their judgments in subsequent replications. In addition, scenario-based problems are relatively time consuming to administer, which places further practical constraints on the number of replicate judgments that can be made. Another limitation of scenario-based problems is that there is no objective value against which correspondence can be evaluated. As a result, scenario based problems are only amenable to the evaluation of coherence. One could argue from a practical standpoint that improving correspondence is more important than improving coherence. A set of judgments could achieve coherence but depart drastically from their true values. For this reason, the ability of the interventions in Experiment 4 to improve correspondence may have been missed.

Researchers have been interested in errors such as the conjunction fallacy because they demonstrate simple but compelling violations of probability theory. For example, Tversky and Kahneman (1983) referred to the conjunction rule as the most basic rule of probability theory. Despite this appeal, errors are very crude measures of judgment performance and consequentially have poor statistical properties. It is well established that dichotomizing continuous variables results in a considerable loss of information and statistical power (Dawson, & Weiss, 2012; MacCallum, Zhang, Preacher & Rucker, 2002). Consequentially, there is some possibility that the inconsistent relationship between noise and errors may have been due to the crude means with which judgments are commonly evaluated. Evaluating the entire distribution of judgments may provide unique opportunities to investigate the integration rules and cognitive processes through which the judgments are formed. For example, when conjunctive probabilities are formed through multiplicative integration of component probabilities, the resulting subjective probability distribution tends to be skewed. By contrast, additive integration of component probabilities produces a subjective probability distribution that is less skewed with greater variance.

Future Directions

Notwithstanding the shortcomings of the PTV model, perhaps its most important contribution was emphasizing the variance of judgments. A common practice in psychology is to simply dismiss the variance as unimportant noise and focus instead on measures of central tendency, particularly the mean. In the PTV model, the variance is an integral component of judgment from which several novel predictions were derived, including the variance sum law and tests of the addition law and Bayes' theorem. Tremendous progress in perceptual (Ratcliff & Smith, 2004) and preferential decision making (Roe et al., 2001) and categorization (Nosofsky & Palmeri, 1997) has resulted from simultaneously modeling multiple aspects of cognitive output, such as choice and reaction time. One benefit of simultaneously modeling choice and reaction time is that it imposes constraints on the model. As a consequence, more confidence can be placed on a model if it is able to account for both cognitive outputs as opposed to only one. Along similar lines, considering the variance in judgment may help constrain and refine models and provide insights into the cognitive processing that underlie probability judgments. If a model is a good approximation of the true generating process, it should be able to capture general trends in the variance as well as the central tendency. Richer and even more detailed information could be extracted by considering subjective probability distributions in their entirety.

New methods for studying probability judgment are required in order to obtain richer information about the subjective probability distributions. I plan to develop an alternative judgment task to circumvent the shortcomings associated with scenario-based problems. The new task would replace scenarios with perceptual stimuli. The perceptual stimuli are 10X10 matrices with cells that feature the following stimulus dimensions: color (white vs. grey) and symbol (dot vs. no dot). Importantly, the dimensions can be combined to create stimuli with component, joint and conditional events. For example, a conjunctive event could be represented as a grey cell that has a dot. Using perceptual stimuli has several advantages over scenario-based problems. One general advantage is the increase in experimental control and flexibility. The experimenter can control the appearance of the stimuli, presentation time, the proportion of events and the absolute number of events. Another advantage is that the stimuli can be presented multiple times because they are less memorable. Alternatively, slight variations in the event coordinates can be introduced for given stimulus type (e.g. 30 dotted cells, 10 grey cells squares, 3 grey and dotted cells) to further increase the number of replicate judgments. Collecting multiple judgments will enable researchers to assess subjective probability distributions and become less reliant on crude measures of judgment performance, such as the conjunction fallacy. Subjective probability distributions may provide richer data with which various integration rules and cognitive processes can be compared.

Another benefit of using perceptual stimuli is that correspondence and coherence could be evaluated simultaneously in the same task. A major limitation of Experiment 4 was that interventions could only be evaluated in terms of coherence. As a result, it is unknown whether the interventions were successful in improving correspondence. Another benefit of using perceptual stimuli is that the increased number of replicate judgments would make it possible to assess the functional form of the PTV model. The medians could be estimated to determine whether the failure of the averaging interventions was due to regression.

Are People Naïve Probability Theorists?

The notion that people are rational is deeply embedded in Western culture. More than 2,000 years ago, Aristotle characterized humans as rational animals (Se Code, 2003). Rationality remains a core assumption in current economic models, such as Expected Utility Theory. One assumption of normative theories of decision making is that people reason about uncertainty according to probability theory. The PTV model posits that people basically follow the rules of probability theory but noise produces the errors in judgments. The notion that errors in judgment are simply due to noise adds an interesting and complex dimension to the debate about rationality. One might ponder whether people can be considered rational if the underlying cognitive system is noisy but consistent with probability theory. In an unlikely reality in which judgments are free of noise, people may adhere to the rules of probability theory. Under most realistic conditions, some degrees of error are likely. Another question is whether rational agents must be able to articulate the rules of probability theory. The PTV model assumes that the judgments are not made explicitly. According to the model, people are not aware of the rules of probability theory that are generating their judgments. Are people rational if their behavior is not purposeful? Nonetheless, the results of the four Experiments suggest that philosophers need not preoccupy themselves with these questions so quickly. Even if different auxiliary assumptions are adopted to increase the explanatory scope of the PTV model, the model's account of noise and errors was not sufficiently supported. The available evidence suggests that people are not naïve probability theorists.

References

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

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. Behavioral and Brain Sciences, 30(3), 241-254.

Bonini, N., Tentori, K., & Osherson, D. (2004). A different conjunction fallacy. Mind & Language, 19(2), 199-210.

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment. Part I: New theoretical developments. Journal of Behavioral Decision Making, 10(3), 157-171.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Psychology Press.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5(4), 559-583.

Costello, F. J. (2009a). How probability theory explains the conjunction fallacy. Journal of Behavioral Decision Making, 22(3), 213-234.

Costello, F. J. (2009b). Fallacies in probability judgments for conjunctions and disjunctions of everyday events. Journal of Behavioral Decision Making, 22(3), 235-251.

Costello, F., & Watts, P. (2013). Surprisingly Rational: Evidence that people follow probability theory when judging probabilities, and that biases in judgment are due to noise. arXiv preprint arXiv:1211.0501.

Crisp, A. K., & Feeney, A. (2009). Causal conjunction fallacies: The roles of causal strength and mental resources. The Quarterly Journal of Experimental Psychology, 62(12), 2320-2337.

Dawson, N. V., & Weiss, R. (2012). Dichotomizing Continuous Variables in Statistical Analysis A Practice to Avoid. *Medical Decision Making*, *32*(2), 225-226.

Dougherty, M. R., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. Psychological Review, 106(1), 180-209.

Dougherty, M. R., & Hunter, J. E. (2003). Hypothesis generation, probability judgment, and individual differences in working memory capacity. *Acta psychologica*, *113*(3), 263-282.

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over-and under-confidence: The role of error in judgment processes. Psychological Review, 101(3), 519.

Fantino, E., Kulik, J., Stolarz-Fantino, S., & Wright, W. (1997). The conjunction fallacy: A test of averaging hypotheses. Psychonomic Bulletin & Review, 4(1), 96-101.

Fisher, C. R., & Wolfe, C. R. (2011). Assessing semantic coherence in conditional probability estimates. Behavior research methods, 43(4), 999-1002.

Fisk, J. E., & Pidgeon, N. (1996). Component probabilities and the conjunction fallacy: Resolving signed summation and the low component model in a contingent approach. Acta psychologica, 94(1), 1-20.

Gavanski, I., & Roskos-Ewoldsen, D. R. (1991). Representativeness and conjoint probability. Journal of Personality and Social Psychology, 61(2), 181-194.

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. Psychological Review, 103, 592–596.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. Psychological review, 102(4), 684-704.

Hammond, K. R. (2000). Coherence and correspondence theories in judgment and decision making. In T. Connolly, H. R. Arkes, & K. R. Hammond (Eds.), Judgment and decision making: An interdisciplinary reader (2nd ed., pp. 53-65). Cambridge: Cambridge University Press

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. Journal of Behavioral Decision Making, 12, 275–305.

Herzog, S. M., & Hertwig, R. (2009). The Wisdom of Many in One Mind Improving Individual Judgments With Dialectical Bootstrapping. Psychological Science, 20(2), 231-237.

Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. Behavior Research Methods, 16(2), 96-101.

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. Cognition, 84(3), 343-352.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. Psychological review, 80(4), 237-251.

Kahneman, D., & Tversky, A. (1979). Prospect theory: an analysis of decision under risk. Econometrica,47, 263–291.

Lewandowsky, S., & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. Sage.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*,*7*(1), 19.

Marr, D., & Vision, A. (1982). A computational investigation into the human representation and processing of visual information. WH San Francisco: Freeman and Company.

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. Judgment and Decision Making, 6(4), 283-294.

Neil Bearden, J., Wallsten, T. S., & Fox, C. R. (2007). Contrasting stochastic and support theory accounts of subadditivity. Journal of Mathematical Psychology, 51(4), 229-241.

Nilsson, H., Winman, A., Juslin, P., & Hansson, G. (2009). Linda is not a bearded lady: Configural weighting and adding as the cause of extension errors. Journal of Experimental Psychology: General, 138(4), 517.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. Psychological Review; Psychological Review, 104(2), 266.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological review*, *111*(2), 333.

Reyna, V.F., & Brainerd, C.J. (1995). Fuzzy-trace theory: An interim synthesis. Learning and Individual Differences, 7, 1–75.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making.*Psychological review*, *108*(2), 370.

Se Code, A. (2003). Aristotle's logic and metaphysics. *Routledge History of Philosophy, 2*, 40-75.

Stolarz-Fantino, S., Fantino, E., & Kulik, J. (1996). The conjunction fallacy: Differential incidence as a function of descriptive frames and educational context. Contemporary Educational Psychology, 21(2), 208-218.

Stolarz-Fantino, S., Fantino, E., Zizzo, D. J., & Wen, J. (2003). The conjunction effect: New evidence for robustness. American Journal of Psychology, 116(1).

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wis- dom shapes business, economies, socities, and nations. Random House of Canada.

Tentori, K., Bonini, N., & Osherson, D. (2004). The conjunction fallacy: a misunderstanding about conjunction?. Cognitive Science, 28(3), 467-477.

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.

Vul, E., & Pashler, H. (2008). Measuring the crowd within probabilistic representations within individuals. *Psychological Science*, *19*(7), 645-647.

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. Psychological Review, 101, 490–504.

Wedell, D. H., & Moro, R. (2008). Testing boundary conditions for the conjunction fallacy: Effects of response mode, conceptual focus, and problem type. Cognition, 107(1), 105-136.

Wolfe, C. R. (1995). Information seeking on Bayesian conditional probability problems: A fuzzy- trace theory account. Journal of Behavioral Decision Making, 8(2), 85-108.

Wolfe, C. R., & Britt, M. A. (2008). The locus of the myside bias in written argumentation. Thinking & Reasoning, 14(1), 1-27.

Wolfe, C. R., & Reyna, V. F. (2010). Semantic coherence and fallacies in estimating joint probabilities. Journal of Behavioral Decision Making, 23(2), 203-223.

Wolfe, C. R., & Reyna, V. F. (2010). Assessing semantic coherence and logical fallacies in joint probability estimates. Behavior research methods, 42(2), 373-380.

Wolfe, C.R., Fisher, C.R. & Reyna, V.F. (2012). Semantic Coherence and Inconsistency in Estimating Conditional Probabilities. Journal of Behavioral Decision Making: DOI: 10.1002/bdm.1756

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets?. Experimental Psychology (formerly Zeitschrift für Experimentelle Psychologie), 50(2), 97-106.

Appendix 1

Model Fitting Details

In this section, I describe the model fitting procedure used to compute the predicted error and semantic coherence rates for joint and conditional probability judgment. As described in Costello (2009a), subjective probabilities are derived from a normal distribution that represents subjective confidence. Subjective probability distributions are produced via a response rule that maps the confidence distribution on the interval [0,1] through a logistic function. To make the model more psychologically plausible, we assume subjective probabilities are rounded to the nearest multiple of five, as commonly found in practice (see Erev, Wallsten & Budescu 1994). Quantitative model predictions were derived by sampling from the subjective probability distributions for each of the 34 problems using group level data. Let $\overline{S(k)_{p,obs}}$ and $\sigma_{S(k)_{p,obs}}$ be the observed mean and standard deviation of event k in problem p, formed by averaging judgments across participants at time 1 and time 2. To find the subjective probability distribution for event k in problem p, N = 100,000 simulated subjective confidence judgments were sampled from a normal distribution $C(k)_{pi} \sim N(\mu_{C(K)_p}, \sigma_{C(K)_p})$. Next, the simulated subjective confidence judgments were transformed to subjective probabilities using the following logistic function: $\widehat{S(k)}_{pi} = \frac{e^{C(k)_{pi}}}{1+e^{C(k)_{pi}}}$.
Parameters $\mu_{C(K)_p}$ and $\sigma_{C(K)_p}$ of the subjective confidence distribution where adjusted using the Nelder-Mead algorithm until the mean and standard deviation of the resulting subjective probability distribution, $\overline{\widehat{S(k)_p}}$and $\widehat{\sigma_{S(k)_p}}$, approximated the observed mean and standard deviation as closely as possible.

Quantitative predictions for the error and semantic coherence rates were estimated by sampling N = 100,000 times from the relevant estimated subjective probability distributions, $\widehat{S(k)}_p$, and computing the relative frequency with which the errors and semantic coherence occurred. The predicted conjunction and disjunction fallacy rates were computed as :$F_{pi} = S\widehat{(B|A)}_{pi}\widehat{S(A)}_{pi} - \widehat{S(B)}_{pi}$ and $P\widehat{(CF)}_p = P\widehat{(DF)}_p = \frac{\sum_{r=1}^{N}\begin{cases} F_{pi} > 0 \to 1 \\ \text{Else } 0 \end{cases}}{N}$, where CF and DF denote the conjunction and disjunction fallacy, respectively. Similarly, the double conjunction and disjunction fallacies were computed as: $G_{pi} = S\widehat{(A|B)}_{pi}\widehat{S(B)}_{pi} - \widehat{S(A)}_{pi}$ and $P\widehat{(DCF)}_p = P\widehat{(DDF)}_p = \frac{\sum_{r=1}^{N}\begin{cases} G_{pi} > 0 \to 1 \\ \text{Else } 0 \end{cases}}{N}$. The predicted minimum conjunction error was computed as: $H_{pi} = -1 + \widehat{S(A)}_{pi} + \widehat{S(B)}_{pi} - S\widehat{(A|B)}_{pi}\,\widehat{S(B)}_{pj}$ and $P\widehat{(MCE)}_p = \frac{\sum_{r=1}^{N}\begin{cases} H_{pi} > 0 \to 1 \\ \text{Else } 0 \end{cases}}{N}$, where $i \neq j$. The predicted maximum disjunction error was computed as $H_{pi} = \left[\widehat{S(A)}_{pi} + \widehat{S(B)}_{pi} - S\widehat{(B|A)}_{pi}\,\widehat{S(A)}_{pi}\right] - \left[\widehat{S(A)}_{pj} + \widehat{S(B)}_{pj}\right]$ and $P\widehat{(MCE)}_p = \frac{\sum_{r=1}^{N}\begin{cases} H_{pi} > 0 \to 1 \\ \text{Else } 0 \end{cases}}{N}$. Semantic coherence was also estimated by computing the relative frequency with which it occurred in the simulated judgments (for details see Wolfe & Renya, 2010; Fisher & Wolfe, 2011). The model was fit by minimizing the sum of the squared differences between the predicted and observed error and semantic coherence rates for each problem p. $\widehat{\mathrm{S(k)_p}}$and $\widehat{\sigma_{\mathrm{S(k)_p}}}$ were free to vary within $\pm$ .02 of their observed values. This allowed the model to adjust for sampling error while being sufficiently constrained. One challenge in estimating the error and semantic coherence rates through simulation is that it destabilizes the parameter space; a slightly different rate will be obtained each time with the same parameters. Two measures were undertaken to minimize this problem. First, the predicted error rates were rounded to two decimal places. Second, a large number of simulated judgments (N = 100,000) was used to further increase the stability of the estimates.

The model fitting procedure for conditional probability judgment in Experiment 2 was identical, except the rates were computed for the minimum conditional errors, the conditional reversal, the conversion error and semantic coherence for conditional probability judgment. The predicted rate for the minimum conditional for A|B was, $\mathrm{L_{pi}} = \frac{\widehat{\mathrm{S(A)}}_{\mathrm{p}\imath} + \widehat{\mathrm{S(B)}}_{\mathrm{p}\imath} - 1}{\widehat{\mathrm{S(B)}}_{\mathrm{p}\jmath}} - \mathrm{S}\widehat{\mathrm{(A|B)}}_{\mathrm{p}\imath}$ and $\mathrm{P(M\widehat{CE\ A}|B)_p} = \frac{\sum_{\mathrm{r=1}}^{\mathrm{N}} \begin{cases} \mathrm{L_{pi}} > 0 \rightarrow 1 \\ \quad \mathrm{Else\ 0} \end{cases}}{\mathrm{N}}$. The minimum conditional error for B|A is computed similarly: $\mathrm{M_{pi}} = \frac{\widehat{\mathrm{S(A)}}_{\mathrm{p}\imath} + \widehat{\mathrm{S(B)}}_{\mathrm{p}\imath} - 1}{\widehat{\mathrm{S(A)}}_{\mathrm{p}\jmath}} - \mathrm{S}\widehat{\mathrm{(B|A)}}_{\mathrm{p}\imath}$ and $\mathrm{P(M\widehat{CE\ B}|A)_p} = \frac{\sum_{\mathrm{r=1}}^{\mathrm{N}} \begin{cases} \mathrm{M_{pi}} > 0 \rightarrow 1 \\ \quad \mathrm{Else\ 0} \end{cases}}{\mathrm{N}}$. The conversion error is calculated as: $\mathrm{O_{pi}} = 1$ if $\widehat{\mathrm{S(A)}}_{\mathrm{p}\imath} \neq \widehat{\mathrm{S(B)}}_{\mathrm{p}\imath}$ and $\mathrm{S}\widehat{\mathrm{(A|B)}}_{\mathrm{p}\imath} = \mathrm{S}\widehat{\mathrm{(B|A)}}_{\mathrm{p}\imath}$ and $\mathrm{S}\widehat{\mathrm{(A|B)}}_{\mathrm{p}\imath} = \mathrm{S(B|}\widehat{\mathrm{A)}}_{\mathrm{p}\imath} \neq 0$ are true and $\mathrm{O_{pi}} = 0$ otherwise. Thus, the probability of a conversion error is $\mathrm{P}\widehat{\mathrm{(CE\ )}}_{\mathrm{p}} = \frac{\sum_{\mathrm{r=1}}^{\mathrm{N}} \mathrm{O_{pi}}}{\mathrm{N}}$. A conditional reversial is coded as $\mathrm{Q_{pi}} = 1$ if both $\widehat{\mathrm{S(A)}}_{\mathrm{p}\imath} > \widehat{\mathrm{S(B)}}_{\mathrm{p}\imath}$ and $\mathrm{S}\widehat{\mathrm{(A|B)}}_{\mathrm{p}\imath} < \mathrm{S}\widehat{\mathrm{(B|A)}}_{\mathrm{p}\imath}$, $\mathrm{Q_{pi}} = 0$ otherwise. Thus, the probability of a conditional reversal is $\mathrm{P}\widehat{\mathrm{(CR)}}_{\jmath} = \frac{\sum_{\mathrm{r=1}}^{\mathrm{N}} \mathrm{Q_{pi}}}{\mathrm{N}}$