AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Review

© 2021 American Psychological Association
ISSN: 0033-295X

2021, Vol. 128, No. 6, 1088–1111
https://doi.org/10.1037/rev0000299

The Similarity-Updating Model of Probability Judgment and Belief Revision

Rebecca Albrecht[1], Mirjam A. Jenny[1, 2, 3], Håkan Nilsson[4], and Jörg Rieskamp[1]

[1] Department of Psychology, University of Basel
[2] Science Communication, Robert Koch Institute, Berlin, Germany
[3] Harding Center for Risk Literacy, University of Potsdam
[4] Department of Psychology, Uppsala University

People often take nondiagnostic information into account when revising their beliefs. A probability judgment decreases due to nondiagnostic information represents the well-established "dilution effect" observed in many domains. Surprisingly, the opposite of the dilution effect called the "confirmation effect" has also been observed frequently. The present work provides a unified cognitive model that allows both effects to be explained simultaneously. The suggested similarity-updating model incorporates two psychological components: first, a similarity-based judgment inspired by categorization research, and second, a weighting-and-adding process with an adjustment following a similarity-based confirmation mechanism. Four experimental studies demonstrate the model's predictive accuracy for probability judgments and belief revision. The participants received a sample of information from one of two options and had to judge from which option the information came. The similarity-updating model predicts that the probability judgment is a function of the similarity of the sample to the options. When one is presented with a new sample, the previous probability judgment is updated with a second probability judgment by taking a weighted average of the two and adjusting the result according to a similarity-based confirmation. The model describes people's probability judgments well and outcompetes a Bayesian cognitive model and an alternative probability-theory-plus-noise model. The similarity-updating model accounts for several qualitative findings, namely, dilution effects, confirmation effects, order effects, and the finding that probability judgments are invariant to sample size. In sum, the similarity-updating model provides a plausible account of human probability judgment and belief revision.

Keywords: probability judgment, belief updating, similarity, dilution effect

Supplemental materials: https://doi.org/10.1037/rev0000299.supp

Judging probabilities correctly and making good decisions under uncertainty are crucial skills and can affect any area of people's lives, be it finance, education, or health care. What is the probability that my newly acquired stocks will rise within the next month? And what is the probability that my mother will recover from her hip injury within the next year? Such probability judgments are usually based on several sequentially acquired pieces of information. Accordingly, initial probability judgments are made and then dynamically revised on the basis of new incoming information. We suggest the similarity-updating model, which explains the cognitive process of how people make and revise probability judgments and why people deviate from probability theory. Specifically, the model clarifies how people deal with nondiagnostic information and explains contradictory findings on how people use nondiagnostic information.

Characteristics of Human Probability Judgments and Belief Revision

Past research has shown that people's probability judgments are often inconsistent with normative probability theory. For instance, well-described phenomena such as conservatism (Dougherty et al., 1999; Edwards, 1968), base-rate neglect (e.g., Bar-Hillel, 1980), sub- and superadditivity (Dougherty & Hunter, 2003; Macchi et al., 1999; Tversky & Koehler, 1994), the conjunction fallacy (e.g., Tversky & Kahneman, 1983), and order effects (Hogarth & Einhorn, 1992) represent violations of probability theory. Various models have been proposed to describe these effects and the cognitive processes underlying probability judgments, including mathematical models following Bayesian principles (Chater et al., 2006; Griffiths et al., 2010; Sanborn & Chater, 2016; Tenenbaum et al., 2006).

One interesting violation of probability theory is the well-established dilution effect (e.g., Nisbett et al., 1981). According to this effect, people sometimes revise their beliefs by decreasing their initial probability judgment on the basis of nondiagnostic new information that should be ignored. Although the dilution effect represents a robust phenomenon, the opposite finding—a confirmation effect—is sometimes also observed. People sometimes become

This article was published Online First July 22, 2021.

This research was supported by grants from the Swiss National Science Foundation (100014_126721/1 and 100014_138174/1) to Jörg Rieskamp. The first and second authors contributed equally to this work. Data, models, and model fits presented in this article are available at OSF: https://osf.io/x28um.

Results from Studies 1–3 and a first version of the model were presented as posters and presentations to the judgment and decision making community during conferences in Spain, Germany, Austria, Switzerland, and the United States and were a part of the PhD thesis by Mirjam A. Jenny (Basel, 2013). The data from all four studies and the current version of the model were a part of the PhD thesis by Rebecca Albrecht (Basel, 2020).

Correspondence concerning this article should be addressed to Rebecca Albrecht, Department of Psychology, University of Basel, Missionsstrasse 62a, Basel 4055, Switzerland. Email: rebecca.albrecht@unibas.ch

more certain in their beliefs after receiving nondiagnostic information (e.g., LaBella & Koehler, 2004). Past research suggested various theoretical accounts, such as modified Bayesian models or averaging-and-adjustment models (e.g., Hogarth & Einhorn, 1992), to explain some of the effects, but they cannot explain dilution and confirmation effects simultaneously.

The Impact of Nondiagnostic Information and Presentation Order

When an initial belief has to be revised in light of new, nondiagnostic evidence, theories based on probability theory predict that the initial belief is not changed, contrary to the dilution and confirmation effect.

The Dilution Effect

According to the dilution effect, an initial probability judgment $p(X|E_1)$ that an event X occurs given diagnostic evidence E_1 is *larger* than the revised probability judgment $p(X|E_1, E_2)$ after additional nondiagnostic evidence E_2 occurs. In Shanteau's (1975) classic numerical dilution study, for example, participants observed samples of red and white beads being drawn with replacement from either a box consisting of 70 white and 30 red beads or a box with 30 white and 70 red beads. After the presentation of each sample, participants had to estimate the probability that the sample came from the 70/30 box. Participants' mean probability judgments given a diagnostic sample decreased when they drew a nondiagnostic sample afterward.

The dilution effect has been observed in various areas (Macrae et al., 1992; McKenzie et al., 2002; Meyvis & Janiszewski, 2002; Peters et al., 2007), and in several everyday-life scenarios: For instance, in legal decision making, confidence in a verdict decreased after an independent verdict was received (McKenzie et al., 2002). In social reasoning, students' estimates about other students changed after they received nondiagnostic information (Peters & Rothbart, 2000). In auditor judgments, inexperienced auditors were affected by nondiagnostic information (Shelton, 1999), and auditors generally underweighted diagnostic information (Waller & Zimbelman, 2003).

Factors Mitigating the Dilution Effect

The dilution effect does not always occur, meaning that an initial probability judgment $p(X|E_1)$ that an event X occurs given diagnostic evidence E_1 sometimes *equals* the combined probability $p(X|E_1, E_2)$ after receiving nondiagnostic evidence E_2. Specifically, experts in relatively well-defined domains infrequently show the dilution effect. For instance, only the judgments by relatively less experienced auditors were diluted by nondiagnostic information, whereas judgments of experienced auditors were not (Shelton, 1999). Likewise in legal decision making, people were relatively immune to the dilution effect when they had much diagnostic information (Smith et al., 1998–1999).

The dilution effect also seems to be mitigated by the mode and type of stimulus presentation. In perceptual decision making, the dilution effect occurs less frequently and/or less strongly if naturalistic stimuli that promote automatic processing are used (Hotaling et al., 2015). The dilution effect is often not observed in situations where diagnostic information mixes nondiagnostic and diagnostic features (LaBella & Koehler, 2004; Sanborn et al., 2020). In social reasoning, the dilution effect is eliminated or even reversed by the evidence's typicality in a given situation (Peters & Rothbart, 2000). For example, when participants predicted the number of books a fraternity member would read outside of class assignments, the dilution effect occurred only after presenting information that was atypical for a fraternity member (does not like parties) and was reversed for typical information (is extroverted).

The Confirmation Effect

The confirmation effect represents the opposite of the dilution effect, so that an initial probability judgment $p(X|E_1)$ that an event X occurs given diagnostic evidence E_1 is *smaller* than the combined probability $p(X|E_1, E_2)$ after presenting additional nondiagnostic evidence E_2. Generally speaking, the confirmation effect describes that people often look for information to confirm rather than disprove their previous hypothesis (Jones & Sugden, 2001; Wason, 1968). Similarly, people tend to interpret information in a way that confirms their hypotheses (Lord et al., 1979; Plous, 1991). For example, the stepwise evolution of the preference paradigm (cf., Russo et al., 1998; for an overview see Russo, 2015) describes how people tend to interpret new, nondiagnostic information as being in favor of their original hypothesis when they update their preferences repeatedly.

LaBella and Koehler (2004) investigated in detail in which situations people show a confirmation effect versus a dilution effect in a repeated probability judgment task. They observed the dilution effect for probability judgments that were based on a mix of diagnostic and nondiagnostic information. However, in a belief-revision task, where participants had to revise an initial probability judgment, LaBella and Koehler observed on average a confirmation effect when a diagnostic sample was followed by a mixed nondiagnostic sample and none when it was followed by a neutral one including only nondiagnostic features.

Order Effects

Sequential probability judgments based on two pieces of nondiagnostic evidence tend to be influenced by presentation order (see Hogarth & Einhorn, 1992, for an early overview and Trueblood & Busemeyer, 2011, for a more current review). Order effects describe the finding that the probability $p(X|E_1, E_2)$ for event X given evidence E_1 before evidence E_2 is not equal to the probability $p(X|E_2, E_1)$ where E_2 was presented before E_1. Order effects have been shown in various areas, including the standard "bookbag-and-poker-chip" task (Shanteau, 1970), risky monetary gambles (Hertwig et al., 2004), legal evidence in a jury trial (Furnham, 1986; Walker et al., 1972), and clinical evidence in a medical case (Bergus et al., 1998). For example, Bergus et al. (1998) found that physicians judged the probability that a patient was suffering from a urinary tract infection to be higher if they received indicative information last compared to when they received inconclusive evidence last.

Cognitive Theories for Probability Judgments and Belief Revision

In the following, we describe theories to explain human probability judgments and belief updating. We start with the Bayesian model following normative probability theory and explain how it has been extended to explain some probability judgment phenomena. Afterward, we introduce the probability-theory-plus-noise model (Costello & Watts, 2014, 2016) and the quantum probability model (Busemeyer et al., 2011). Then we introduce different approaches to modeling probability judgments and belief updating, especially based on similarities and weighting-and-adding mechanisms.

Bayesian Models

According to Bayesian theory, belief updating starts with a prior belief, which is then updated on the basis of subsequent observations following Bayes's theorem. The prior belief is updated according to the likelihood of the observed information leading to the posterior belief. When the prior belief can be represented by a probability, the updating process leads to a posterior probability. This process repeats itself with each new piece of evidence. In principle, standard Bayesian models cannot explain how people's probability judgments change through nondiagnostic evidence or the order in which evidence is presented.

However, there are some characteristics predicted by the Bayesian model: First, according to Bayes's theorem, probability estimates increase if additional, converging evidence supports an initial hypothesis.[1] For example, if a physician is sequentially presented with *converging evidence* suggesting that a patient has pneumonia, their probability estimate that the patient has pneumonia should continually increase. Second, the law of large numbers postulates that if an experiment is repeated a large number of times, the average of the results will be close to the expected value of the experiment. However, people are often *insensitive to sample sizes* following the "law of small numbers," according to which people think even small samples represent the underlying population well (Kahneman, 2011; Tversky & Kahneman, 1971). Third, a statistical principle related to the law of large numbers is the *regression to the mean* (Galton, 1886). Accordingly, after extreme observations of a random variable have been made, subsequent observations should fall closer to the variable's expected value.

Inductive Confirmation

To explain certain aspects of human probability judgments, researchers have proposed several extensions to Bayesian models. For example, inductive confirmation (Carnap, 1962; Tentori et al., 2013) is a mechanism included in Bayesian models to explain the conjunction fallacy. This mechanism assumes that the combination of an initial belief (or a prior distribution) and a new piece of evidence also depends on how much the new information differs from the prior. For example, assume a person draws a hidden card from a card deck and is asked to guess if the card shows a king. The initial probability of the card showing a king is 1/13. With new information that the card shows a picture, the probability of it also showing a king increases to 1/3, which increases the credibility of the card showing a king dramatically, although the probability of it showing a king is still less than the probability that the card does not show a king (2/3; Tentori et al., 2013).

The Probability-Theory-Plus-Noise Model

The Bayesian model provides a normative solution for probability judgments, but it might not present a plausible cognitive model (Tversky & Kahneman, 1974; but see Chater et al., 2006; Griffiths et al., 2010; Sanborn, & Chater, 2016; Tenenbaum et al., 2006). One promising alternative model for probability judgments is the probability-theory-plus-noise (PT+N) model (Costello & Watts, 2014, 2016) that can explain a range of probability judgment phenomena such as conservatism, subadditivity, the conjunction fallacy, and the disjunction fallacy. The PT+N model assumes that single probability judgments are the result of a sampling process that draws instances from memory. The probability of sampling an instance of a certain type A is the actual probability that an event of type A occurs, $p(A)$. However, there is a chance $d < .5$ that a sampled instance is read incorrectly, meaning that with probability d an event A is incorrectly read as $\neg A$.

Quantum Probability Theory

A different approach to explaining probability judgments and belief updating assumes that the cognitive process is not based on (traditional) probability theory. One increasingly popular explanation for belief updating is based on quantum probability theory (QPT; Pothos & Busemeyer, 2013; Pothos et al., 2013, 2015; Pothos & Trueblood, 2015; Trueblood et al., 2014). According to QPT, probabilities are computed geometrically via the projection of state vectors onto different (cognitive) subspaces and by computing the squared length of such projections (Trueblood & Busemeyer, 2011). The process of belief updating is the result of a change in perspective modeled in a vector space. Each dimension in the vector space represents the joint probability of a hypothesis and a piece of evidence. A belief that a hypothesis is true is then a point in the vector space represented differently depending on one's perspective. A change in perspective is mathematically modeled with a unitary transformation that is not commutative and thus leads to order effects (Busemeyer et al., 2011; Trueblood & Busemeyer, 2011). However, as of now there is no QPT model that can explain fallacies associated with nondiagnostic evidence, specifically, the dilution and the confirmation effect.

Similarity-Based Probability Judgments

Recently, similarity-based approaches have been used to explain subjective probability judgments in the form of the similarity heuristic (Read & Grushka-Cockayne, 2011), the representativeness as prototype similarity (Nilsson et al., 2005), and representativeness as relative likelihood (Nilsson et al., 2005). Generally, similarities can be identified conceptually according to rules or more automatically on the basis of perceived relations between objects. There are four traditional approaches to similarity (Goldstone & Son, 2005; Hahn, 2014): Geometric models represent items in a metric space and calculate similarity based on the geometric distance of objects. Feature-based approaches (e.g., Tversky, 1977) assume that features are represented as binary contrasts (e.g., black and nonblack), and thus similarity can be assessed by simple feature matching. Models based on structural alignment stem from the field of analogical reasoning and extend feature-based models by taking the structural alignment of features into account. Finally, transformational models assess similarities based on the number of operations one has to apply to transform one object into another.

[1] Note that this is true only if the pieces of evidence are independent.

Recently, quantum probability theory has been introduced to model similarity processes in cognition (Pothos & Busemeyer, 2013; Pothos et al., 2013, 2015; Pothos & Trueblood, 2015; Trueblood et al., 2014).

Weighting-and-Adding Models for Belief Updating

Alternative approaches to belief updating are weighting-and-adding theories that have repeatedly been demonstrated to describe people's controlled judgments well (Anderson, 1981, 1996; Juslin et al., 2008; Lopes, 1985, 1987; Roussel et al., 2002; Shanteau, 1970, 1972, 1975), explain order effects (Hogarth & Einhorn, 1992), and, recently, describe conjunctive probability judgments (Jenny et al., 2014; Nilsson et al., 2009, 2013). Even the integration of sensory input from different modalities is assumed to happen through a weighting-and-adding process (Ernst & Bülthoff, 2004). Hogarth and Einhorn (1992) have described belief updating with a weighting-and-adding process, which have also been discussed in the context of the dilution effect (LaBella & Koehler, 2004). Weighted-additive integration is mathematically equivalent to reinforcement learning, where people form expectancies or beliefs based on the information and feedback they accumulated in the past. They update their expectancies according to the prediction error, represented by the difference between the expectancies and the present information or feedback (cf., Sutton & Barto, 1998). Reinforcement learning theories have also accumulated substantially converging evidence at a neural level (e.g., Tobler et al., 2006).

Psychological Explanations for the Sensitivity to Nondiagnostic Information

Psychological explanations for the dilution and the confirmation effect are sparse. Theories built directly on probability theory, like Bayesian models and the PT+N model, naturally cannot explain the effects. Whether or not a model based on quantum probability theory can, however, has yet to be explored. Because such an exploration needs extensive theorizing and testing it is not part of the current work. A weighting-and-adding process has been thought to produce the dilution effect (Shanteau, 1975; Troutman & Shanteau, 1977), but these models cannot account for the confirmation effect (LaBella & Koehler, 2004). Verbal similarity theories (e.g., Nisbett et al., 1981; but see Peters & Rothbart, 2000) have provided another theoretical explanation for the dilution effect. According to Nisbett et al. (1981), for example, a nondiagnostic piece of evidence decreases the perceived similarity between a hypothesis and previously presented evidence, which dilutes belief in the hypothesis.

As alternative theoretical explanations, expectancy and representativeness have been discussed (Tetlock & Boettger, 1989; Troutman & Shanteau, 1977). According to expectancy theory, people form hypotheses about how subsequent samples are likely to look after the presentation of a first sample. Certainty decreases if these expectations are not met by a nondiagnostic sample. This account predicts that a first sample that is nondiagnostic does not have any influence on subsequent samples because it does not elicit a directed hypothesis. However, contrary to the expectancy account, nondiagnostic first samples have been found to influence subsequent judgments (Troutman & Shanteau, 1977). According to the representativeness account, people predict outcomes that are representative of an option. Therefore, similarity between an option and an outcome depends on the number of common features (Tetlock & Boettger, 1989).

The Similarity-Updating Model

The proposed similarity-updating model combines a similarity-based process and a weighting-and-adding process, thereby explaining several established phenomena from the domain of probability judgments and belief updating. The initial judgment is based on a similarity process, specifically the idea that the belief that a hypothesis is thought to be true is an increasing function of the similarity between the evidence and the hypothesis. This initial belief determined by similarity is then updated following two psychological processes: (a) a weighting-and-adding process for combining two judgments and (b) a similarity-based confirmation process modifying the combined judgment. In this model, the dilution effect is the result of a combination of a weighting-and-adding mechanism, as proposed by Shanteau (1975), and a similarity-based process, in line with the ideas behind representativeness (Tetlock & Boettger, 1989; Troutman & Shanteau, 1977) and verbal similarity theories (Nisbett et al., 1981). The confirmation effect is the result of a similarity-based process alone and counteracts the weighting-and-adding process.

Formation of Initial Beliefs

In a nutshell, people first form two similarity judgments, which are based on the metric distance between a piece of evidence E_1 and hypothesis A, and the distance between the same evidence E_1 and hypothesis B and then compare these similarities to compute a probability judgment that hypothesis A is true. According to our model, first a distance in the metric space (e.g., Nosofsky & Johansen, 2000) is computed:

$$d_{ij} = \left[\sum\nolimits_m w_m \times |x_{im} - x_{jm}|^r\right]^{\frac{1}{r}}, \tag{1}$$

where x_{im} is the value of the ith piece of evidence on a psychological dimension m, x_{jm} is the value of hypothesis j on the psychological dimension w_m is the weight put on the dimension m, and r defines the metric ($r = 1$ for the city block metric, $r = 2$ for the Euclidean metric).[2]

This distance is transformed into a similarity between the evidence i and the hypothesis j by a nonlinearly decreasing function:

$$s_{ij} = e^{(-c \times d_{ij}^l)}, \tag{2}$$

where c is a sensitivity parameter and l determines the form of the similarity gradient ($l = 1$ for the exponential similarity gradient, $l = 2$ for the Gaussian similarity gradient).[3,4] The sensitivity

[2] Note that this definition of similarity assumes an equal number of features in i and j. Considering the task at hand, if we assume that sampling a certain feature (e.g., green cards) immediately brings one to reject a deck because it does not contain that feature, we need to (a) extend the missing features with an F (false) and (b) extend our notion of similarity:

$$s_{i,j}^{\text{ext}} = \begin{cases} s_{i,j} \text{ if } \forall A_{i_k} \in i = (i_1, \ldots, i_n), \forall j_k \in s = (j, \ldots j_n) : (i_k \wedge j_k) \\ 0 \quad \text{else} \end{cases}.$$

The conjunction $(i \wedge j)$ becomes false if one conjunct is false. Intuitively, false in the present paradigm means that a sample s cannot have been drawn from a deck A. In that case, the similarity is 0 and, thus, the probability for that deck is also 0.

[3] Similarity judgments based on such separable features are usually better described by a city-block metric as opposed to a Euclidean metric (Shepard, 1987), which is why r was fixed to 1. We did not find a reason to assume that these dimensions are weighted unequally, so w was fixed to 1/3 in our model (as we had three color dimensions in our task).

[4] According to Shepard (1987), an exponential similarity gradient is preferred in the case of discriminable stimuli, which is why we fixed l to 1.

$$p(A|E_1, E_2) = \begin{cases} p_{\text{avg}}(A|E_1, E_2) + c(A|E_1, E_2) \times (1 - p_{\text{avg}}(A|E_1, E_2)), & \text{if} c(A|E_1, E_2) \geq 0 \\ p_{\text{avg}}(A|E_1, E_2) + c(A|E_1, E_2) \times (p_{\text{avg}}(A|E_1, E_2) - 0.5), & \text{if} c(A|E_1, E_2) < 0 \end{cases}, \qquad (6)$$

parameter represents the subjective perception of similarity relative to distance. If c is high, then objectively small distances between evidence and hypotheses result in low similarity judgments (Appendix A compares likelihoods and similarities).[5]

The similarities between the evidence E_1 and hypothesis A ($s_{E_1,A}$) and the evidence E_1 and hypothesis B ($s_{E_1,B}$) are then transformed into probabilities (Luce, 1959):

$$p(A|E_1) = \frac{1}{1 + e^{\theta(s_{E_1,B} - s_{E_1,A})}}, \qquad (3)$$

where θ is a free parameter that determines how strongly hypothesis A is favored over hypothesis B if the evidence speaks for hypothesis A. The larger this parameter's value, the more clearly hypothesis A is favored; the smaller θ is, the more conservative (closer to .50) estimates of probabilities become.

Updating Process

In light of new information, a new similarity judgment is formed according to Equations 1 and 2 and transformed into probability judgments according to Equation 3. The updating process is based on two different psychological processes: (a) a weighting-and-adding process for combining two judgments and (b) a similarity-based confirmation process modifying the combination of single probabilities. The first component follows the work of Hogarth and Einhorn (1992) that assumes there is a belief-adjustment process where new information is integrated with old information in a weighted-additive fashion. The second component explicates the ideas sketched by LaBella and Koehler (2004) by transferring the functional definition of inductive confirmation (Tentori et al., 2013) to a similarity-based model for probability judgments.

Weighting-and-Adding Process

In a first step, the two probability judgments are combined in a weighting-and-adding process as proposed by Hogarth and Einhorn (1992):

$$p_{\text{avg}}(A|E_1, E_2) = (1 - \tau) \times p(A|E_1) + \tau \times p(A|E_2), \qquad (4)$$

where τ is a recency parameter with values between 0 and 1 that measures how much weight is put on the more recent piece of evidence, $p(A|E_2)$.[6] The τ parameter is comparable to the weight parameter in the value-updating model for risky choice (Hertwig et al., 2005).

Confirmation Process

In a second step, the result of the weighting-and-adding process is modified by a confirmation mechanism. In a nutshell, if people believe hypothesis A is likely to be true given evidence E_1 [i.e., $p(A|E_1) > 0.5$ according to Equation 3], they have a tendency to interpret a new piece of evidence E_2 relative to what they have learned about hypothesis A through E_1 (LaBella & Koehler, 2004). According to Tentori et al. (2013), inductive confirmation can be defined formally as a function $c(h, e)$ that is positive [$c(h, e) > 0$] if the probability of a hypothesis h after an additional piece of evidence e is greater than the prior probability [$P(h|e) > P(e)$]; negative [$c(h, e) < 0$] if the probability of the hypothesis h after an additional piece of evidence e is smaller than the prior probability; and 0 if they are equal. Several different instantiations of this general framework have been proposed (e.g., Crupi & Tentori, 2010; Fitelson, 2006; Tentori et al., 2007).

We apply this idea to the context of subjective probability judgments that are the result of a similarity process by considering the difference in similarity estimates between the two pieces of evidence E_1 and E_2 and the chosen hypothesis ($s_{E_2,A} - s_{E_1,A}$). More precisely, if the new piece of evidence is more similar to the favored hypothesis than the previous piece of evidence, that is, ($s_{E_2,A} - s_{E_1,A}$) > 0, then the hypothesis is *confirmed*, leading to an increase in the probability estimate. If, however, it is less similar ($s_{E_2,A} - s_{E_1,A}$) < 0 and does not confirm the previous evidence, the probability estimate decreases:

$$c(A|E_1, E_2) = (s_{E_2,A} - s_{E_1,A}). \qquad (5)$$

The final predicted probability judgment is the averaged probability, $p_{\text{avg}}(A|E_1, E_2)$, adjusted by the confirmation process, $c(A|E_1, \text{E}_2)$: (See above Equation 6)

The adjustment is weighted by ($p_{\text{avg}}(A|E_1, E_2) - 0.5$) or ($1 - p_{\text{avg}}(A|E_1, E_2)$) because the size of the adjustment should depend on the predicted averaged probability and how close it is to possible minimum and maximum response values (in our case responses were restricted to values between 50% and 100%). If the averaged probability estimate is already very high, a positive adjustment would naturally be smaller than if the averaged probability was around .50.

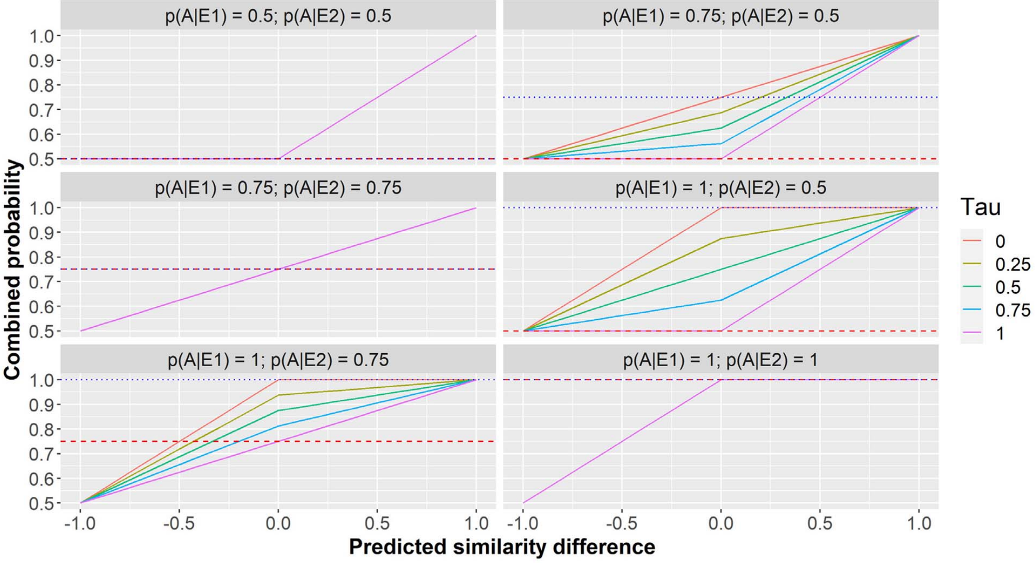
Figure 1 illustrates how the predicted, combined probability in favor of hypothesis A, $p(A|E_1, E_2)$ in Equation 6, depends on the difference between the two single probabilities $p(A|E_1)$, $p(A|E_2)$, and on their distance to the maximum and minimum response values (cf. Equations 4 and 6), the difference between similarities $s_{E_2,A}$, $s_{E_1,A}$ (Equation 5) and parameter τ (Equation 4).

[5] Note that the similarity-updating model uses the general concept of similarity to model judgments and applies meaningful defaults for parameter values where possible; which of the two distance matrices (city-block metric or Euclidean metric) and which form of the similarity gradient (exponential or Gaussian) is more appropriate depend on the types of stimuli used.

[6] Note that Equation 4 can be reformulated in the tradition of reinforcement learning models ($p_t(A) = p_{t-1}(A) + \beta[p(A) - p_{t-1}(A)]$) where β is the weighting parameter with values between 0 and 1, which indicates the weight on the difference between the two probability judgments—in other words, on the prediction error of the first judgment relative the second judgment, the latter being based on additional information (Sutton & Barto, 1998). Also note that if the weighting parameter takes the function of $\beta = 1/(n - 1)$, Equation 4 produces the current or running mean over all previously encountered pieces of evidence.

Figure 1

Predictions of Combined Probabilities, $p(A|E_1, E_2)$, Relative to the Predicted Similarity Between the Two Pieces of Evidence and the Hypothesis ($s_{E_2,A}$ and $s_{E_1,A}$) and Different Combinations of Constituent Probabilities $p(A|E_1)$ and $p(A|E_2)$



Note. Colored lines correspond to different values of parameter τ (Equation 4). The blue dotted line is the maximum and the red dashed line is the minimum of the two constituent probabilities. The weighting-and-adding models always predict values within the range of single probabilities. The similarity-updating model predicts values outside the range of single probabilities, depending on the similarity distance. See the online article for the color version of this figure.

Explaining the Dilution and the Confirmation Effect

The similarity-updating model predicts the dilution effect as the result of two possible sources. First, as proposed by Shanteau (1975), the dilution effect generally occurs because of the nature of the weighting-and-adding mechanism (Equation 4). In addition, the dilution effect might be increased if the new, nondiagnostic piece of evidence E_2 is less similar to hypothesis A than the diagnostic piece of evidence E_1. For an example, consider the upper right panel in Figure 1, with probability $p(A|E_1) = 0.75$ and the probability $p(A|E_2) = 0.5$. Assuming a similarity difference of 0, the dilution effect is observed if parameter τ is not equal to 0 (if $\tau = 0$ the second piece of evidence would be completely ignored). The dilution effect increases with a decreasing similarity difference (i.e., the first piece of evidence is more similar to the favored hypothesis than the second, nondiagnostic piece of evidence). However, with an increasing similarity difference, the dilution effect decreases, up to the point where it is reversed and a confirmation effect is observed (probability judgments above the blue dotted line in Figure 1). On average, the model predicts that dilution effects are observed more frequently and should be larger compared to confirmation effects.

With the similarity-based confirmation mechanism, the similarity updating model explains a number of findings from the literature. For example, if the nondiagnostic evidence is neutral, in the sense that it includes mainly information that is not represented in both hypotheses, the dilution effect is predicted to be higher (or more likely to happen) than if the nondiagnostic evidence is mixed, meaning that it includes information used to describe the hypotheses, where confirmation effects are more likely (LaBella & Koehler, 2004). The reason is that mixed nondiagnostic samples are on average more similar to the hypotheses than neutral nondiagnostic samples. Similarly, findings that the typicality of evidence mitigates or even inverts the dilution effect (leading to the confirmation effect; Peters & Rothbart, 2000) might be a result of the similarity between a held hypothesis (fraternity members do not read) and how similar a new piece of evidence is to that hypothesis (does not like parties). Our model is in principle also able to explain why individuals show the dilution effect in some trials but not in others.

Empirical Investigation

Paradigm

To test our similarity-updating model, we implemented the classic bookbag-and-poker-chip paradigm as a computerized card game (Edwards, 1968). This task was designed specifically to investigate deviations from Bayesian theory in human belief formation and the belief-updating process, such as the conservatism bias (Corner et al., 2010; Edwards, 1968; Peterson et al., 1965; Peterson & Miller, 1965). Since its development, several versions of it have been used as a standard task to assess people's probability judgments (e.g., Jenny et al., 2014; Nilsson et al., 2013; Shanteau, 1975) and specifically to assess the dilution effect (e.g., LaBella & Koehler, 2004; Troutman & Shanteau, 1977). We chose this paradigm with stochastic events to rule out that nondiagnostic evidence could carry semantic information about hypothesis *A* or *B* (cf., Peters & Rothbart, 2000).

On each trial, participants received two novel card decks (*A* and *B*) and samples were drawn with replacement from a randomly chosen and undisclosed deck. Each deck totaled 100 cards and contained blue, red, and green cards. Above the cards, numbers between 10 and 80 indicated how many cards per color each deck contained. After receiving the first sample, participants were asked to indicate the deck the sample had been drawn from and to estimate the probability that the chosen deck had generated the sample (by pointing on a scale between the values 50% and 100%). After receiving a second sample, participants were asked to identify the deck from which *both* samples came and state the probability that their deck was the source of the samples.

In all experimental studies, mixed and neutral nondiagnostic samples were tested. For example, assume that deck *A* consists of 30% blue, 20% red, and 50% green cards and deck *B* 20% blue, 30% red, and 50% green cards. A nondiagnostic sample that is neutral to these two decks would consist of only green cards, because both decks have the same number of green cards. A mixed nondiagnostic sample, on the other hand, might include two blue cards, two red cards, and three green cards. It is easy to see that the neutral nondiagnostic evidence is less similar to (or typical for) both decks than the mixed nondiagnostic sample.

In Study 1, two samples of seven cards each were presented sequentially and both samples were visible on the screen during the whole trial. In Study 2, the first sample was removed from the screen when the second sample appeared. In Study 3, three samples were sequentially presented. In Study 4, half of the trials included samples with 14 instead of seven cards.

Model Testing

We tested our model qualitatively and quantitatively against the Bayesian model (Appendix B) and the PT+N model (Costello & Watts, 2014, 2016; Appendix C) in four different experimental studies. The similarity-updating model makes different predictions from the alternative models. As discussed in detail above and to the best of our knowledge, it is the only model to explain the dilution effect and the confirmation effect within one framework. Second, the Bayesian model and the PT+N model predict that probabilities increase in light of sequentially presented converging evidence; the similarity-updating model, however, predicts that probability judgments can decrease if later converging evidence is less predictive than earlier evidence. Second, the Bayesian model and the PT+N model predict on average no effects of presentation order of the combined judgments, because noise causes the probability judgments to be centered around the predicted mean. In contrast, the differential weighting of sequentially presented pieces of evidence leads the similarity-updating model to produce order effects. Third, the similarity-updating model is insensitive to changes in sample size while the Bayesian model and the PT+N model predict judgments become more extreme and have decreasing trial-by-trial variance with increasing sample size.

We also compared the models to a benchmark baseline model. Per person, this model predicts the mean of the observed first probability judgments as a first probability judgment and the mean of the observed second probability judgments as a second probability judgment in all trials. Any cognitive model that claims to provide a good account of probability judgments needs to outcompete the baseline model as a plausibility check.[7] For all models we assumed an error process so that the predicted judgment would vary around the most likely point estimate (cf. Budescu et al., 1997; Juslin et al., 1997). We used a normalized truncated normal probability density likelihood function to link the models' point predictions with people's judgments.

In Studies 3 and 4 we additionally tested how generalizable the predictions of the different models are by estimating the models on a subset of the data and generalizing the predictions of the models to the remaining data set. Study 3 tested the generalizability of the proposed belief-updating mechanism and Study 4 tested the generalizability across different sample sizes.

Study 1: Full Display of Samples

The dilution effect has been shown in many different domains (Hackenbrack, 1992; LaBella & Koehler, 2004; Macrae et al., 1992; McKenzie et al., 2002; Meyvis & Janiszewski, 2002; Peters & Rothbart, 2000; Shanteau, 1975; Shelton, 1999; Smith et al., 1998–1999; Troutman & Shanteau, 1977; Waller & Zimbelman, 2003). However, sometimes the opposite effect is observed, namely, a confirmation effect (LaBella & Koehler, 2004; Russo et al., 1998). Study 1 aimed to replicate findings of how people update beliefs when facing nondiagnostic information, especially the dilution effect. Additionally, Study 1 evaluates quantitatively and qualitatively the similarity-updating model.

Method

Participants

Twenty-five undergraduate students ($Mdn_{\text{age}} = 22$ years, 76% women, 24% men) at the University of Basel participated. Participants were compensated with either course credit or book vouchers worth 15 Swiss francs (CHF). Additionally, they received a performance-contingent bonus ($Mdn = 2.10$ CHF).

Materials

The experiments were computerized. Participants were presented with a diverse set of randomly ordered "games" involving two decks of cards. For the distributions of cards in the first deck, all combinations of three underlying probabilities of 10%, 20%, 30%, 40%, 50%, 60%, 70%, and 80% were used. These probabilities—for example, 20%/50%/30% red, blue, and green cards—always added up to 100%. To construct the second deck, one of the probabilities was held constant, and the other two switched positions, resulting in 30%/50%/20% red, blue, and green cards, for example.

The sample distributions for 81% of all trials were determined by randomly drawing seven times from a Dirichlet distribution with the underlying probability distribution of the picked deck. The Dirichlet distribution is a multivariate generalization of the beta distribution

[7] To additionally test the ecological validity of the similarity-updating model, we ran a simulation that parallels simulations by Juslin et al. (2009). This simulation was intended to test if the model leads to good judgments in our environment as compared to a normative solution as given by likelihood computation and Bayes's theorem. In sum, the similarity-updating model's binary predictions are well adapted to the environment and their accuracy increases with decreasing sampling error.

and takes a symmetrical (i.e., "uniform") shape when its parameters are all set to 1. Sampling values from a three-parameter Dirichlet distribution (with all three parameters set to 1) produces three values between 0 and 1 that sum up to 1. For 19% of all trials, one sample was randomly sampled and the other sample was tweaked such that its likelihood of being drawn from deck *A* was identical to its likelihood of coming from deck *B*. These nondiagnostic samples consisted of either only cards of the color that was equally represented in the two decks (neutral nondiagnostic samples) or a certain number of this card and an equal number of the other two cards (mixed nondiagnostic samples). To simplify the task for the participants, the samples were sorted according to color. All games were presented twice, once with the original order and once with a switched order of samples. This means that all nondiagnostic samples were presented first once, which allowed us to elicit single probability judgments for them. Appendix D shows a visual representation of how stimuli were presented to the participants.

Procedure

Study 1 involved 86 rounds consisting of one game with two samples each. The two samples were presented sequentially, with both samples visible on the screen at the end of a trial. In each round, participants were asked to choose the deck that they thought was more likely to have generated the samples they drew. Additionally, they stated the probability that their chosen deck and not the other one had generated the sample. At the end of the experiment, one round was randomly picked and participants won 2.50 CHF if they had chosen the right deck in that round. The participants received an additional reward with a maximum of 5 CHF for the probability judgment that they had provided after having seen two samples in that round. The reward was based on an inverse Brier score (Brier, 1950) by subtracting the squared difference between the judged probability and the outcome score from 1. If, for example, deck *A* was the right choice, the outcome score for deck *A* was 1, and if the participant assigned this deck a probability of 70%, then the inverse Brier score was $1 - (.70 - 1)^2 = .91$. This score was multiplied by five, rounded to one decimal point, and paid in Swiss francs.

After providing informed consent, participants read through the instructions on the screen and received a printed version of the instructions (see Appendix E, German version translated to English), which they could hold on to throughout the experiment. The instructions particularly stressed that samples were drawn with replacement and that the probability judgment always concerned the chosen deck. Participants were also informed about the success-dependent bonus with a maximum of 7.5 CHF. The procedure included five practice trials and afterward participants could ask remaining questions concerning the task and procedure, if necessary. At the end of the experiment, we checked if participants understood the task by asking them to write down how they had solved the task. Twenty-two of the 25 participants had understood the task. For three participants, it was not clear if they had understood that within one trial both samples were drawn from the same deck.

Results

Participants' General Performance

To assess participants' general performance we compared their responses with the normative solution following Bayes's theory. Table 1 shows the performance after the first sample (if it was diagnostic) and after both samples had been presented for Study 1 and also for the three following studies. Participants chose the correct deck in 82% of all trials after the first sample and in 82% after the second sample was presented. The normative solution fared around 10% better in both cases. In 20% of trials, participants changed their minds between the first and the second sample. The majority (65%) of these switches resulted in correct choices. Interestingly, the root mean square deviation between participants' probability judgments and the normatively correct probabilities increased (0.17–0.21) and the correlation between them decreased (.67–.51) across samples, showing a decrease in judgment accuracy. Apparently people's updating process distorted the second probability judgments. Nevertheless, although people's probability judgments differed considerably from the normative solution, they correlated with it and allowed participants to choose the correct deck in most trials.

Reaction to Nondiagnostic Evidence

First Sample. In principle, people were able to detect and correctly identify nondiagnostic samples if they were presented first, suggesting that the effect of nondiagnostic information was driven by the belief-updating process and not by the process of judging single probabilities. On average, participants correctly identified a first nondiagnostic sample (response 50%) in 88% (median; range: 0%–100%) of all trials. However, this varied between different types of nondiagnostic evidence: For neutral samples, participants correctly identified the nondiagnostic samples in a median of 100% of the trials (interquartile range [IQR] = 90%–100%) and for mixed samples correct identification decreased to a median of 67% (IQR = 17%–83%) of all trials.

Second Sample. In this analysis, we focused on trials in which the second sample was nondiagnostic (its likelihood of coming from either deck was .50), participants' judgments after only the first samples were >.50 (treating it as diagnostic and leaving room for a decreased second judgment), and participants did not change their choice between samples, the latter because choice changes based on nondiagnostic samples when the first samples were diagnostic could potentially be due to guessing and not only to perceiving a

Table 1
Percentage of Trials Where the Correct Deck Was Identified

Study	Sample	Participants	Normative	Correlation ρ	RMSD
1	1	82%	90%	.67	0.17
	2	82%	91%	.51	0.21
2	1	83%	90%	.58	0.19
	2	83%	91%	.45	0.22
3	1	75%	77%	.61	0.19
	2	73%	75%	.50	0.23
	3	75%	81%	.22	0.22
4	1	82%	89%	.55	0.20
	2	83%	93%	.33	0.25

Note. RMSD = Root Mean Square Deviation. The normative solution was calculated using likelihood and Bayes's theorem. The correlation shows the median Spearman correlation coefficient between participants' probability judgments and the normatively correct probabilities and the respective RMSD.

nondiagnostic piece of information as diagnostic or to a specific information-integration process.

In total, analyzed nondiagnostic trials accounted for 14% of all trials. Table 2 shows the percentage of these trials in which participants showed a dilution effect, a confirmation effect, or no effect. Note that the latter is predicted by the Bayesian and the PT+N model. The similarity-updating model can predict all events, but in principle it would predict a larger proportion of dilution effects and a small proportion of confirmation effects and null effects. The results show that participants clearly showed the dilution effect in around 65% of these trials. Additionally, more than half (68%) of the participants showed the dilution effect in more than half of these trials.

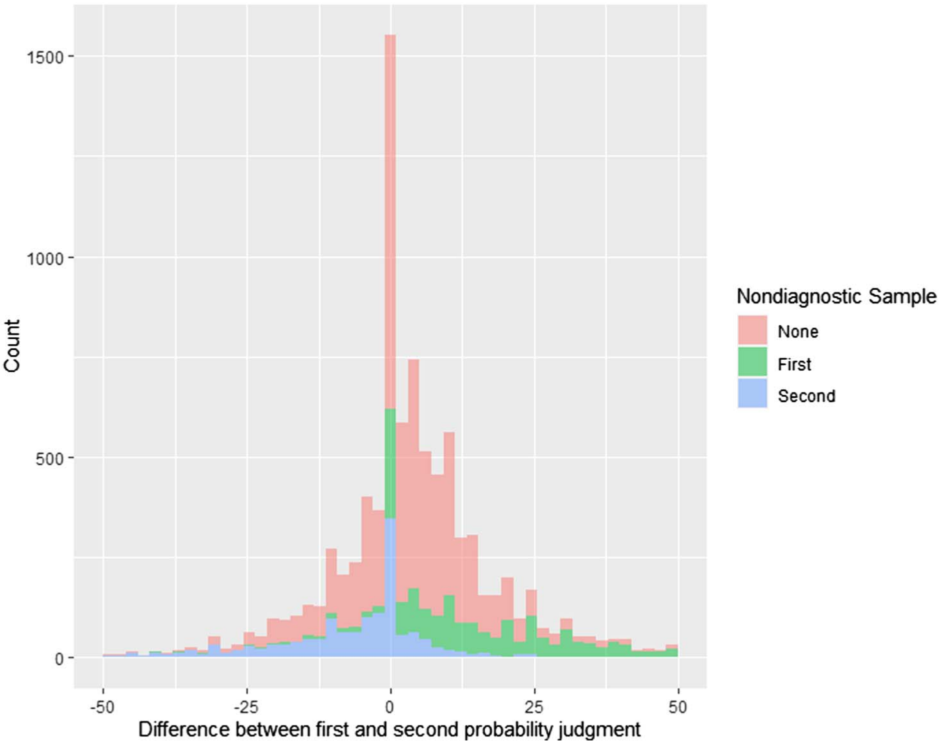
Dilution Effect. We performed a regression analysis to examine the strength of the dilution effect between the first and second probability judgment. In principle, if the participants ignored nondiagnostic information, the slope of the regression line would have a value of 1, whereas a smaller positive value would on average show the dilution effect, and a smaller negative value would on average show a confirmation effect. The observed slope was .54, indicating a strong dilution effect. Figure 2 shows the difference between participants' first and second probability judgments in all four studies, when the first sample, the second sample, or neither of the samples were nondiagnostic. When the second sample was diagnostic the majority of differences was positive regardless of whether the first sample was diagnostic or nondiagnostic, meaning that the second, combined probability judgment was greater than the first. When the second sample was nondiagnostic, however, the situation was reversed, with a clear majority of second probability judgments being smaller than the first, representing a dilution effect. This clearly shows that the dilution effect depends on the second sample being nondiagnostic and that the dilution effect cannot simply be explained as a regression to the mean effect.

There are different factors that have an impact on the dilution effect. When the first probability judgment was lower than or equal to the median probability judgment (74%) over all trials and participants, dilution effects were observed in 56% of the trials. When the first probability judgment was higher than the median, dilution effects were observed in 71% of the trials. Thus, the rate of dilution effects depended on the size of the first probability judgment. This, together with the finding that people recognized nondiagnostic

Table 2
Percentage of Dilution and Confirmation Trials

Effect	Study 1 *Mdn* (range)	Study 2 *Mdn* (range)	Study 3 *Mdn* (range)	Study 4 *Mdn* (range)
Dilution	64% (8%, 100%)	67% (7%, 93%)	62% (0%, 100%)	75% (0%, 100%)
Confirmation	21% (0%, 92%)	23% (0%, 67%)	20% (0%, 100%)	11% (0%, 75%)
None	0% (0%, 53%)	8% (0%, 57%)	0% (0%, 60%)	8% (0%, 61%)

Figure 2
Histogram of Differences Between Participants' First and Second Probability Judgments by Presentation of Nondiagnostic Sample in All Trials in All Four Studies



Note. See the online article for the color version of this figure.

information correctly in the first sample, indicates that if participants also identified the second, nondiagnostic decks as nondiagnostic, the integration of the two decks may have led to the dilution effect.

Dilution Versus Confirmation. For neutral nondiagnostic samples, dilution effects were observed in 64% of the trials, whereas for mixed nondiagnostic samples, dilution effects were observed in 61% of the trials. On average, the dilution effect was greater for neutral nondiagnostic samples (mean difference between the first and second probability judgment of −8%) compared to mixed nondiagnostic samples (mean difference of −3%). In line with LaBella and Koehler's (2004) studies, we also found more confirmation effects for mixed nondiagnostic evidence (*Mdn* = 40%; IQR = 0%–50%) than for neutral nondiagnostic evidence (*Mdn* = 15%; IQR = 0%–40%). The similarity-updating model explains these results: Mixed nondiagnostic samples were on average more similar to both decks (and thus the correct deck) than neutral nondiagnostic samples. As a result, the similarity-based confirmation mechanism led to smaller increases or even decreases in probability judgments for neutral samples (Equation 5).

In sum, the results replicate the dilution effect and differences between different types of nondiagnostic evidence. Additionally, our analyses clearly show that the dilution effect is not due to regression to the mean. The Bayesian model and the PT+N model can neither account for the dilution effect nor for different probability judgments following mixed and neutral nondiagnostic samples. Thus, the qualitative results support the similarity-updating model.

Averaging and Adjustment

Pure averaging models predict that all combined probabilities after the second sample was presented must lie between the single probability estimates for Samples 1 and 2. Bayesian models, on the other hand, predict that the combined probability estimates must always be larger than both single probability estimates when both samples are in favor of the same hypothesis. If one piece of evidence is nondiagnostic, the Bayesian model predicts that the combined probability judgment solely depends on the diagnostic piece of evidence. The similarity-based confirmation mechanism predicts that judgments could be smaller or larger than the single probability estimates under the condition that a second sample is much more (or less) similar to the chosen deck than the first sample. The predicted percentage of judgments outside the range of single probabilities depends on the similarity between a deck and a sample, which depends on values of free parameters (cf. Figure 1) and usually differs between people and tasks.

In our experiments each trial was presented twice to participants with a switched order of the samples. Thus, we could estimate the single probability assigned to a configuration of cards presented as a second sample and compare it with a trial where this configuration was presented as a first sample. On average, 47% (median; IQR = 42%–55%) of all combined probability judgments after participants saw two samples lay between the two single probability judgments (including the two single probabilities), 12% (median; IQR = 8%–21%) were smaller than both single probability judgments, and 36% (median; IQR = 30%–45%) were larger.

To specifically test if the similarity-based confirmation mechanism is consistent with participant responses, we additionally compared the predicted similarity difference (based on estimated individual parameter values, cf. Table 3) with the difference between the two observed probability judgments within one trial ($r = .23, p < .01$). Figure 3 shows a graphical representation of this correlation across all four studies. In sum, the similarity-based confirmation mechanism predicts the difference between first and second probability judgments well, and the similarity-updating model is most consistent with our findings.

Table 3
Median Parameter Values of All Models in All Studies

Model	Parameter	Study 1	Study 2	Study 3	Study 4
Similarity updating	c	0.13	0.44	0.57	0.08
	θ	39.77	14.59	10.01	95.69
	τ	.56	.60	.49	.52
	σ	0.13	0.14	0.11	0.12
Bayesian	σ	0.24	0.27	0.31	0.33
PT+N	d	.19	.19	.25	.19
	σ	0.20	0.22	0.19	0.25
Baseline	σ	46,910.56	2,357	0.41	0.61

Note. PT+N = probability-theory-plus-noise model.

Order Effects

The Bayesian model and the PT+N model predict trial order invariance in that the order in which the samples are presented does not influence the final probability judgments. In contrast, the similarity-updating model predicts order effects, which result from the unequal weighting of the first and the second sample. Indeed, the median absolute deviation between the two second probability judgments for tasks with the same configuration presented in inverse order was 16% (IQR = 13%–17%) over all participants. Thus, the fact that participants' probability judgments were affected by sample order speaks against the Bayesian model and the PT+N model on a qualitative level (assuming probabilities are taken at face value). Table 4 shows the order effects for all four studies.

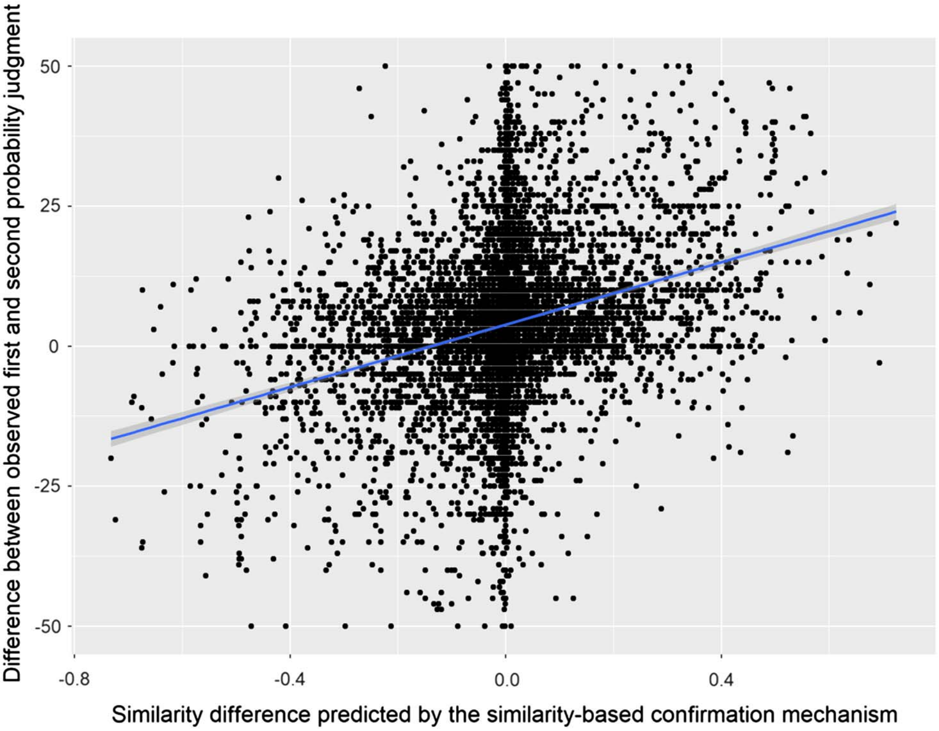
Converging Evidence

We investigated converging evidence by looking at all trials in which both samples stemmed from one of the two decks with a likelihood of >.50, in which participants picked the same deck after both samples, and in which the normative solution predicted the same deck after all samples. These trials constituted 53% of the whole set in Study 1. In 83% (median) of the subset of trials in which according to the normative solution the second sample was more diagnostic than the first one, participants' final probability judgments exceeded their first judgments (range: 67%–91%). In 36% (median) of the subset of trials in which the first sample was more diagnostic than the second one according to the normative solution, participants' judgments decreased between the initial and the final estimate (range: 30%–61%). Table 4 shows the results for converging evidence in all four studies.

Thus, in contrast to the normative solution but in line with the similarity-updating model, converging evidence did not always lead participants to increase their probability judgments, and how participants treated converging evidence depended on the presentation

Figure 3

Correlation of the Difference in Participants' Probability Judgments in All Four Studies Between the First and the Second Judgment (Combined Probability Estimate for Samples 1 and 2 Minus the Probability Estimate for Sample 1) and the Similarity Difference Predicted by the Similarity-Updating Model (Equation 1)



Note. Predictions based on estimated individual parameter values (cf. Table 3). See the online article for the color version of this figure.

order of the samples, thus providing us with qualitative evidence in favor of the similarity-updating model.

Quantitative Model Comparison

We estimated all models[8] on the basis of the participants' individual complete data using maximum likelihood estimation and compared models by the Bayesian information criterion (BIC; Schwarz, 1978), which takes model complexity into account. The BICs for the four competitor models are listed in Table 5 and the median optimal parameter values are listed in Table 3. The median value of the weight parameter τ was .56, indicating that the second piece of information was weighted more than the first, meaning that we observed a recency effect.

The similarity-updating model,[9] clearly outperformed the competing models. The median BIC of the similarity-updating model across all participants was -242 while for the PT+N model it was -139. Also, the similarity-updating model explained the responses of 24 of the 25 participants best according to model selection based on BIC. One person was best described by the PT+N model. Neither the Bayesian model (median BIC of -98) nor the baseline model (median BIC of 4) performed well in comparison. In sum, the first experiment provides strong evidence that people's probability judgments are better described with the similarity-updating model than with the PT+N or the Bayesian model.

Discussion

Study 1 shows qualitative and quantitative findings supporting the similarity-updating model as compared to the competing models. The dilution effect, effects of different types of nondiagnostic samples, the existence of order effects, and the results on converging evidence are

[8] Data, models, and model fits presented in this article are available at OSF: https://osf.io/x28um.

[9] With the similarity-updating model, we tested a specific instantiation of a similarity process and a specific belief updating mechanism. In a preliminary, explorative analysis we evaluated variants of both processes as alternatives to the here presented versions using the data of Study 1. As an alternative similarity process, we tested the similarity heuristic (Read & Grushka-Cockayne, 2011), representativeness as prototype similarity (Nilsson et al., 2005), and representativeness as relative likelihood (Nilsson et al., 2005). As alternative updating processes we tested the Sigma model (Juslin et al., 2008) and a traditional belief-updating mechanism (Hogarth & Einhorn, 1992; which in our setting is mathematically equivalent to simple averaging). This preliminary work provided empirical evidence for the similarity-updating model. A detailed description of this analysis is available as Supplemental Material.

Table 4

Order Effects and Effects on Converging Evidence

Effect	Study 1 *Mdn* (IRQ)	Study 2 *Mdn* (IRQ)	Study 3 *Mdn* (IRQ)	Study 4 *Mdn* (IRQ)
Order	16% (13%–17%)	17% (14%, 19%)	7% (6%, 9%)	17% (16%, 23%)
Converging: Increase	83% (67%, 91%)	84% (61%, 95%)	80% (69%, 88%)	83% (65%, 92%)
Converging: Decrease	36% (30%, 61%)	26% (11%, 39%)	13% (0%, 27%)	32% (23%, 69%)

Note. IQR = Interquartile range. Order effects show the median absolute deviation between the two second probability judgments for tasks with the same configuration presented in inverse order. Converging evidence is calculated over a subset of trials where all samples stemmed from one of the two decks with a likelihood of >.50, in which participants picked the same deck after all samples, and in which the Bayesian model picked the same deck after all samples. Included percentage of trials was 53% in Study 1, 54% in Study 2, 38% in Study 3 (because condition extended to the third sample), and 53% in Study 4. Increase shows the trials in which according to the normative solution the last sample was more diagnostic than the other samples. Decrease shows the trials in which the first sample was more diagnostic than the other samples according to the normative solution.

all in line with the predictions of the similarity-updating model. A quantitative model comparison based on BIC supports these results.

Study 2: Sequential Display of Two Samples

In Study 1, the first sample was still present at the time of the presentation of the second sample. Thus, the updating process could be based on description and performed with perfect memory. However, this is not always the case, as people often receive bits of information sequentially, have access to only one piece of evidence at a time, and have to remember previous information. The goal of the second study was to replicate the results from Study 1 in a setting in which the two samples were never simultaneously presented.

Method

Participants

Twenty-six undergraduate students (Mdn_{age} = 23.0 years, 58% women, 42% men) at the University of Basel participated and were compensated with either course credit or book vouchers worth 15 CHF. Additionally, they received a performance-contingent bonus (*Mdn* = 2.65 CHF).

Materials

In contrast to Study 1, in Study 2, samples were presented in a truly sequential manner. Additionally, a screen between rounds announced the next round and instructed participants to start the next round by pressing the letter “W” (which stood for the German word *weiter*, which means “proceed” in this context).

Procedure

The procedure in Study 2 was identical to that in Study 1 except that the first sample disappeared when the second one was presented. All participants understood the task.

Results and Discussion

Table 1 shows participants’ general performance, Table 2 the results for nondiagnostic evidence, and Table 4 the results for order effects and converging evidence across all four studies. Participants showed similar results for order effects and converging evidence to results found in Study 1.

Participants’ General Performance

Participants performed equally as well as in Study 1 (Table 1). In 17% of trials, participants changed their mind between the first and the second sample. The majority (69%) of these switches resulted in correct choices.

Reaction to Nondiagnostic Evidence

First Sample. The median percentage of correctly identified first nondiagnostic samples averaged around 81% (range: 6%–100%). If the nondiagnostic sample was a neutral sample, participants correctly identified it in a median of 100% of the trials (IQR = 73%–100%). This performance was lower for nondiagnostic mixed samples (*Mdn* = 50%; IQR = 33%–100%). We found confirmation effects with a median of 11% of the trials (IQR = 0%–30%) for neutral samples and a median of 33% (IQR = 20%–60%) for mixed samples.

Second Sample. In Study 2, 14% of trials met the criteria to be analyzed as second, nondiagnostic samples. Table 2 shows the percentage of these trials in which participants showed a dilution effect, a confirmation effect, or no effect. The slope of a simple linear regression was .52, illustrating, on average, a dilution effect. One participant showed a dilution effect in 0% of all trials, but this was because this person responded “.50” on all trials. More than half of the remaining participants (56%) showed the dilution effect in more than half of these trials. As in Study 1, participants showed fewer dilution effects (49%) when their first probability judgment was lower than or equal to the median probability judgment over all trials and participants (*Mdn* = 74%) than when their first probability judgment was higher (68%). Further, participants showed dilution effects in 61% of the analyzed trials if the nondiagnostic sample was neutral, but in 53% of mixed nondiagnostic trials. The median downward adjustment of

Table 5

The Models’ Bayesian Information Criteria (BICs) for the Probability Judgments in All Studies

		Model			
Study	Result	Similarity updating	Bayesian	PT+N	Baseline
1	Median BIC	−242	−98	−139	4
	n	24	0	1	0
2	Median BIC	−247	−84	−133	4
	n	25	1	0	0
3	Median BIC	−408	−83	−184	−29
	n	23	0	1	0
4	Median BIC	−251	−55	−96	−3
	n	23	1	1	0

Note. PT+N = Probability-theory-plus-noise model; BIC = *Bayesian Information Criteria.*

the probabilities after the presentation of the nondiagnostic piece of information in these trials was 11% (IQR = 5%–23%). We could thus replicate our findings from Study 1 irrespective of whether the initial samples were present when the subsequent samples were shown.

Averaging and Adjustment

On average, 48% (median; IQR = 39%–55%) of all combined, second probability judgments lay between the two single, first probability judgments. Around 8% (median; IQR = 5%–14%) were smaller than both single probability judgments and 42% (median; IQR = 28%–50%) were larger. The correlation of the similarity-based confirmation mechanism's predictions (based on estimated parameter values, cf. Table 3) and the difference between the two observed probability judgments within one trial was again positive ($r = .36$, $p < .01$; cf. Figure 3).

Quantitative Model Comparison

In Study 2, the first sample disappeared as soon as the second sample appeared. We hypothesized that this manipulation would not change the cognitive process behind people's updating behavior. We estimated all models in the same ways as in Study 1. The median BIC over all participants for the four models was 4 for the baseline model, −84 for the Bayesian model, −133 for the PT+N model, and −247 for the similarity-updating model, indicating that the latter provided the best model fit. The median optimal parameter values of all models are listed in Table 3. According to model selection based on the BIC, 25 of 26 participants were best described by the similarity-updating model and one person was best described by the Bayesian model.

The median value of the weight parameter τ, at .60, was a bit higher than in Study 1 (.56). This indicates that the second piece of information was weighted more than the first in Study 2, where the first piece of evidence was removed from the screen when the second one was presented.

Study 3: Sequential Display of Three Samples

Study 2 provided strong evidence that the similarity-updating process describes not only description-based updating well, but also truly sequential updating including a memory component, in that the first sample or the first judgment had to be retrieved from memory when making a final judgment. Additionally, we wanted to test whether the similarity-updating model could also describe a longer updating process well, one that is based on more than two samples. In Study 3, participants received three samples in total. This allowed us to test the models against each other in a longer, more complex sequential belief-updating situation.

Method

Participants

Twenty-four undergraduate students ($Mdn_{\text{age}} = 23.50$ years, 79% women, 21% men) participated in Study 3. Participants were compensated with either course credit or book vouchers worth 15 CHF. Additionally, they received a performance-contingent bonus ($Mdn = 2.30$ CHF).

Materials

The experimental setup of Study 3 was identical to that of Study 2 with samples being presented sequentially. The only difference was that in each round, three samples were drawn and presented to the participants. As in Studies 1 and 2, participants were confronted with a diverse set of randomly ordered games containing all combinations of three underlying probabilities of 10%, 20%, 30%, 40%, 50%, 60%, 70%, and 80%. The sample frequencies for 75% of all trials were determined by drawing random samples of seven cards from a binary distribution with the underlying probability of the respective deck. For 25% of all trials, one sample was tweaked such that its likelihood of being drawn from deck *A* was identical to the likelihood of being drawn from deck *B*. A randomly determined quarter of the regular trials were repeated three times with different orders of the samples. All trials including nondiagnostic evidence were repeated three times with the nondiagnostic sample appearing first, second, or third.

Procedure

The procedure in Study 3 was identical to that of the previous studies and included 84 rounds plus five practice rounds. Samples were presented sequentially and a sample disappeared when a subsequent one was presented. Twenty-three of the 24 participants understood the task.

Results and Discussion

Table 1 shows participants' general performance, Table 2 the results for nondiagnostic evidence, and Table 4 the results for order effects and converging evidence across all four studies. Results for order effects and converging evidence are comparable to those in Studies 1 and 2.

Participants' General Performance

In Study 3 participants performed as well as in Studies 1 and 2 (Table 1). In 29% of trials, participants changed their mind between the first and the second sample or between the second and the third. The majority (59%) of these switches resulted in correct choices.

Reaction to Nondiagnostic Evidence

First Sample. The percentage of correctly identified nondiagnostic samples averaged around 89% (median; range: 0%–100%). If the nondiagnostic sample was neutral, participants correctly identified it in a median of 100% of the trials (IQR = 83%–100%). This performance was lower for nondiagnostic mixed samples (Mdn = 84%; IQR = 54%–100%). Similarly to the previous studies, we found confirmation effects in a median of 17% (IQR = 0%–33%) of trials after neutral samples and a median of 31% (IQR = 11%–51%) of trials after mixed samples in Study 3.

Second Sample. We analyzed all trials where either the second sample was nondiagnostic, participants did not change their choice between Samples 1 and 2, and their first probability judgment was ≠.50; or the third sample was nondiagnostic, participants did not change their choice between Samples 2 and 3, and their second probability judgment was ≠.50. This included 22% of all trials. Table 2 shows the percentage of these trials in which participants showed a dilution effect, a confirmation effect, or no effect. The slope of a simple linear regression between the first and the second probability judgment was .52 and the slope for a linear regression between the second and the third sample was .82, again illustrating the dilution effect. Sixty-seven percent of the participants showed the dilution effect in more than half of the analyzed trials between the first (second) and the second (third) sample. The median downward adjustment of the probabilities after the presentation of a nondiagnostic piece of information in these trials was 7% (IQR = 3%–14%). As in Studies 1 and 2, participants showed fewer dilution effects (46%) when their first probability judgment was lower than or equal to the median probability judgment over all trials and participants (66%) than when their first probability judgment was higher (69%). Similarly, between Samples 2 and 3, they showed fewer dilution effects (45%) when their second probability judgment was lower than or equal to the median probability judgment over all trials and participants (68%) than when their second probability judgment was higher (55%). Further, if the nondiagnostic sample was neutral, dilution effects were observed in 60% of the dilution trials, whereas in all other dilution trials, dilution effects were observed 58% of the time between Samples 1 and 2, and 71% and 50% of the time between Samples 2 and 3.Thus, we could replicate the dilution effect not only irrespective of whether the initial samples were present when the subsequent samples were shown but also irrespective of the number of samples presented.

Averaging and Adjustment

In Study 3 not all trials were presented in reverse order, thus the following analysis is based on 44% of the data. On average, 50% (median; IQR = 35%–56%) of all combined, second probability judgments lay between the two single, first probability judgments, 41% (median; IQR = 33%–49%) were larger, and 9% (median; IQR = 25%–37%) were smaller. The correlation of the similarity-based confirmation mechanism's predictions (based on estimated parameter values, cf. Table 3) and the difference between the two observed probability judgments within one trial was again positive ($r = .23$, $p < .01$; cf. Figure 3).

Quantitative Model Comparison

Parameter estimation and model selection were done as in the previous two studies, including all probability judgments per trial. The median BIC over all participants for the four models was −29 for the baseline model, −83 for the Bayesian model, −184 for the PT+N model, and −408 for the similarity-updating model. The median best fit parameter values of all models are listed in Table 3. Twenty-three of 24 participants were best described by the similarity-updating model according to model selection based on the BIC, and one was best described by the PT+N model. So the similarity-updating model is the best model overall and also describes the individual participants' probability judgments best.

Study 3 also tested the predictive power of the similarity-updating model regarding generalizations to independent judgments ignored for parameter estimation. To this end, we estimated the model's parameters on the basis of the first two judgments in each trial and predicted the third judgment. The median deviance over all participants was −15 for the Bayesian model, −40 for the PT+N model, and −97 for the similarity-updating model. According to model selection based on minimum deviance, the Bayesian model was chosen for three of 24 participants, the PT+N model also for three participants, and the similarity-updating model for 18 participants. These results show that the similarity-updating model not only accounts for the qualitative patterns found in all studies and describes the probability judgments well on an individual level, but also predicts repeated probability judgments well in comparison to other models.

Study 4: Varying Sample Size

Study 4's goal was to test whether sample size has an impact on probability judgments. The similarity-updating model, in contrast to the Bayesian or the PT+N model, predicts that sample size should not influence the probability judgments because single similarity judgments are agnostic to sample size. Study 4's setting was identical to Study 2's. Around two thirds of the nondiagnostic and half of the diagnostic trials were chosen randomly and the samples were doubled to directly compare pairs of trials with seven and 14 cards.

Method

Participants

Twenty-five students (Mdn_{age} = 25 years, 64% women, 36% men) at the University of Basel participated and were compensated with either course credit or 20 CHF. Additionally, they received a performance-contingent bonus with a maximum of 7.5 CHF.

Materials

Study 4 was based on the materials of Study 2. From Study 2 we randomly selected 60% of the nondiagnostic trials and 50% of the diagnostic trials. For each of the selected trials we tested one unchanged version (sample size of seven cards) and one version with doubled sample size (14 cards). As in the previous studies, all seven or 14 cards were shown on the screen, making the increased sample size explicit (cf., Appendix D). Around 5% of the decks with a sample size of 14 had to be changed because they inconsistently showed one or both decks with only 10 cards of a certain color but a sample with more than 10 cards with that color. In these tasks we switched two values in the decks so that the samples were consistent with the information given by the decks.

Procedure

The procedure in Study 4 was identical to that in Study 2. All participants understood the task.

Results and Discussion

Table 1 shows participants' general performance, Table 2 the results for nondiagnostic evidence, and Table 4 the results for order effects and converging evidence across all four studies. Participants showed similar results for order effects and converging evidence to those in Studies 1–3.

Participants' General Performance

Participants performed equally as well as in all previous studies (Table 1). In 20% of trials, participants changed their mind between the first and the second sample. The majority (66%) of these switches resulted in correct choices.

Differences in Sample Size

All cognitive models that follow probability theory and sampling ideas in a broader sense predict differences between observations with different sample sizes. Bayesian models predict (independently of the assumed error component) a more extreme probability estimate for the correct deck. Frequentist models predict a lower variance. The similarity-updating model predicts no differences between sample sizes.

We first tested the predictions of the similarity-updating model (no difference due to sample size) against the predictions of frequentist models (lower variance) by using a Bayesian *t*-test (Morey & Rouder, 2018). We found substantial evidence in favor of the null hypothesis, that there is no difference in the mean variance per participant between sample sizes, with a Bayes factor (*BF*) of $BF_0 = 3.55$ for the first probability judgments and $BF_0 = 3.56$ for the second probability judgments, as predicted by the similarity-updating model. The median standard deviation across participants for the first probability judgment was 12% (IQR = 10%–13%) for samples with seven cards and 12% (IQR = 9%–14%) for samples with 14. For the second probability judgment, the standard deviation across participants was 10% (IQR = 8%–13%) for samples with seven cards and 10% (IQR = 9%–12%) for samples with 14 cards.

Next, we tested the predictions of the similarity-updating model against the predictions of the Bayesian model, that there is a difference between the two sample sizes, using a Bayesian *t*-test on median judgments. The Bayesian *t*-test gave substantial evidence in favor of the null hypothesis, that there is no difference between sample sizes. The BF_0 for the first probability judgment was 3.56 and for the second probability judgment also 3.60. These results support the similarity-updating model and speak against the Bayesian model and the PT+N model. The median first probability judgment across participants was 73% (IQR = 60%–76%) for a sample size of seven cards and 73% (IQR = 64%–78%) for a sample size of 14 cards. The median second probability judgment across participants was 75% (IQR = 63%–82%) for a sample size of seven and 75% (IQR = 64%–85%) for a sample size of 14 cards. All these results clearly show that we did not observe any differences in probability judgments between the different sample sizes, which speaks in favor of the similarity-updating model and against the Bayesian model and the PT+N model.

Reaction to Nondiagnostic Evidence

First Sample. The percentage of correctly identified nondiagnostic first samples averaged around 89% (median; IQR: 67%–94%). If the nondiagnostic sample was a neutral sample, participants correctly identified it in a median of 100% of the trials (IQR = 75%–100%). This performance was lower for nondiagnostic mixed samples (*Mdn* = 83%; IQR = 33%–83%). We found confirmation effects in a median of 8% of the trials (IQR = 0%–17%) after neutral samples and a median of 0% (IQR = 0%–33%) after mixed samples.

Second Sample. In Study 4, 16% of trials met the criteria to be analyzed as second, nondiagnostic samples. Table 2 shows the percentage of these trials in which participants showed a dilution effect, a confirmation effect, or no effect. The slope of a simple linear regression was .72, illustrating the dilution effect. More than two thirds of the participants (68%) showed the dilution effect in more than half of these trials. Further, participants showed dilution effects in 82% of trials with neutral nondiagnostic second samples and in 75% of trials with mixed nondiagnostic second samples.

Averaging and Adjustment

On average, 57% (median; IQR = 52%–63%) of all combined, second probability judgments lay between the two single, first probability judgments, as predicted by pure averaging models. Eleven percent (median; IQR = 3%–17%) were smaller than both single probability judgments and 33% (median; IQR = 25%–37%) were larger. The correlation of the similarity-based confirmation mechanism's predictions (based on estimated parameter values, cf. Table 3) and the difference between the two observed probability judgments within one trial was again positive ($r = .20$, $p < .01$; cf. Figure 3).

Quantitative Model Comparison

In Study 4 we estimated all models on the basis of participants' probability judgments following a maximum likelihood estimation approach. The median BIC over all participants for the four models was −3 for the baseline model, −55 for the Bayesian model, −96 for the PT+N model, and −251 for the similarity-updating model, indicating that the latter provided the best model fit. The median optimal parameter values of all models are listed in Table 3. According to model selection based on the BIC, 23 of 25 participants were best described by the similarity updating model, one was best described by the Bayesian model, and one by the PT + N model.

As a second test of the predictive power of the models we estimated parameters for all trials with a sample size of 7 and predicted the probability judgments for the trials with a sample size of 14. The median deviances over all participants were −34 for the Bayesian model, −33 for the PT+N model, and −131 for the similarity-updating model. As in the model selection based on the BIC, 24 of 25 participants were best predicted by the similarity-updating model, and one by the PT+N model according to the minimum deviance.

To summarize the results of the four studies, we observed a stable dilution effect (in 60% or more of all considered trials, Table 2). When looking at all first and second probability judgments over all studies, the slope of a simple linear regression line was .58, illustrating, on average, the dilution effect. We also found a smaller proportion of confirmation effects, a finding that in principle can be explained by our similarity-updating model. Additional results, such as order effects and effects on converging evidence (Table 4), are stable across studies and speak in favor of the similarity-updating model. According to model selection based on the BIC, over 95% of all participants were best described by the similarity-updating model, 3% by the PT+N model, and 2% by the Bayesian model.

An additional analysis of predictive power reveals that the results of the model comparison are not due to overfitting.

General Discussion

Research has shown that when people judge probabilities, they do not always behave according to Bayesian theory (e.g., Bar-Hillel, 1980; Edwards, 1968; Jenny et al., 2014; Nisbett et al., 1981; Tversky & Kahneman, 1983; but see Chater et al., 2006; Gigerenzer & Hoffrage, 1995; Griffiths et al., 2010; Sanborn & Chater, 2016; Tenenbaum et al., 2006). Belief updating prescribes that beliefs and probability judgments should not be influenced by nondiagnostic information. In contrast, people's beliefs and judgments often change upon the presentation of nondiagnostic information. The dilution effect, a special case of this influence of nondiagnostic information on people's beliefs, has been observed in a plethora of studies in fields ranging from social reasoning to accounting. The opposite effect, a confirmation effect, has also been observed. Yet, few studies have provided thorough cognitive explanations as to why people show the dilution effect, and there is no single framework to explain why people sometimes show a dilution effect and in other situations show a confirmation effect.

Introducing and Testing the Similarity-Updating Model

To explain the cognitive processes behind people's belief-updating behavior we developed a new cognitive model, the similarity-updating model, which is inspired by models from the judgment and decision-making and categorization literature. One innovation of this theory is its synthesis of cognitive models from different fields of research: People's subjective probability judgments are modeled with a similarity process, which has been used in the categorization literature (e.g., Nosofsky & Johansen, 2000), and people's belief updating is modeled with a weighting-and-adding process, which has been used in judgment and decision-making research (e.g., Hogarth & Einhorn, 1992; Jenny et al., 2014; Juslin et al., 2009; Nilsson et al., 2009, 2013). Combined probability judgments are affected by a similarity-based confirmation process modeled after the principles underlying inductive confirmation (Tentori et al., 2013). This model not only bridges different fields of research but also is based on concepts that have already been validated in the respective fields of research.

The similarity-updating model is able to explain why people show the dilution effect most of the time but sometimes also show a confirmation effect. In the words of LaBella and Koehler (2004): "One possible modification of such models . . . is to allow the subjective evaluation of the implications of a piece of evidence to depend on what evidence has already been encountered" (p. 1086). We propose that this subjective evaluation is a by-product of the similarity-based process underlying the formation of constituent probability judgments formally described by our similarity-based confirmation mechanism. Importantly, the confirmation mechanism biases the formation of the combined probability, that is, the weighting-and-adding process, explaining why the dilution effect is observed more often than the confirmation effect.

Further, our model settles a debate started by Nisbett et al. (1981), in which they argued that similarity rather than averaging processes produce the dilution effect. In our view, it is the fact that people seem to follow a combination of both similarity and averaging processes when making their judgments that leads them to produce the dilution effect. The subtleties of this only become clear when the two processes are combined into one overarching model. The combination shows that although a similarity process can explain the individual judgments, the updating process of weighting and averaging eventually leads to the dilution effect. Our model also accounts for the criticism by Trueblood and Busemeyer (2011) that adding and averaging models cannot account for the strength of different pieces of evidence relative to each other, which our model does with the addition of the similarity-based confirmation mechanism.

The similarity-updating model predicts that the sample size should not affect people's judgments. The Bayesian model and the PT+N model predict that probability judgments become more extreme and less noisy with increasing sample size. The results of Study 4 confirm the prediction of the similarity-updating model. The idea that in probability judgments people give too little weight to sample sizes has a long tradition in psychology (Tversky & Kahneman, 1971) and has also been observed in more recent work (e.g., Hoffart, Olschewski, et al., 2019; Hoffart, Rieskamp, et al., 2019).

Although people's probability judgments differed considerably from the normative solution, they nevertheless correlated well with this solution and people were able to choose the correct deck in most trials. Consistently over all studies and participants, we observed dilution trials at a rate of approximately 60%. The greater participants' first probability judgments, the more dilution effects they produced. Overall, participants correctly identified a nondiagnostic first deck as such. Extrapolating from this and assuming that they also were quite well able to identify a second, nondiagnostic deck, it seems likely that it was people's belief-updating mechanism rather than the way they judged individual probabilities that led them to produce the dilution effect. On average, their second probability judgment was also smaller only when the second sample was nondiagnostic. This suggests that the effect is not a falsely qualified regression to the mean effect, which would be explained by the PT+N model.

We tested the similarity-updating model thoroughly within a cognitive modeling framework by estimating it on the basis of people's trial-by-trial probability judgments. An additional strength of our model test is that we examined the model's ability to predict people's behavior and compared it with a Bayesian model, the PT+N model, and a random baseline model. The models were tested against each other in four experimental studies, in which people provided subjective probability judgments and revised them in light of additional information in a card game. Additionally, we evaluated the models' generalization ability by fitting participants' responses only to a subset of the data and predicting responses for the excluded data.

The process behind people's first and subsequent probability judgments in this task was best described by the similarity-updating model as compared to a Bayesian model, the PT+N model, and a random baseline model. The similarity-updating model not only provided the best model fit over all participants but also described the individual behavior of almost all participants best. These results held irrespective of whether the first sample was still present or removed at the time of the second sample being presented, for updating situations in which participants could make use of an additional third sample, and also for different sample sizes. The generalization test in Studies 3 and 4 additionally showed that the superior model fit was not due to overfitting or excessive model flexibility.

Implications

Subjective Probability Judgment

Among other areas, similarity has previously informed research on quantitative estimations (e.g., Juslin et al., 2008; von Helversen et al., 2014; von Helversen & Rieskamp, 2009). In line with related findings (Nilsson et al., 2005; Read & Grushka-Cockayne, 2011), we have shown that introducing the concept of similarity to the study of subjective probability judgments provides important insights. Alternative models of similarity are the evidential support accumulation model (Koehler et al., 2003), cue-based relative frequency, and the probabilities from exemplars (PROBEX) model (Juslin & Persson, 2002). The first two models were not tested in the present studies because there is not a straightforward application of these models to probability judgment tasks. Further, applied to our task, which does not involve memory processes across trials, the PROBEX model boils down to the probability judgment part of the similarity-updating model with a probability judgment function, which differs slightly from Equation 3. This makes testing this model superfluous as long as no memory processes are involved.

Belief Updating

A weighting-and-adding process has been successfully applied to explain people's behavior in conjunctive probability estimation (Jenny et al., 2014; Nilsson et al., 2009, 2013) and they have been demonstrated to perform better than normative as well as alternative cognitive models. In the present article we have shown that they can also well describe people's belief-updating processes, which have previously been described with belief-updating models (Hogarth & Einhorn, 1992) and the sigma model (Juslin et al., 2008).

A potential alternative approach to modeling belief-updating is based on QPT (Busemeyer et al., 2011; Trueblood et al., 2017), and there is evidence suggesting that QPT explains several nonnormative behaviors. QPT can be seen as a generalized and relaxed version of Bayesian probability theory that explains nonnormative behaviors usually with a context or background that gives rise to certain mathematical representations. For example, order effects are explained by the idea that people adopt different perspectives when evidence is presented and a shift between contexts is a noncommutative operation (Trueblood & Busemeyer, 2011). However, the dilution effect, as usually tested with the bookbag-and-poker-chip task, does not include different contexts or backstories. As such, while interesting, specifying a model following QPT that explains the dilution effect is not straightforward and not in scope of the current work.

Similarity-Based Confirmation

Confirmation effects have a long-standing tradition in psychology and economics and have been shown to affect people's judgments and decision making in a variety of contexts (Jones & Sugden, 2001; Lord et al., 1979; Plous, 1991; Wason, 1968). However, computational cognitive models that intend to explain this effect are rare. Similarly, confirmation effects have rarely been investigated in the domain of probability judgments (but see LaBella & Koehler, 2004). We propose a similarity-based confirmation mechanism that is a part of the belief-updating process. This belief-updating mechanism can be seen as an instance of an anchoring-and-adjustment process (Chapman & Johnson, 2002; Epley & Gilovich, 2001, 2006; Hogarth & Einhorn, 1992; Tversky & Kahneman, 1974). Anchoring-and-adjustment models have recently been discussed as a resource-efficient way to combine the result of different cognitive functions (Albrecht et al., 2020; Lieder et al., 2018; Millroth et al., 2019). In our model, people average the probabilities of two sequentially presented pieces of evidence (Hogarth & Einhorn, 1992) but then adjust the probability as a result of a similarity bias. The more similar the second piece of evidence is to a favored hypothesis, the higher the adjustment and, thus, the resulting probability judgment.

Reduced Dilution by Expertise

It has been shown that the dilution effect decreases with expertise in several domains (e.g., Shelton, 1999; Smith et al., 1998–1999). According to the similarity-updating model, the magnitude of decrease after a second sample has been presented depends on, among other things, the perceived importance of the second sample. If this importance is zero, the probability judgment does not change when the second sample is presented. A decrease in the perceived importance thus implies a reduction in the observed dilution in probability judgments. For example, an experienced auditor or a judge might consider the importance of a new piece of evidence as rather small in comparison to a body of evidence already gathered.

Bridging the Gap Between Related Fields

In general, the similarity-updating model can be applied whenever the probability of an event has to be judged given multiple pieces of information. One such domain is causal reasoning (Trueblood & Pothos, 2014; Trueblood et al., 2017). After learning a causal structure (aspect *A*/*B* causes event *E*) participants were asked to judge how likely event *E* is to occur given the presence/absence of aspects *A* and *B*. The similarity-updating model can also account for some of the effects in causal reasoning. Order effects in causal reasoning as well as the "memoryless" effect (the probability that an aspect is present depends on only the most recent information; Trueblood & Pothos, 2014) can be explained by the recency component in the updating mechanism.

Another related domain is the literature on probability judgments of conjunctive events (e.g., Tversky & Kahneman, 1983). Here participants are asked to combine $p(A)$, that is, the probability of *A*, and $p(B)$ into a judgment of $p(A \text{ and } B)$ and the typical finding is that $p(A \text{ and } B)$ tends to fall between the two constituent probabilities, that is, if $p(A) < p(B)$, then the typical finding is that $p(A) < p(A \text{ and } B) < p(B)$. This judgment pattern is often referred to as the conjunction error/fallacy, as it violates the conjunction rule of probability theory. Notably, the similarity-updating model can predict the conjunction error with its averaging mechanism (for similar arguments see, Fantino et al., 1997; Nilsson et al., 2009, 2013) and through the adaptation of inductive confirmation (Tentori et al., 2013).

Bridging Cognitive Psychology and Judgment and Decision-Making Research

Applying the concept of similarity to the area of judgment and decision making via the similarity-updating model fits the movement of applying concepts from more basal processes such as perception to higher order processes such as similarity judgments. This so-called mindful judgment and decision-making research has led to a more detailed understanding of judgment and decision-making phenomena (Weber & Johnson, 2009), for example, by modeling recognition within the cognitive architecture ACT-R (Schooler & Hertwig, 2005), modeling confidence judgments with evidence-accumulation models (Pleskac & Busemeyer, 2010), and modeling inferential choices with evidence-accumulation models (Lee & Cummins, 2004).

Considering the similarity-updating model in the context of the dilution effect facilitated the realization that pieces of information that a researcher or experimenter uses because they are nondiagnostic will not necessarily be perceived and treated as nondiagnostic by the experiment participants. Although in our experiment nondiagnostic first samples were often perceived as such, nondiagnostic second samples still influenced the initial probability judgments based on first diagnostic samples. This was due to people's tendency to put considerable weight on the second piece of (potentially nondiagnostic) evidence. Thus, by looking at the first probability judgments based on nondiagnostic samples and by inspecting the weight parameter of the estimated similarity-updating model (τ), it is possible to distinguish whether the dilution effect is caused by distorted probability estimation or by the belief-updating mechanism. Additionally, we propose that the similarity between evidence and hypothesis influences how people aggregate sequential probability judgments.

Open Questions

Belief-updating can occur in an immediate, online fashion where one's belief changes within seconds. Alternatively, it can also occur much more slowly, over the course of hours, days, or longer periods of time. The structure of the typical dilution task allows one to consider belief updating formed in a rapid sequence. Updating a belief in such an online fashion is an important skill, as people often have to adapt accurately and quickly update their beliefs based on sequentially observed information and make good corresponding decisions in many professional situations. Thus, understanding how people form such belief updates crucially broadens the understanding of complex human cognition as a whole. This raises the question of whether the similarity-updating model also explains people's belief-updating well if the updating happens over a longer period of time. This is an important question to address in future research.

Conclusions

People's probability judgments in a belief-revision task in which they experience the occurrence rate of events through sampling can be better described by a similarity-updating model consisting of a similarity and a weighting-and-adding process than by a Bayesian model or other models based on probability theory. It seems that, on average, people show the dilution effect because when forming and updating their probability judgments, they use a rule that lets them integrate nondiagnostic information into the judgment via a weighting-and-adding process. However, they sometimes show a confirmation effect because a similarity-based confirmation process affects this weighting-and-adding process. These findings are in line with previous findings (e.g., Anderson, 1981; Hogarth & Einhorn, 1992; Nilsson et al., 2009; Shanteau, 1970) that a similarity and a weighting-and-adding process affect people's probability judgments. Although following a similarity-updating process leads people to take nondiagnostic information into account and produce the dilution effect, it still leads them to make generally good predictions and receive good decision outcomes. The similarity-updating model can describe the underlying cognitive process of people's probability judgments that often lead to accurate decisions despite violating probability theory.

References

Albrecht, R., Hoffmann, J. A., Pleskac, T. J., Rieskamp, J., & von Helversen, B. (2020). Competitive retrieval strategy causes multimodal response distributions in multiple-cue judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(6), 1064–1090. https://doi.org/10.1037/xlm0000772

Anderson, N. H. (1981). *Foundations of information integration theory*. Academic Press.

Anderson, N. H. (1996). *A functional theory of cognition*. Lawrence Erlbaum.

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Bergus, G. R., Chapman, G. B., Levy, B. T., Ely, J. W., & Oppliger, R. A. (1998). Clinical diagnosis and the order of information. *Medical Decision Making*, *18*, 412–417. https://doi.org/10.1177/0272989X9801800409

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

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*, 157–171. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<157:AID-BDM260>3.0.CO;2-#

Busemeyer, J. R., Pothos, E. M., Franco, R., & Trueblood, J. S. (2011). A quantum theoretical explanation for probability judgment errors. *Psychological Review*, *118*, 193–218. https://doi.org/10.1037/a0022542

Carnap, R. (1962). *Logical foundations of probability* (2nd ed.). University of Chicago Press.

Chapman, G. B., & Johnson, E. J. (2002). Incorporating the irrelevant: Anchors in judgments of belief and value. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 120–138). Cambridge University Press. https://doi.org/10.1017/CBO9780511808098.008

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287–291. https://doi.org/10.1016/j.tics.2006.05.007

Corner, A., Harris, A., & Hahn, U. (2010). Conservatism in belief revision and participant skepticism. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual meeting of the cognitive science society* (pp. 1625–1630). Cognitive Science Society.

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review*, *121*(3), 463–480. https://doi.org/10.1037/a0037010

Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106–133. https://doi.org/10.1016/j.cogpsych.2016.06.006

Crupi, V., & Tentori, K. (2010). Irrelevant conjunction: Statement and solution of a new paradox. *Philosophy of Science*, *77*, 1–13. https://doi.org/10.1086/650205

Dougherty, M. R. P., Gettys, C. F., & Ogden, E. E. (1999). MINERVA-DM: A memory processes model for judgments of likelihood. *Psychological Review*, *106*, 180–209. https://doi.org/10.1037/0033-295X.106.1.180

Dougherty, M. R. P., & Hunter, J. (2003). Probability judgment and subadditivity: The role of working memory capacity and constraining retrieval. *Memory & Cognition*, *31*, 968–982. https://doi.org/10.3758/BF03196449

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal representation of human judgment* (pp. 17–52). Wiley.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–396. https://doi.org/10.1111/1467-9280.00372

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*, 162–169. https://doi.org/10.1016/j.tics.2004.02.002

Fantino, E., Kulik, J., Stolarz-Fantino, S., & Wright, W. (1997). The conjunction fallacy: A test of the averaging hypotheses. *Psychonomic Bulletin & Review*, *4*, 96–101. https://doi.org/10.3758/BF03210779

Fitelson, B. (2006). Logical foundations of evidential support. *Philosophy of Science*, *73*, 500–512. https://doi.org/10.1086/518320

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. *The Journal of General Psychology*, *113*, 351–357. https://doi.org/10.1080/00221309.1986.9711045

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263. https://doi.org/10.2307/2841583

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Goldstone, R. L., & Son, J. Y. (2005). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *Handbook of thinking and reasoning* (pp. 13–36). Cambridge University Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Hackenbrack, K. (1992). Implications of seemingly irrelevant evidence in audit judgment. *Journal of Accounting Research*, *30*, 126–136. https://doi.org/10.2307/2491095

Hahn, U. (2014). Similarity. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(3), 271–280. https://doi.org/10.1002/wcs.1282

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539. https://doi.org/10.1111/j.0956-7976.2004.00715.x

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2005). The role of information sampling in risky choice. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 72–91). Cambridge University Press. https://doi.org/10.1017/CBO9780511614576.004

Hoffart, J. C., Olschewski, S., & Rieskamp, J. (2019). Reaching for the star ratings: A Bayesian-inspired account of how people use consumer ratings. *Journal of Economic Psychology*, *72*, 99–116. https://doi.org/10.1016/j.joep.2019.02.008

Hoffart, J. C., Rieskamp, J., & Dutilh, G. (2019). How environmental regularities affect people's information search in probability judgments from experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(2), 219–231. https://doi.org/10.1037/xlm0000572

Hogarth, R., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*, 1–55. https://doi.org/10.1016/0010-0285(92)90002-J

Hotaling, J. M., Cohen, A. L., Shiffrin, R. M., & Busemeyer, J. R. (2015). The dilution effect and information integration in perceptual decision making. *PLOS ONE*, *10*(9), Article e0138481. https://doi.org/10.1371/journal.pone.0138481

Jenny, M. A., Rieskamp, J., & Nilsson, H. (2014). Inferring conjunctive probabilities from noisy samples: Evidence for the configural weighted average model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*, 203–217. https://doi.org/10.1037/a0034261

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53–85. https://doi.org/10.1146/annurev.psych.60.110707.163633

Jones, M., & Sugden, R. (2001). Positive confirmation bias in the acquisition of information. *Theory and Decision*, *50*(1), 59–99. https://doi.org/10.1023/A:1005296023424

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*, 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Juslin, P., Nilsson, H., & Winman, A. (2009). Probability theory, not the very guide of life. *Psychological Review*, *116*, 856–874. https://doi.org/10.1037/a0016979

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making*, *10*, 189–209. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<189:AID-BDM258>3.0.CO;2-4

Juslin, P., & Persson, M. (2002). PROBabilities from Exemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science*, *26*, 563–607. https://doi.org/10.1207/s15516709cog2605_2

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

Koehler, D. J., White, C. M., & Grondin, R. (2003). An evidential support accumulation model of subjective probability. *Cognitive Psychology*, *46*, 152–197. https://doi.org/10.1016/S0010-0285(02)00515-7

LaBella, C., & Koehler, D. J. (2004). Dilution and confirmation of probability judgments based on nondiagnostic evidence. *Memory & Cognition*, *32*, 1076–1089. https://doi.org/10.3758/BF03196883

Lee, M. D., & Cummins, T. D. R. (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review*, *11*, 343–352. https://doi.org/10.3758/BF03196581

Lieder, F., Griffiths, T. L., Huys Q. J. M., & Goodman, N. D. (2018). Empirical evidence for resource-rational anchoring and adjustment. *Psychonomic Bulletin & Review*, *25*(2), 775–784. https://doi.org/10.3758/s13423-017-1288-6

Lopes, L. L. (1985). Averaging rules and adjustment processes in Bayesian inference. *Bulletin of the Psychonomic Society*, *23*, 509–512. https://doi.org/10.3758/BF03329868

Lopes, L. L. (1987). Procedural debiasing. *Acta Psychologica*, *64*, 167–185. https://doi.org/10.1016/0001-6918(87)90005-9

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis*. Wiley.

Macchi, L., Osherson, D., & Krantz, D. H. (1999). A note on superadditive probability judgment. *Psychological Review*, *106*, 210–214. https://doi.org/10.1037/0033-295X.106.1.210

Macrae, C. N., Shepherd, J. W., & Milne, A. B. (1992). The effects of source credibility on the dilution of stereotype-based judgments. *Personality and Social Psychology Bulletin*, *18*, 765–775. https://doi.org/10.1177/0146167292186013

McKenzie, C. R. M., Lee, S. M., & Chen, K. K. (2002). When negative evidence increases confidence: Change in belief after hearing two sides of a dispute. *Journal of Behavioral Decision Making*, *15*, 1–18. https://doi.org/10.1002/bdm.400

Meyvis, T., & Janiszewski, C. (2002). Consumers' beliefs about product benefits: The effect of obviously irrelevant product information.

The Journal of Consumer Research, *28*, 618–635. https://doi.org/10.1086/338205

Millroth, P., Guath, M., & Juslin, P. (2019). Memory and decision making: Effects of sequential presentation of probabilities and outcomes in risky prospects. *Journal of Experimental Psychology: General*, *148*(2), 304–324. https://doi.org/10.1037/xge0000438

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs* (R package version0.9.12-4.2) [Computer software]. https://CRAN.R-project.org/package=BayesFactor

Nilsson, H., Olsson, H., & Juslin, P. (2005). The cognitive substrate of subjective probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 600–620. https://doi.org/10.1037/0278-7393.31.4.600

Nilsson, H., Rieskamp, J., & Jenny, M. A. (2013). Exploring the overestimation of conjunctive probabilities. *Frontiers in Psychology*, *4*, 101. https://doi.org/10.3389/fpsyg.2013.00101

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*, 517–534. https://doi.org/10.1037/a0017351

Nisbett, R., Zukier, H., & Lemley, R. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, *13*, 248–277. https://doi.org/10.1016/0010-0285(81)90010-4

Nosofsky, R. M., & Johansen, M. K. (2000). Exemplar-Based accounts of "multiple-system" phenomena in perceptual categorization. *Psychonomic Bulletin & Review*, *7*, 375–402. https://doi.org/10.1007/BF03543066

Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., & Mertz, C. K. (2007). Less is more in presenting quality information to consumers. *Medical Care Research and Review : MCRR*, *64*, 169–190. https://doi.org/10.1177/10775587070640020301

Peters, E., & Rothbart, M. (2000). Typicality can create, eliminate, and reverse the dilution effect. *Personality and Social Psychology Bulletin*, *26*, 177–187. https://doi.org/10.1177/0146167200264005

Peterson, C. R., & Miller, A. J. (1965). Sensitivity of subjective probability revision. *Journal of Experimental Psychology*, *70*, 117–121. https://doi.org/10.1037/h0022023

Peterson, C. R., Schneider, R. J., & Miller, A. J. (1965). Sample size and the revision of subjective probabilities. *Journal of Experimental Psychology*, *69*, 522–527. https://doi.org/10.1037/h0021720

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*, 864–901. https://doi.org/10.1037/a0019737

Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology*, *21*(13), 1058–1082. https://doi.org/10.1111/j.1559-1816.1991.tb00459.x

Pothos, E. M., Barque-Duran, A., Yearsley, J. M., Trueblood, J. S., Busemeyer, J. R., & Hampton, J. A. (2015). Progress and current challenges with the quantum similarity model. *Frontiers in Psychology*, *6*, 205. https://doi.org/10.3389/fpsyg.2015.00205

Pothos, E. M., & Busemeyer, J. R. (2013). Can quantum probability provide a new direction for cognitive modeling? *Behavioral and Brain Sciences*, *36*, 255–274. https://doi.org/10.1017/S0140525X12001525

Pothos, E. M., Busemeyer, J. R., & Trueblood, J. S. (2013). A quantum geometric model of similarity. *Psychological Review*, *120*(3), 679–696. https://doi.org/10.1037/a0033142

Pothos, E. M., & Trueblood, J. S. (2015). Structured representations in a quantum probability model of similarity. *Journal of Mathematical Psychology*, *64–65*, 35–43. https://doi.org/10.1016/j.jmp.2014.12.001

Read, D., & Grushka-Cockayne, Y. (2011). The similarity heuristic. *Journal of Behavioral Decision Making*, *24*, 23–46. https://doi.org/10.1002/bdm.679

Roussel, J.-L., Fayol, M., & Barrouillet, P. (2002). Procedural vs. direct retrieval strategies in arithmetic: A comparison between additive and multiplicative problem solving. *The European Journal of Cognitive Psychology*, *14*, 61–104. https://doi.org/10.1080/09541440042000115

Russo, J. E. (2015). The predecisional distortion of information. In E. A. Wilhelms & V. F. Reyna (Eds.), *Neuroeconomics, judgment, and decision making* (pp. 91–110). Psychology Press.

Russo, J. E., Meloy, M. G., & Medvec, V. H. (1998). Predecisional distortion of product information. *JMR, Journal of Marketing Research*, *35*, 438–452. https://doi.org/10.1177/002224379803500403

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893. https://doi.org/10.1016/j.tics.2016.10.003

Sanborn, A. N., Noguchi, T., Tripp, J., & Stewart, N. (2020). A dilution effect without dilution: When missing evidence, not non-diagnostic evidence, is judged inaccurately. *Cognition*, *196*, Article 104110. https://doi.org/10.1016/j.cognition.2019.104110

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*, 610–628. https://doi.org/10.1037/0033-295X.112.3.610

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Shanteau, J. C. (1970). An additive model for sequential decision making. *Journal of Experimental Psychology*, *85*, 181–191. https://doi.org/10.1037/h0029552

Shanteau, J. C. (1972). Descriptive versus normative models of sequential inference judgment. *Journal of Experimental Psychology*, *93*, 63–68. https://doi.org/10.1037/h0032509

Shanteau, J. C. (1975). Averaging versus multiplying combination rules of inference judgment. *Acta Psychologica*, *39*, 83–89. https://doi.org/10.1016/0001-6918(75)90023-2

Shelton, S. (1999). The effect of experience on the use of irrelevant evidence in auditor judgment. *The Accounting Review*, *74*, 217–224. https://doi.org/10.2308/accr.1999.74.2.217

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323. https://doi.org/10.1126/science.3629243

Smith, H. D., Stasson, M. F., & Hawkes, W. G. (1998-1999). Dilution in legal decision making: Effect of non-diagnostic information in relation to amount of diagnostic evidence. *Current Psychology: A Journal for Diverse Perspectives on Diverse Psychological Issues*, *17*(4), 333–345. https://doi.org/10.1007/s12144-998-1015-6

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning*. MIT Press.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Tentori, K., Crupi, V., Bonini, N., & Osherson, D. (2007). Comparison of confirmation measures. *Cognition*, *103*, 107–119. https://doi.org/10.1016/j.cognition.2005.09.006

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General*, *142*, 235–255. https://doi.org/10.1037/a0028770

Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology*, *57*, 388–398. https://doi.org/10.1037/0022-3514.57.3.388

Tobler, P. N., O'doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Human neural learning depends on reward prediction errors in the blocking paradigm. *Journal of Neurophysiology*, *95*, 301–310. https://doi.org/10.1152/jn.00762.2005

Troutman, C. M., & Shanteau, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior and Human Performance*, *19*, 43–55. https://doi.org/10.1016/0030-5073(77)90053-8

Trueblood, J. S., & Busemeyer, J. R. (2011). A quantum probability account of order effects in inference. *Cognitive Science*, *35*, 1518–1552. https://doi.org/10.1111/j.1551-6709.2011.01197.x

Trueblood, J. S., & Pothos, E. M. (2014). A quantum probability approach to human causal reasoning. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the cognitive science society* (pp. 1616–1621). Cognitive Science Society.

Trueblood, J. S., Pothos, E. M., & Busemeyer, J. R. (2014). Quantum probability theory as a common framework for reasoning and similarity. *Frontiers in Psychology*, *5*, 322. https://doi.org/10.3389/fpsyg.2014.00322

Trueblood, J. S., Yearsley, J. M., & Pothos, E. M. (2017). A quantum probability framework for human probabilistic inference. *Journal of Experimental Psychology: General*, *146*(9), 1307–1341. https://doi.org/10.1037/xge0000326

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352. https://doi.org/10.1037/0033-295X.84.4.327

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*(2), 105–110. https://doi.org/10.1037/h0031322

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*, 547–567. https://doi.org/10.1037/0033-295X.101.4.547

von Helversen, B., Herzog, S. M., & Rieskamp, J. (2014). Haunted by a doppelgänger: Irrelevant facial similarity affects rule-based judgments. *Experimental Psychology*, *61*(1), 12–22. https://doi.org/10.1027/1618-3169/a000221

von Helversen, B., & Rieskamp, J. (2009). Models of quantitative estimations: Rule-based and exemplar-based processes compared. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 867–889. https://doi.org/10.1037/a0015501

Walker, L., Thibaut, J., & Andreoli, V. (1972). Order of presentation at trial. *The Yale Law Journal*, *82*, 216–226. https://doi.org/10.2307/795112

Waller, W. S., & Zimbelman, M. F. (2003). A cognitive footprint in archival data: Generalizing the dilution effect from laboratory to field settings. *Organizational Behavior and Human Decision Processes*, *91*, 254–268. https://doi.org/10.1016/S0749-5978(03)00024-4

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

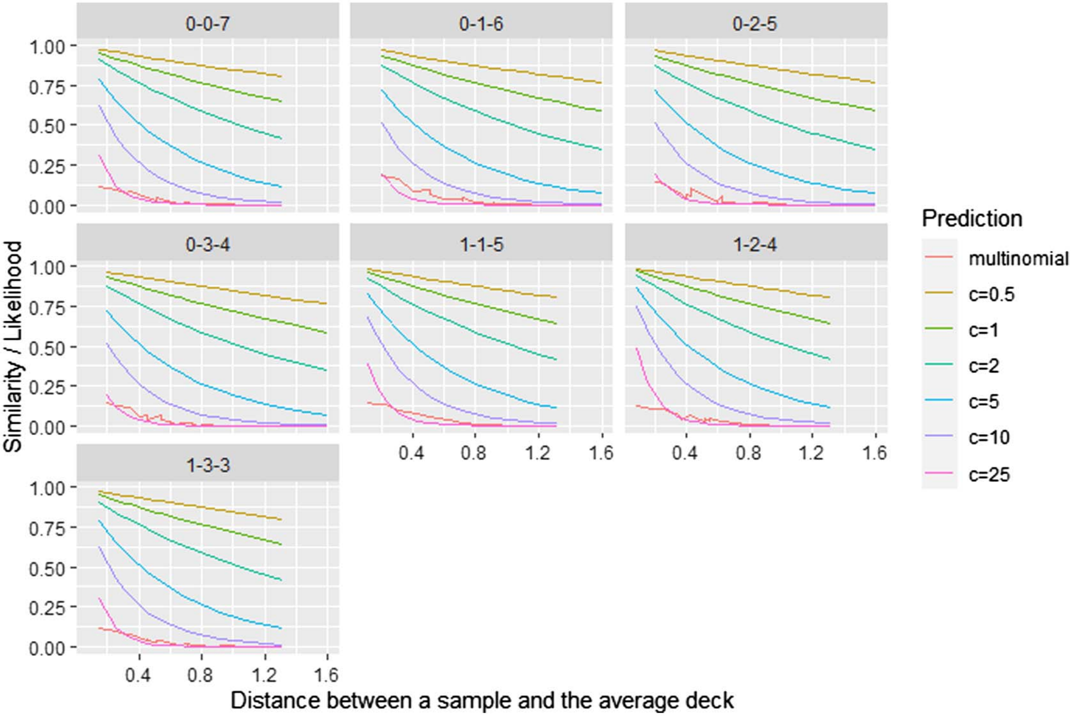
Appendix A

Comparison of Likelihood and Similarity

Figure A1 shows the similarity (for different values of the sensitivity parameter *c*) and likelihood for a sample relative to the city-block distance between different samples and all tested decks. Each grid cell shows the results for another, representative sample. Grid cell 0-0-7, for example, shows the results (similarity/likelihood) for a sample with 0 blue, 0 green and 7 red cards relative

Figure A1

Comparison of Likelihood and Similarity



Note. Likelihood is calculated with the multinomial distribution. Similarity is calculated with different values of parameter *c* relative to the distance between an average deck and different types of samples (grid). See the online article for the color version of this figure.

(*Appendices continue*)

to the distance between this sample and all possible tested decks. The presented sample types are representative for all samples, because city-block distance is symmetrical, meaning that the graphs for 0-0-7 and 0-7-0 are identical. The results show that both the likelihood and the similarity decrease with the distance between a sample and a deck. However, for small distances, the likelihood is much lower on average than similarity and consequently the decrease is flatter compared to similarity. Similarity can only approximate likelihood for some sample types and high values of the sensitivity parameter *c*.

Appendix B

Detailed Implementation of the Bayesian Model

The first step to calculate the probability of hypothesis *A* given a specific piece of evidence is to calculate the likelihood of observing the sample under hypothesis *A*. In our card game example, this is the likelihood of sampling the cards of sample E$_1$ out of deck *A*:

$$p(E_1|A) = p_{A1}^{f_{1.1}} \times p_{A2}^{f_{1.2}} \times p_{A3}^{f_{1.3}}. \qquad \text{(B1)}$$

where p_{A1}, p_{A2}, and p_{A3} are the probabilities of the different colors in deck *A* and $f_{1.1}$, $f_{1.2}$, and $f_{1.3}$ are the frequencies of the respective colors observed in sample E$_1$. The posterior probability of sample E$_1$ coming from deck *A* is then computed by

$$p(A|E_1) = \frac{p(E_1|A) \times p(A)}{p(E_1|A) \times p(A) + p(E_1|B) \times p(B)}, \qquad \text{(B2)}$$

where $p(E_1|A)$ and $p(E_1|B)$ are the likelihoods of receiving sample E$_1$ out of deck *A* and deck *B* and $p(A)$ and $p(B)$ are the prior probabilities of categories *A* and *B*, respectively. We implemented this model fixing the prior probabilities to .50, assuming that prior to having seen any data, participants would be indifferent about the categories. This posterior probability becomes a new prior probability in light of which additional information will be processed. According to Bayesian theory, this new prior probability is updated in light of new evidence as follows by first computing the likelihood of observing sample E$_2$ out of deck *A*:

$$p(E_2|A) = p_{A1}^{f_{2.1}} \times p_{A2}^{f_{2.2}} \times p_{A3}^{f_{2.3}}, \qquad \text{(B3)}$$

where p_{A1}, p_{A2}, and p_{A3} are the probabilities of the different colors in deck *A* and $f_{2.1}$, $f_{2.2}$, and $f_{2.3}$ are the frequencies of the respective colors observed in sample E$_2$. The posterior probability that both sample E$_1$ and sample E$_2$ come from deck *A* is then computed by

$$p(A|E_1, E_2) = \frac{p(A|E_1) \times p(E_2|A)}{p(A|E_1) \times p(E_2|A) + p(B|E_1) \times p(E_2|B)}, \qquad \text{(B4)}$$

where $p(A|E_1)$ and $p(B|E_1)$ are the posterior probabilities for decks *A* and *B* given that sample E$_1$ was observed (and thus the new prior probabilities), and $p(E_2|A)$ and $p(E_2|B)$ are the likelihoods of receiving sample E$_2$ out of deck *A* and *B*. Note that in this example, because the second sample is just as likely to come from deck *A* as from deck *B*, $p(A|E_1, E_2) = p(A|E_2)$. Equation B4 can be rearranged to an odds format where

$$\begin{aligned}\frac{p(A|E_1, E_2)}{p(B|E_1, E_2)} &= \frac{p(A)}{p(B)} \times \frac{p(E_1|A)}{p(E_1|B)} \times \frac{p(E_2|A)}{p(E_2|B)} \\ &= \frac{p(E_1|A)}{p(E_1|B)} \times \frac{p(E_2|A)}{p(E_2|B)}, \text{if} p(A) = p(B) = .50. \qquad \text{(B5)}\end{aligned}$$

Appendix C

Detailed Implementation of the PT+N Model

Our implementation of the PT+N model, in principle, follows the implementation of the Bayesian model introduced in Appendix B, with the exception that we assume that there is a chance that samples are perceived incorrectly, more specifically, a chance *d* that the color of a card is misperceived, leading to an incorrect count.

In our setting, the error chance *d* modulates how a sample of *n* cards is actually perceived. To model this, we first calculate the chance for every unordered sample of size *N* given the real sample E_1. For an arbitrary sample E_1 the probability of accidentally perceiving E_1 as $E_1^{'}$ given *d* is

$$\begin{aligned}P_{\text{perceive}}(E_1^{'}|E_1, d) = &\frac{N!}{\text{dist}(E_1, E_1^{'})! \cdot (n - \text{dist}(E_1, E_1^{'}))!} \\ &\cdot ((1-d)^{N-\text{dist}(E_1, E_1^{'})} \cdot d^{\text{dist}(E_1 E_1^{'})}). \qquad \text{(C1)}\end{aligned}$$

The distance (*dist*) between the two samples E_1 and $E_1^{'}$ is given by the number of different cards. The likelihood that a sample E_1 stems from deck *A* is calculated by multiplying the probability of perceiving every unordered sample, $P_{\text{perceive}}(E_1^{'}|E_1, d)$, with the probability that this perceived sample stems from the deck, $P(E_1^{'}|A)$,

$$P(E_1|A) = \sum_{E_1^{'}} P_{\text{perceive}}(E_1^{'}|E_1, d) \cdot P(E_1^{'}|A) \qquad \text{(C2)}$$

The probability that the misperceived sample $E_1^{'}$ stems from deck *A* is given by

$$p(E_1^{'}|A) = p_{A1}^{f_{1.1}} \times p_{A2}^{f_{1.2}} \times p_{A3}^{f_{1.3}} \qquad \text{(C3)}$$

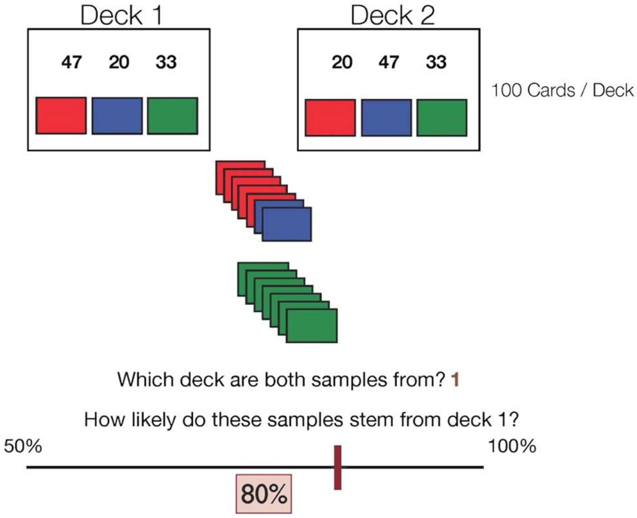
(*Appendices continue*)

The updating process is calculated analogously to the Bayesian cognitive model (described in Appendix B). Please note that the Bayesian cognitive model is nested in this implementation of the PT+N model ($d = 0$). The higher the error chance d, the higher are the deviations of the model's predictions from the predictions of the Bayesian cognitive model. However, the deviations are symmetrical and cannot in principle predict stimulus-dependent deviations such as the dilution effect.

Appendix D

Figure D1

Visual Presentation of Stimuli at the End of a Trial in Study 1



Which deck are both samples from? 1

How likely do these samples stem from deck 1?

Note. In Study 1, Sample 1 stayed visible when the second sample was presented. In all other studies, Sample 1 (and Sample 2, respectively in Study 3) was removed from the screen when a new sample appeared. Initially, the slider was not visible to avoid anchor effects. In Study 4, samples with a size of 14 showed 14 instead of seven cards. See the online article for the color version of this figure.

Appendix E

Instructions of the Four Studies

In the following, you will be presented with two decks of cards. Both decks consist of three types of card: red, blue, and green cards. Each deck consists of 100 cards. The numbers above the decks indicate how many red, blue, and green cards the decks contain. Proceed by hitting any key.

One of the decks will be randomly picked and two [three] samples will be sequentially drawn (with replacement) from this deck. The cards that are drawn in each sample are replaced before drawing additional samples. Thus, within one game, the decks always consist of the number of cards indicated above the decks. The composition of the decks will vary between games. Proceed by hitting any key.

Your task will be to indicate which deck you think the two [three] samples were drawn from. Both samples were drawn from the same deck. Additionally, you will assess the probability that the samples stem from the deck you picked. **PLEASE NOTE THAT THE PROBABILITY JUDGMENTS ALWAYS RELATE TO THE DECK THAT YOU PICKED.** Proceed by hitting any key.

You will receive 15 CHF (or 2 course credits) for your participation. In the end, one of the games you played will be picked at

(*Appendices continue*)

random and played out. If you picked the right deck in this game, then you receive 2.50 CHF in addition to the participation fee as a bonus. Proceed by hitting any key.

Additionally, you can win a bonus, which is contingent on the accuracy of your probability judgment. The better your judgment, the higher your bonus will be (max. 5 CHF). Thus, you will receive a bonus for your choice AND your probability judgment. Proceed by hitting any key.

Please address the instructor now if anything is unclear. Note that you will always first pick a deck by hitting either key “1” or “2” and then provide your probability judgment. **PLEASE NOTE THAT WITHIN ONE GAME, BOTH [ALL THREE] SAMPLES WILL BE DRAWN FROM THE SAME DECK**.[10] First, a couple of practice trials will follow, which will not count.[11] Proceed by hitting any key.

Are you ready for the real experiment? Now every game counts.[10] Please address the instructor if anything is unclear. If you are ready, then please hit any key.

Received September 15, 2018
Revision received April 1, 2021
Accepted April 2, 2021 ■

[10] This sentence was added to the instructions of Studies 2, 3 and 4 to make sure participants thoroughly understood the task.

[11] This sentence was used only in Studies 2, 3 and 4.