Journal of Behavioral Decision Making
J. Behav. Dec. Making, (2010)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/bdm.686

Probabilistic Reasoning in Prediction and Diagnosis: Effects of Problem Type, Response Mode, and Individual Differences

DOUGLAS H. WEDELL*
Department of Psychology, University of South Carolina, USA

ABSTRACT

In prediction, subset relations require that the probability of conjoined events is never higher than that of constituent events. However, people's judgments regularly violate this principle, producing conjunction errors. In diagnosis, the probability of a hypothesis normatively is often higher for conjoined cues. An online survey used a within-subjects design to explore the degree to which participants ($n = 347$) differentiated diagnosis and prediction using matched scenarios and both choice and estimation responses. Conjunctions were judged more probable than a constituent in diagnosis (76%) more often than prediction (64%) and in choice (84%) more often than direct estimation (57%), with no interaction of type of task and response mode. Correlation, regression, and path analyses were used to determine the relationships among individual difference variables and the diagnosis and prediction tasks. Among the correlation findings was that time spent on the task predicted higher conjunction probabilities in diagnosis but not prediction and that class inclusion errors predicted increased conjunction errors in choice but not estimation. Need for cognition and numeracy were only minimally related to reasoning about conjunctions. Results are consistent with the idea that people may misapply diagnostic reasoning to the prediction task and consequently commit the conjunction error. Copyright © 2010 John Wiley & Sons, Ltd.

KEY WORDS conjunction error; heuristics; probabilistic reasoning; Bayesian updating; individual differences

INTRODUCTION

People often make probability assessments based on conjoined events. For example, a person might wonder how likely is it to have a cold given one is sneezing and coughing? Or, conversely, how likely is it to be sneezing and coughing if one has a cold? On the face of it, these two questions are quite similar and might well prompt the same answer. Each involves a hypothesis, ''having a cold,'' and each involves conjoined events, ''sneezing and coughing.'' However, these two questions reflect opposite types of reasoning tasks.

* Correspondence to: Douglas H. Wedell, Department of Psychology, University of South Carolina, Columbia, SC 29208, USA.
E-mail: wedell@sc.edu

The first reflects diagnostic or backwards conditional reasoning, in which one is estimating the likelihood of a hypothesis given a conjunction of events, *p*("Has a cold" | "Sneezing" and "Coughing"). The second reflects predictive or forward conditional reasoning in which one is estimating the likelihood of a conjunction of events given a hypothesis, *p*("Sneezing" and "Coughing" | "Has a cold"). Although these terms have been used more generally, in this paper I will use the term diagnosis to refer to reasoning about hypotheses from outcomes and prediction to refer to reasoning about outcomes from a hypothesis.[1]

Both diagnosis and prediction are often required in daily life; however, people may well have difficulty distinguishing between them and hence may inappropriately apply reasoning processes associated with one task to the other. Extensive literatures have centered on each of these two types of probabilistic judgment tasks. Researchers have examined diagnostic probability judgments using a belief updating paradigm, with responses evaluated relative to the normative application of Bayes' theorem (Edwards, Lindman, & Savage, 1963; Fischhoff & Beyth-Marom, 1983). Although people typically update probability estimates in the direction consistent with the normative standard, updating is generally insufficient, referred to as conservatism (Phillips & Edwards, 1966). Updating sometimes inappropriately ignores base rate information (Kahneman & Tversky, 1973), and people may fail to properly integrate diagnostic and non-diagnostic information (Nisbett, Zukier, & Lemley, 1981; Tetlock & Boettger, 1989).

Research on predictive probability judgments has also demonstrated that participants can roughly follow normative patterns for integrating probabilistic information, such as multiplying probabilities for co-occurrences of independent events (Anderson & Shanteau, 1970). However, once again people often exhibit clear violations of normative predictive reasoning that include inappropriate covariation assessment (Chapman & Chapman, 1969; White, 2003), failure to appropriately consider sample size (Tversky & Kahneman, 1971), and inappropriate assessments of conjunctions of events (Tversky & Kahneman, 1983).

The research presented here explored the working hypothesis that conjunction errors in predictive judgment may derive from a tendency for people to inappropriately apply diagnostic reasoning when predictive reasoning is required. This hypothesis has received partial support in prior research (Wolford, Taylor, & Beck, 1990). One rationale for this hypothesis is that the way in which people naturally experience the world is geared toward the use of diagnostic judgment rather than predictive judgment. In diagnosis, people experience conditions in the world and seek to explain them. For example, one experiences sneezing and coughing and looks for their cause (Is it a cold, an allergy, or dust in the air?). From an evolutionary perspective, diagnostic reasoning may be more primary: When one encounters a configuration of conditions associated with a dangerous situation, it is important to derive the cause and hence be alerted to an appropriate response. If the diagnostic mode is more fundamental, then it may be that people automatically tend to apply this mode of thinking when considering forward reasoning or predictive judgments. To examine this possibility, a primary manipulation in the current study was to present the same scenarios in both a predictive and a diagnostic context and determine whether people treat these differently. Strong evidence for confusion of the two would occur if responses were indistinguishable for the two decision contexts. On the other hand, one might expect some participants to distinguish between these while others may not. In this case, a key question becomes the degree to which individual differences in processing of these two types of judgments can be delineated. Consequently, the current study included measures of individual differences and related these to the two types of judgments.

The basic idea explored in the current study is that conjunction errors may arise because people essentially commit a conversion error, judging the probability of the hypotheses given the conjunction of events, $\mathrm{p}(\mathrm{H} \mid \mathrm{E}_1 \cap \mathrm{E}_2)$, a diagnostic judgment, instead of the probability of the conjunction of events given the

[1] While diagnosis is typically used in the literature to reflect backward conditional reasoning, prediction tends to be a more broadly used umbrella term that can apply to both forward and backward reasoning. Thus, for example, in their classic article on the psychology of prediction, Kahneman and Tversky (1973) apply the term prediction to Bayesian updating problems as well as forward conditional reasoning problems. The exclusionary use of these terms in the present article is for expository purposes only.

DOI: 10.1002/bdm

hypothesis, $p(E_1 \cap E_2 \mid H)$, a predictive judgment. For example, instead of judging how likely it is that Linda is a bank teller and active in the feminist movement given the prior description of Linda, people judge how likely the prior description of Linda is given that she is a bank teller and active in the feminist movement. It should be noted that this explanation would only apply to conjunction errors in which a model or hypothesis for generating events is explicitly stated, as in the famous Linda problem. Indeed Tversky and Kahneman (1983) distinguished two paradigms for generating the conjunction error, which they labeled the M→A paradigm and the A→B paradigm (where M stands for the model of the situation, and A and B are events linked to that model). The former involves explicitly stating a model, such as the description of Linda, and then pairing a conjunction of events in which a base event is not likely given the model but the added event is. Tversky and Kahneman argued that people use a representativeness heuristic in which they judge the similarity of the events to the model in order to generate probabilities, with the added likely event incrementing similarity and hence probability judgments.

The A→B paradigm does not explicitly state a model and hence may not be easily formulated as a conditional probability that can be stated in both a forward predictive form and a backward diagnostic form. For example, Tversky and Kahneman (1983) found forecasters judged the likelihood of an earthquake in California causing a massive flood that kills over 1000 people in 1983 as greater than the likelihood of a massive flood in North America that kills over 1000 people in 1983. They argued that this type of conjunction error arises because the added event (an earthquake in this case) makes it easier to imagine the consequent event and hence leads to a higher probability assessment. Several researchers have demonstrated the robustness of this type of conjunction error that cannot easily be represented as a conditional probability (Sides, Osherson, Bonini, & Viale, 2002). The current research is based on the M→A model rather than the A→B model and hence is not designed to explain the latter type of conjunction errors.

Before describing experimental procedures in greater detail, I will consider more carefully how predictive and diagnostic situations differ. To do so I will present the typical paradigm used to assess conjunction errors and describe how it might be altered to create a diagnostic judgment paradigm. I then describe individual difference measures collected in the current study and how they might relate to task performance.

PREDICTIVE AND DIAGNOSTIC EVALUATIONS OF CONJUNCTIONS

In their classic work on the conjunction error in predictive judgment, Tversky and Kahneman (1983) demonstrated that when predicting events, people often indicate that the conjunction of two events is more likely than that of one of the constituent events, a violation of the extensional properties of these events. Although this classic effect has been demonstrated in numerous studies (Bar-Hillel & Neter, 1993; Tentori, Bonini, & Osherson, 2004; Wedell & Moro, 2008), one may question the applicability of the effect to real world situations. One reason is because the type of judgment being examined may not be typical of judgments made in the real world. An example of this mismatch is reflected in one of the medical reasoning studies described by Tversky and Kahneman (1983, pp. 301–302) They tested over 100 physicians with five medical scenarios that provided specific patient treatment information and asked the physicians to rank order the likelihood of resulting outcomes from the treatments. Averaging across the five scenarios, they found 91% of physicians committed the conjunction error, ranking the probability of the conjunction of two outcomes (e.g., dyspnea and hemiparesis) more probable than one of the constituent outcome (e.g., hemiparesis) given the preceding event (a pulmonary embolism 10 days after a cholecystechomy).

Why did physicians within their own area of expertise commit the conjunction error? One possibility is that this type of task is simply not representative of tasks in which physicians usually engage. In their practice, physicians are much more likely to have to diagnose a disease given presenting symptoms, i.e., determine a hypothesized cause for the observed effects. Thus, the forward predictive reasoning task presented by Tversky and Kahneman was different in kind from the backward diagnostic reasoning task in which

DOI: 10.1002/bdm

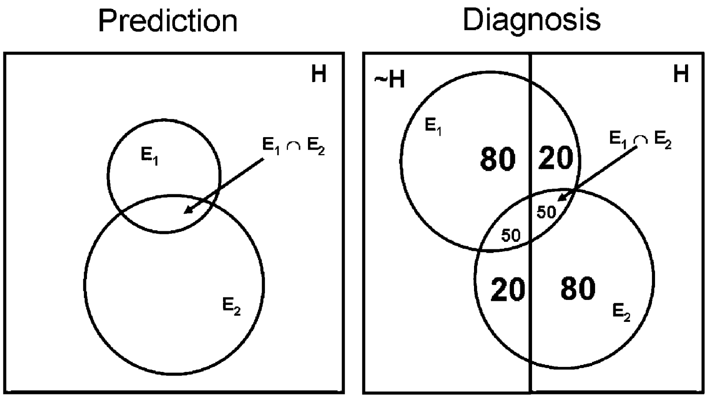

Figure 1. Diagrammatic illustration of the difference between forward conditional reasoning used in prediction (left panel) and backwards conditional reasoning used in diagnosis (right panel). In prediction, subset relations make the conjunction of events less probable than constituent events. In diagnosis, likelihood of a hypothesis may be increased by conjoining information. Numbers in the right panel reflect relative percentages of the area overlapping with the two hypotheses

physicians are highly practiced. In diagnostic reasoning, Bayes' theorem applies. As such, the conjunction of an unlikely symptom with a likely symptom can increase the probability of a hypothesis relative to the unlikely symptom alone, and hence the conjunction rule does not apply.

Figure 1 provides a schematic illustration that distinguishes between the two types of reasoning situations. A key difference is that in prediction one is considering the likelihood of events given a single hypothesis, whereas in diagnosis one is considering the likelihood of two or more hypotheses given the occurrence of events. The left panel illustrates the predictive situation in which the hypothesis dictates how the subsequent events might be sampled. Returning to our original example, the hypothesis implies random sampling of individuals from the population of those currently having a cold. Event E_1 then might reflect the probability of sampling those in the population who are sneezing and event E_2 might reflect the probability of sampling those in the population who are coughing, with the intersection of coughing and sneezing being represented as the overlap between these two sets. The intersection probability in this case may be represented as follows:

$$p(E_1 \cap E_2|H) = \frac{p(E_1 \cap H)\, p(E_2|E_1 \cap H)}{p(H)} \tag{1}$$

This multiplicative expression highlights the idea that the intersection probability must be less than or equal to the constituent probability. The expression, $p(E_1 \cap H)$, represents the area shown encircled by E_1 in the left panel of Figure 1. This area is then multiplied by the expression, $p(E_2 \mid E_1 \cap H)$, which reflects the proportion of the E_1 enclosed area that overlaps with E_2. Since this expression must be less than or equal to 1.0, the intersection probability cannot exceed the probability of event 1. Thus, one should never judge the conjunction to be more probable than one of the constituent events.

The right panel of Figure 1 illustrates the diagnostic reasoning situation in which one evaluates a hypothesis against its complementary hypothesis. Using the same example, H represents the hypothesis of having a cold and ~H represents the hypothesis of not having a cold. In considering the probability of H, one uses Bayes' theorem:

$$P(H|E_1 \cap E_2) = \frac{P(H)\, P(E_1 \cap E_2|H)}{P(H)\, P(E_1 \cap E_2|H) + P(\sim H)\, P(E_1 \cap E_2| \sim H)} \tag{2}$$

DOI: 10.1002/bdm

The numbers shown in Figure 1 indicate the percentage of samples of an event linked to either H or ~H. So for example, E_1 favors ~H over H by a ratio of 80 to 20. Conversely, E_2 favors H over ~H by a ratio of 80 to 20. In the example shown here, the two events are assumed to be independent so that the intersection likelihood ratio is the product of these two ratios and can be expressed as 1:1 or 50:50. We may link E_1 in the right panel of Figure 1 to the symptom "sneezing" and assume that it is diagnostic of not having a cold, $p(\mathrm{H} \mid \mathrm{E}_1) = .20$. We can further link E_2 to the symptom of "coughing" and assume that it is diagnostic of having a cold, $p(\mathrm{H} \mid \mathrm{E}_2) = .80$. Then from this analysis it is clear that the conjunction of events can lead to a higher probability of the hypothesis than found from one of the constituent events, $p(\mathrm{H} \mid \mathrm{E}_1 \cap \mathrm{E}_2) = .50 > p(\mathrm{H} \mid \mathrm{E}_1) = .20$.

One procedural area of concern explored in the current study is the effect of mode of responding. At the outset of the study of conjunction errors, Tversky and Kahneman (1983) noted that while generally robust to various manipulations, the conjunction error was sometimes strongly reduced by having judges estimate probability or frequency values for each event rather than rank ordering events or simply choosing which event was most likely. Subsequently, researchers have demonstrated the reductions in conjunction errors by switching from choice or ranking procedures to frequency estimates (Hertwig & Gigerenzer, 1999). To unconfound the effects of response focus (frequency versus probability) and response mode (choice versus estimation), Wedell and Moro (2008) factorially combined them in a within-subjects design. They found that engaging in estimation significantly reduced conjunction errors as compared with simply choosing the most likely alternative. However, conjunction errors did not depend on whether participants were considering frequencies or probabilities. For example, conjunction errors were just as likely when events were worded in terms of the number of people in a sample of 100 who fit the description as when they were worded in terms of likelihoods for a given individual. Moreover, the response mode effect was greater for scenarios that combined a likely base event with a likely conjoined event than for scenarios that combined an unlikely base event with a likely conjoined event.

The upshot from this study and others (Hertwig & Chase, 1998; Sloman, Over, Slovak, & Stibel, 2003) is that engaging in the process of making estimates for each option can improve the quality of predictive probabilistic reasoning. This basic result is consistent with the idea that when forced to generate estimates, participants may move away from heuristic processing and tend to engage in more algorithmic or computational processing that may tap important elements of the problem structure. Given the potential ameliorating effects of response mode on errors in probabilistic reasoning, both response modes (choice and estimation) were included in the present study. This manipulation poses the interesting question of whether engaging in direct estimation will have a similarly helpful effect in diagnosis as it does in prediction.

The current study examined these issues by building on the design and materials of Wedell and Moro (2008) in two fundamental ways. First, a within-subjects manipulation of task variables was used to increase power and provide a direct comparison at the individual level. Second, the predictive reasoning problems were constructed in a manner parallel to those used by Wedell and Moro. Three options were presented for each problem. One option described the base event alone (B). A second option described the conjunction of the base event and a likely event (B&L). A third option described the conjunction of the base event and the complement of the likely event (B&~L). This approach was based on prior work by Tentori et al. (2004) and was designed to reduce the chances of participants misinterpreting the base event as implying a conjunction with the negation of the added events. Consistent with this idea, Wedell and Moro reported that participants' interpretations of the base event reflected a marked reduction of misinterpretations compared with when the third option was not included.

One problem the current research used was the Scandinavian reasoning problem, developed by Tentori et al. (2004). This problem was also used by Wedell and Moro (2008) and thus serves as a comparison to prior results. The predictive choice version of the problem is presented below.

> The Scandinavian Peninsula is the European area with the greatest percentage of people with blond hair and blue eyes. This is the case even though every possible combination of hair color and eye color occurs

DOI: 10.1002/bdm

> in those countries. Suppose we choose at random an individual from the Scandinavian population. Which event do you think is most probable?
>
> (B) The individual has blond hair
>
> (B&L) The individual has blond hair and blue eyes
>
> (B&~L) The individual has blond hair and does not have blue eyes

Averaging across the results of two experiments (and probability versus frequency focus), Wedell and Moro reported that 63% of participants committed the conjunction error when choosing among these or similar options. However, when the problem required participants to estimate probabilities or frequencies (out of a sample of 100), only 18% committed the conjunction error. Thus, for this problem, which conjoins two likely events, estimation led to a very large reduction in conjunction errors.

Now consider a diagnostic or backward reasoning version of this task. To make this version concrete, it refers to three people, persons A, B and C. The object is then to pick which of the three is most likely to be Scandinavian based on the description. The wording for the choice form of this problem is presented below.

> The Scandinavian Peninsula is the European area with the greatest percentage of people with blond hair and blue eyes. This is the case even though every possible combination of hair color and eye color occurs in those countries. Suppose you are traveling through Europe and you meet three people. Which of these people do you think is most likely from the Scandinavian Peninsula?
>
> (B) Person A who has blond hair.
>
> (B&L) Person B who has blond hair and blue eyes.
>
> (B&~L) Person C who has blond hair and does not have blue eyes.

In this diagnostic version, the descriptions provide incremental evidence regarding the hypothesis. If characteristic L is diagnostic of being from the Scandinavian Peninsula over and above information of characteristic B, the conjunction (B&L) should be selected as most likely.

A numerical illustration of the differences between the predictive and diagnostic versions of the Scandinavian example is provided in Table 1 along with hypothetical values. In the predictive task, the base

Table 1. Predictive and diagnostic reasoning for Scandinavian problem

Probabilities assumed in this example		
$p(\mathrm{S}) = .10$	$p(\sim\mathrm{S}) = .90$	
$p(\mathrm{BH} \mid \mathrm{S}) = .60$	$p(\mathrm{BE} \mid \mathrm{S}) = .60$	$p(\mathrm{BH\ and\ BE} \mid \mathrm{S}) = .42$
$p(\mathrm{BH} \mid \sim\mathrm{S}) = .30$	$p(\mathrm{BE} \mid \sim\mathrm{S}) = .20$	$p(\mathrm{BH\ and\ BE} \mid \sim\mathrm{S}) = .08$
Predictive reasoning		
Probability of random sampling a member of the Scandinavian population with		
.60 blond hair.		
.42 blond hair and blue eyes		
.18 blond hair and not blue eyes		
Diagnostic reasoning		
Probability that a randomly sampled person is Scandinavian given he or she has		
.182 blond hair.		
.368 blond hair and blue eyes		
.083 blond hair and not blue eyes		

Note p = probability, ~ = Not, | = given, S = Scandinavian, BH = Blond Hair, and BE = Blue Eyes.

DOI: 10.1002/bdm

event probability is estimated by considering what proportion of the Scandinavian population may be blond, p(“Blond” | “Scandinavian”) = .60. To estimate the conjunction of blond hair and blue eyes, one might consider the proportion of blond Scandinavians who also have blue eyes and modify the base probability by this proportion. So if one imagined 70% of blond headed Scandinavians were also blue-eyed, then *p*(“Blond” and “Blue-eyed” | “Scandinavian”) = (.70)(.60) = .42. The probability of the base event with the complement of the added event can be calculated through subtraction: *p*(“Blond” and “Not Blue-eyed”) = *P*(“Blond” | “Scandinavian”)−*p*(“Blond” and “Blue-eyed” | “Scandinavian”) = .60–.42 = .18. Although the prediction version of the problem seems straightforward, it requires that one views the problem in terms of subset relationships. This may be more easily accomplished when one is forced to determine probabilities for each of the three options.

The diagnostic reasoning task would appear more daunting than the prediction task, as Bayes’ theorem indicates one must take into account both hypothesis base rates and implications of the evidence under the different hypotheses. To do so, one must consider the contrast category “Not Scandinavian” and how the relevant attributes and their conjunctions operate for this category. Table 1 presents the resulting probability estimates derived by applying Bayes’ theorem to the component estimates shown in the top of the table. For example, the probability that a person is Scandinavian given he or she has blond hair is calculated as follows: $p(\mathrm{S} \mid \mathrm{BH}) = (.10)(.60)/[(.10)(.60) + (.90)(.30)] = .18$. A basic assumption is that blond hair, blue eyes, and their combination are far less likely in the non-Scandinavian population. As shown in Table 1, the conjoined events of blond hair and blue eyes roughly doubles the likelihood of being Scandinavian compared to the base information alone (blond hair), because the additional information is quite diagnostic.

In predictive problems, it is clear that the probability of either conjunction cannot exceed the probability of the base event, $\mathrm{Max}[p(\mathrm{B}\ \&\ \mathrm{L}), p(\mathrm{B}\ \&\ {\sim}\mathrm{L})] \leq p(\mathrm{B})$. What pattern should be expected for the diagnostic task? Because both the intersection with an event and with its complement are options, the complementary pattern should be expected for the diagnostic task as long as $\mathrm{p}(\mathrm{H} \mid \mathrm{L}) \neq .50$ and events L and B are independent, $\mathrm{Max}[p(\mathrm{H} \mid \mathrm{B}\ \&\ \mathrm{L}), p(\mathrm{H} \mid \mathrm{B}\ \&\ {\sim}\mathrm{L})] > p(\mathrm{H} \mid \mathrm{B})$. Thus, in the current study, selecting the conjunction as more probable is a reasoning error in the prediction situation, but it is the reasonable and normatively expected behavior in the diagnosis situation.

SCENARIOS USED IN THE PRESENT STUDY

In addition to the Scandinavian problem described above, five other scenarios were used in the current study. These five scenarios are presented in Table 2. One of these, the IRS problem, was similar to the Scandinavian problem in that it was not related to issues of security against terrorism. However, the Scandinavian and IRS problems were constructed differently, reflecting likely and unlikely conjunction problems, respectively (Wells, 1985). The IRS problem was the unlikely conjunction and followed the basic recipe described by Tversky and Kahneman (1983) of presenting an unlikely base event (e.g., earning more than $300 000) that would have low similarity to the model and hence incrementing similarity by adding a likely event (e.g., paying less than $30 000 in taxes). In the likely conjunction case of the Scandinavian problem, the base event (e.g., blond hair) is highly probable given the hypothesis and one is simply adding another likely event. Conjunction errors are expected to be less numerous with likely conjunctions if similarity matching of the description to the model is the basis for the conjunction error. Wedell and Moro (2008) confirmed this prediction and also demonstrated that this decrement was greater for estimation than for choice. These two scenarios were included, in part, to determine whether previous results replicate.

The other four scenarios presented in the current study were concerned with security issues and thwarting terrorism. These were developed in a pilot study and used because of their relevance to current affairs as well as to provide a preliminary examination of whether people’s reasoning is better or worse when the scenarios are security based and involve potentially disastrous consequences. One possibility is that when important

DOI: 10.1002/bdm

Table 2. Prediction and diagnosis problems used in current study

Name/type	Problem	Options
IRS Prediction	Which rule, if followed by the internal revenue service, is most likely to result in auditing persons who have cheated the IRS of $10 000 or more?	__ Audit all those who have gross incomes greater than $300 000 __ Audit all those who have gross incomes greater than $300 000 and paid less than $30 000 in taxes __ Audit all those who have gross incomes greater than $300 000 and paid more than $30 000 in taxes
IRS Diagnosis	The internal revenue service is interested in auditing persons who have cheated the IRS. Which of the following persons is most likely to have cheated the IRS of $10 000 or more?	__ Person A who has a gross income greater than $300 000 __ Person B who has a gross income greater than $300 000 and paid less than $30 000 in taxes __ Person C who has a gross income greater than $300 000 and paid more than $30 000 in taxes
Airport Prediction	Imagine you are a security official at an airport and there is a high priority alert (red). Which rule, if followed by security personnel at airports, is most likely to result in detaining real terrorists?	__ Detain all persons entering from Iraq __ Detain all persons entering from Iraq and who are on the Homeland Security watch list __ Detain all persons entering from Iraq and who are not on the Homeland Security watch list
Airport Diagnosis	Imagine you are a security official at an airport and there is a high priority alert (red). There are three people you are considering detaining as suspected terrorists. Based on the information below, which do you think is most likely to be a real terrorist?	__ Person A who is entering from Iraq __ Person B who is entering from Iraq and is on the Homeland Security watch list __ Person C who is entering from Iraq and is not on the Homeland Security watch list
Station Prediction	As a security guard at a crowded station, you are told that there is a high level threat today with a strong likelihood that a terrorist organization is planning a bomb attack today. You are looking for a bomb-carrying terrorist. Which is more likely to be true about this terrorist?	__ He is wearing very bulky clothing __ He is wearing very bulky clothing and is nervously looking around while sweating profusely __ He is wearing very bulky clothing but is not nervously looking around nor sweating profusely
Station Diagnosis	As a security guard at a crowded station, you are told that there is a high level threat today with a strong likelihood that a terrorist organization is planning a bomb attack today. You see three people hanging out just outside the security gate. Which of these would you consider most likely to be a terrorist?	__ Person A who is wearing very bulky clothing __ Person B who is wearing very bulky clothing and is nervously looking around while sweating profusely __ Person C who is wearing very bulky clothing but is not nervously looking around nor sweating profusely
Paris Prediction	A warning goes out that a bomb is likely to be set off in Paris tomorrow. Which is more probable?	__ A person who is not of the Islamic faith will attempt to blow up the central train station in Paris __ A person who is not of the Islamic faith but who is sympathetic to the Arab cause will attempt to blow up the central train station in Paris __A person who is not of the Islamic faith and who is not sympathetic to the Arab cause will attempt to blow up the central train station in Paris

(*Continues*)

Table 2. (Continued)

Name/type	Problem	Options
Paris Diagnosis	The international police (Interpol) suspects a bomb will be set off by terrorists in the central train station of Paris. They currently are watching three suspects. Based on the information given below, which of these seems most likely to be involved in the plot	__ Person A who is not of the Islamic faith __ Person B who is not of the Islamic faith but who is sympathetic to the Arab cause __ Person C who is not of the Islamic faith and who is not sympathetic to the Arab cause
New York Prediction	As the anniversary of 9–11 approaches, Homeland security officials suspect that terrorists may target a prominent location, such as Madison Square Garden. They suspect that al Qaeda is involved. Recent surveillance has intercepted numerous suspicious phone calls from Pakistan that include known code words. Based on this intelligence, indicate which of the following is most likely	__ That a Pakistani terrorist group is involved in the plot __ That a Pakistani terrorist group with ties to al Qaeda is involved in the plot __ That a Pakistani terrorist group with no ties to al Qaeda is involved in the plot
New York Diagnosis	As the anniversary of 9–11 approaches, Homeland security officials suspect that terrorists may target a prominent location, such as Madison Square Garden. They suspect that al Quaeda is involved. Recent surveillance has intercepted numerous suspicious phone calls from Pakistan that include known code words. They are following three suspected terrorist cells that may be involved in the plot. Which cell seems most likely to be involved	__ Cell A which consists of Pakistani terrorists __ Cell B which consists of Pakistani terrorists who have ties to al Qaeda __ Cell C which consists of Pakistani terrorists who do not have ties to al Qaeda

Note: All problems shown in choice mode; estimation mode asks for a probability estimate of each option; top option is the base event; second option is the base and likely events; third option is the base and unlikely events.

consequences such as these are considered, people will take the task more seriously and perhaps demonstrate better reasoning performance. On the other hand, emotional reactions to considering the consequences of terrorist attacks might have the opposite effect and push people to rely further on heuristic processing, hence exacerbating reasoning errors. The current set of scenarios provides the opportunity to make an initial comparison of reasoning with terrorist-related and terrorist-unrelated problems. The terrorism-related scenarios were formulated as unlikely conjuncts, and therefore were expected to show a high likelihood of conjunction errors.[2]

INDIVIDUAL DIFFERENCE MEASURES OF INTEREST

It is reasonable to expect that people may differ in their tendencies to commit the conjunction error or to utilize diagnostic information. Some of these differences may simply be due to motivation to perform the task

[2] Classification of conjuncts as unlikely or likely depends on the model generated by the judge. For problems in which relevant model information is simply provided, such as in the Scandinavian problem, this classification is unambiguous (e.g., the model is that blue eyes and blond hair are highly likely). In other cases, such as the classic Linda problem, the participants must use their world knowledge to construct the model, and hence classification as unlikely or likely is more complex. The security related scenarios fall into this latter category. Thus, although they were constructed to be unlikely conjuncts based on the researcher's evaluation, this classification should be taken with some caution, as no responses were collected to verify how participants perceived these events.

DOI: 10.1002/bdm

while others may be more cognitively based. To assess some of these, the current study required prior participation in a web-based study that assessed numeracy and need for cognition. Numeracy has been defined as the ability to process basic numerical and probability information and has been shown to predict several decision making phenomena (Peters, Vastfjall, Slovic, Mertz, Mazzocco, & Dickert, 2006). Need for cognition refers to the degree to which people prefer to engage in effortful, analytic or complex thought, and it has been linked to debiasing in decision making (Simon, Fagley, & Halleran, 2004). It was hypothesized that those high in numeracy or need for cognition may take a more sophisticated approach to solving these problems and perform better.

The experimental problems were presented in a web-based survey. Included with the predictive and diagnostic problems were two class inclusion problems presented in both choice and estimation response modes. Class inclusion is a cognitive milestone that is typically achieved between 8 and 10 years of age (Winer, 1980). The child may be presented with seven dogs and four cats and asked whether there are more dogs or more animals. A failure of class inclusion occurs when one indicates there are more dogs, apparently failing to include the dogs in the animals category at comparison. The principle of class inclusion shares the same type of set–subset relationship found in the conjunction rule for prediction. Hertwig and Gigerenzer (1999) analyzed verbal protocols and have argued that participants who commit the conjunction error tend to have failed to apply the class inclusion principle. The current study measures the linkage between performance on class inclusion and conjunction errors. Time to complete the web-based problem sets was also recorded and used as a potential predictor of performance (after a log transformation to reduce skewing).

Finally, a subset of the participants came into the lab and was measured on additional individual difference variables of operation span (Turner & Engle, 1989) and Raven's progressive matrices (Raven, Court, & Raven, 1977) to see if these linked to predictive and diagnostic judgment performance. Reduction of conjunction errors for those higher in working memory or abstract reasoning ability would be consistent with work by Stanovich and West (1998), who demonstrated fewer conjunction errors for students who performed well on the Scholastic Aptitude Test (SAT).

METHOD

Participants and design

Three hundred forty seven undergraduates from the University of South Carolina volunteered to participate in this web-based survey and received participant pool credit that could be used toward psychology course requirements or extra credit. These were a subset of a larger sample who volunteered for a web based survey that included the need for cognition and numeracy measures. Additionally, 54 of the participants later came into the laboratory and participated in two computer administered measures, operation span, and Raven's progressive matrices.

The basic design consisted of the within-subjects factorial combination of problem type (prediction and diagnosis) and response mode (choice and estimation). The dependent variable was the number of times the conjunction was chosen or estimated to be more probable than the base event and could vary between 0 and 6 based on the responses to the six scenarios used in each condition. Additionally the six scenarios varied in terms of whether they were security-related terrorist situations (four) or not (two).

Materials

Table 2 presents five of the scenarios used in the study, with the sixth above (the Scandinavian problem). The table shows each scenario in the choice mode, in both prediction and diagnosis forms. In the diagnosis forms, specific options being judged were labeled A, B, and C, as in Person A or Group A, etc. The estimation form

DOI: 10.1002/bdm

of these scenarios differed only in that participants were asked to indicate a probability or likelihood for each of the three options.

Two class inclusion problems were included in both choice and estimation formats. The first was stated (in the choice format) as follows:

> In a jar there are 40 round tokens that are red and 30 round tokens that are blue. Which are there more of?
>
> Red tokens
>
> Blue tokens
>
> Round tokens

The second was stated (in the estimation format) as follows:

> There are six oranges and eight apples on a fruit stand.
>
> Please estimate the probability of randomly selecting an orange.
>
> Please estimate the probability of randomly selecting an apple.
>
> Please estimate the probability of randomly selecting a fruit.

Procedure

Participants accessed the experiment through the participant pool web site. Initial instructions informed them that the study concerned how people reasoned about the world. They were instructed that sometimes they would be asked to indicate which option was more probable and that they should do that by simply clicking on that option. Other times they would be asked to estimate probabilities. Participants were required to report probability estimates as fractions in order to avoid misinterpretation of values that arise when they are free to use percentages and proportions. They were told the following:

> There are several possible formats for indicating probabilities. However, we would like you to simply INDICATE PROBABILITY AS A FRACTION. For example, if you think it is 50% likely, you would indicate this with 1/2 or possibly 50/100. If you think the probability is quite low, say 1 out of 100 000, simply indicate 1/100 000. If you think the probability is almost guaranteed, you might put 9999/10 000. Please do not use decimals or percentages in indicating probability, just fractions.

The 24 problems resulting from the 6 (scenarios) × 2 (tasks) × 2 (response modes) combination along with the four class inclusion problems were presented in random order for each participant. Only one problem appeared on screen at a time and there was no way to go back or review answers to prior problems. For each problem and participant, the order of the three options was randomly varied. Furthermore, participants could not go back and change probabilities for an option after entering them or change choices after entering them. Problems had no time limit, and the only timing measure recorded was the time to complete the full set of problems (to which a log transformation was applied to reduce skewing).

Individual difference measures

Numeracy and need for cognition were assessed in a prior web-based survey that served as a prerequisite to the current study. Numeracy was assessed using the 11-item scale developed by Lipkus, Samsa, and Rimer (2001). Need for cognition was assessed using the 18-item scale developed by Cacioppo, Petty, and Kao (1984).

DOI: 10.1002/bdm

Additionally, participants could sign up to come into the laboratory and take two individual difference measures. Both of these were computer based and self paced. The first was based on the operation span procedure described by Turner and Engle (1989) and it was designed to measure working memory capacity. The second was based on the Raven's progressive matrices test (Raven et al., 1977), and it was designed to measure general fluid intelligence in abstract problem solving. A total of 54 participants completed these tasks.

RESULTS

Analyses of means

The dependent variable in the prediction and diagnosis tasks was the number of times a conjunction was evaluated as having a higher probability than the base event, referred to as a conjunction score. Conjunction scores were determined directly by participants in the choice format and tabulated by the experimenter for the estimation format. Evaluating the probability of a conjunction higher than a base event is an error in the prediction task, but it is a rational outcome in the diagnosis task whenever the combined cues are more diagnostic of the hypothesis than the base cue. Table 3 presents conjunction score percentages for each of the scenarios. The number of conjunction errors in prediction was quite substantial, with more conjunction errors occurring in choice than estimation, replicating results reported by Wedell and Moro (2008). Conjunction score percentages were clearly higher in diagnosis than in prediction, suggesting at least some distinction of the two tasks by participants.

A two-way repeated measures analysis of variance was conducted on the conjunction scores. A significant main effect of response mode, $F(1,346) = 365.4$, $p < .001$ indicated that conjunctions were judged higher than a constituent event more often in choice than in estimation. A significant main effect of problem type, $F(1, 346) = 196.2$, $p < .001$, indicated that conjunctions were judged higher than a constituent event more often in diagnosis than in prediction. The interaction of response mode and problem type was not significant, $F(1, 346) = 3.69$, $p > .05$, indicating that the change in response mode had similar effects for both types of problems. The pattern of effects shown in Table 3 is very similar for both non-security-based and security-based scenarios. Of four *t*-tests comparing conjunction scores for parallel conditions, only conjunction errors in estimation showed a significant difference, with non-security-related scenarios producing fewer conjunction errors (46.5%) than security-related scenarios (52.0%), $t(346) = 2.9$, $p < .01$. This difference may be largely due to the low rate of conjunction errors for the Scandinavian problem in the estimation mode, which is consistent with the generally lower rates of conjunction errors for likely conjunctions (Wedell & Moro, 2008).

Table 3. Percentage Judging the conjunction higher than base as a function of scenario, response mode, and reasoning task ($N = 347$)

	Prediction		Diagnosis	
Scenario	Choice	Estimation	Choice	Estimation
Scandinavian	62.54	28.53	85.01	61.96
IRS	93.37	64.55	95.68	63.69
Airport	74.93	63.98	91.93	75.50
Station	80.40	49.57	88.76	69.16
Paris	80.12	52.45	85.59	59.37
New York	78.10	42.07	85.88	53.03
Mean	78.24	50.19	88.81	63.78
Mean*	73.86	52.28	87.67	69.40

Note: Scenarios described in Table 2.
Mean* = the mean for the 146 participants who showed no class inclusion errors.

DOI: 10.1002/bdm

Class inclusion errors averaged 28.0% across the four class inclusion problems, which is rather high considering this is a basic task that most 10-year olds should pass. Interestingly, class inclusion errors were much more common in the choice versions of the problems (43.7%) than in the estimation versions (12.4%), $t(346) = 11.8$, $p < .001$. Table 3 also shows the mean conjunction score percentages for the subset of 146 participants who made no class inclusion errors. These results are very similar to the full sample, although there are slightly reduced conjunction errors in the choice format for this group.

Correlation, regression and path analyses

To consider individual differences in how people evaluate conjunctions in prediction and diagnosis, correlations were tabulated among the dependent variables and with relevant predictor variables. Table 4 presents these correlations and Table 5 presents the corresponding means and standard deviations for these variables. The correlations among the bottom three rows and rightmost three columns of Table 4 describe how the conjunction scores relate to one another across tasks and response modes. Strikingly, conjunction scores were strongly correlated within response mode but not within type of task. For example, conjunction scores for choice and estimation in the prediction task were only weakly correlated, $r = .22$, but conjunction

Table 4. Correlation between individual difference measures and reasoning measures for full sample ($N = 347$)

	NUM	NFC	LTIME	CLASSE	CSPC	CSPE	CSDC
NFC	0.132*						
LTIME	0.280***	0.028					
CLASSE	−0.296***	−0.112*	−0.229***				
CSPC	−0.005	−0.145**	0.089	0.183**			
CSPE	0.080	−0.116*	0.096	−0.043	0.219***		
CSDC	0.119*	−0.083	0.240***	0.015	0.546***	0.213***	
CSDE	0.141**	−0.089	0.266***	−0.104	0.184**	0.611***	0.261***

NUM, Numeracy; NFC, Need for cognition; LTime, Log of Time spent on tasks; CLASSE, Class inclusion errors; CSPC, Conjunction Scores in Predictive Choice; CSPE, Conjunction Scores in Predictive Estimation; CSDC, Conjunction Scores in Diagnostic Choice; CSDE, Conjunction Scores in Diagnostic Estimation.
$^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$.

Table 5. Means and standard deviations (in parentheses) for full sample and subsample on individual difference variables and reasoning task variables

	NUM	NFC	OSPAN	RAVEN	EQRT	WRT
$N = 347$	80.64 (19.07)	3.230 (0.651)				
$N = 54$	81.15 (19.414)	3.238 (0.718)	32.78 (6.395)	22.94 (3.993)	3,208 (830)	2,147 (774)

	LTIME	CLASSE	CSPC	CSPE	CSDC	CSDE
$N = 347$	2.762 (0.543)	28.03 (28.792)	78.24 (21.006)	50.19 (25.906)	88.81 (17.571)	63.785 (29.194)
$N = 54$	2.780 (0.594)	25.93 (27.882)	74.07 (24.372)	47.53 (27.550)	83.33 (22.663)	60.49 (30.935)

NUM, Numeracy; NFC, Need for cognition; OPSAN, Operation Span Score; RAVEN, Raven's score; EQRT, response time verifying equations in the span task; WRT, response time generating words in the span task; LTIME, Log of Time spent on tasks; CLASSE, Class inclusion errors; CSPC, Conjunction Scores in Predictive Choice; CSPE, Conjunction Scores in Predictive Estimation; CSDC, Conjunction Scores in Diagnostic Choice; CSDE, Conjunction Scores in Diagnostic Estimation.

DOI: 10.1002/bdm

scores for prediction and diagnosis were much more strongly correlated for choice, $r = .55$, and for estimation, $r = .61$. These can be considered strong correlations based on reliability estimates for conjunction scores generated from Cronbach's α that ranged from .46 to .67 in these conditions.

Overall, the pattern of correlations does not appear to support any real differentiation among the four measures based on problem type but only on response mode. This conclusion is supported by a factor analysis implemented using the maximum likelihood path modeling software, RAMONA (SYSTAT Version 12.01.02, SYSTAT Software Inc., 2007). The correlation matrix was adequately modeled by fitting only two path values to explain the six correlations. The resulting model and fit statistics are shown in Figure 2. This analysis makes it clear that the correlations between the four measures can be accounted for by a model that does not include a distinction between prediction and diagnosis but only includes a distinction between estimation and choice.

To investigate further whether there is evidence for a distinction between diagnosis and prediction based on this set of correlations, alternative models were evaluated in which the two diagnostic tasks were linked to a latent "diagnostic reasoning" factor and the two prediction tasks were linked to a latent "predictive reasoning" factor, with these two factors allowed to correlate. A model with a path structure parallel to that shown in Figure 2 fit very poorly ($\chi^2(4) = 128, p < .001$). Another model based on a multi-trait multi-method approach included correlated errors for the two choice tasks and also for the two estimation tasks, reflecting method variance. Once this correlation structure was added, the model produced an equivalent fit to that of the model shown in Figure 2. However, most importantly, this fit was achieved with the correlation between the "diagnostic reasoning" and "predictive reasoning" factors set to 1.0, so that they constituted a single factor. Thus, the alternative approaches to modeling the correlations provided no evidence for a distinction between diagnosis and prediction tasks.

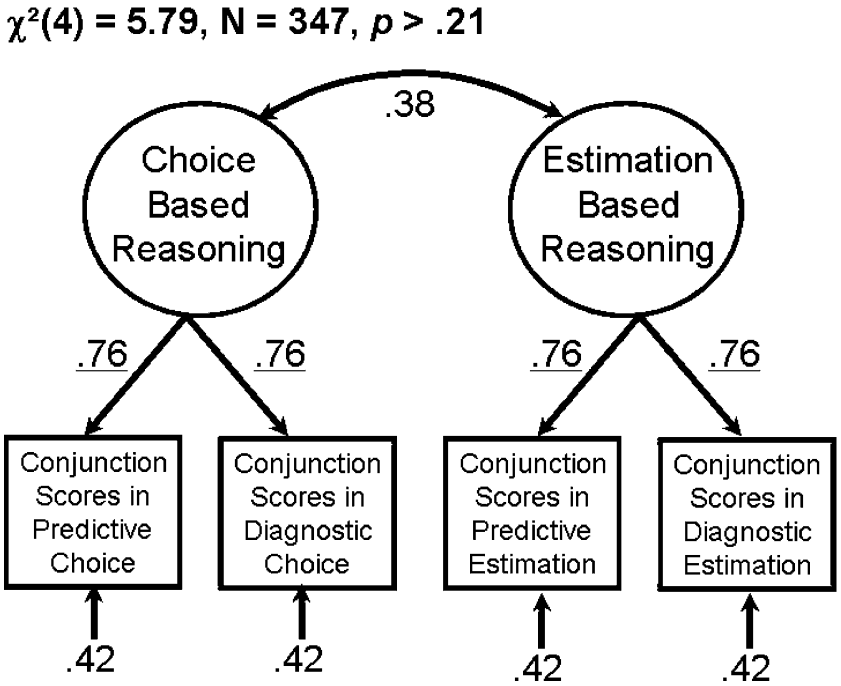
χ²(4) = 5.79, N = 347, p > .21
.38
Choice Based Reasoning
Estimation Based Reasoning
.76
.76
.76
.76
Conjunction Scores in Predictive Choice
Conjunction Scores in Diagnostic Choice
Conjunction Scores in Predictive Estimation
Conjunction Scores in Diagnostic Estimation
.42
.42
.42
.42

Figure 2. Results of a maximum likelihood factor analysis on the correlations among the four sets of conjunctions scores. Underlined values are constrained to be equal to each other so that the six correlations are modeled by two free parameters. Numbers at the bottom represent proportion of error variance for each indicator variable

DOI: 10.1002/bdm

Although the pattern of correlations among the dependent variables does not support a distinction based on problem type (prediction versus diagnosis), the differences in the means did. Differential correlations of the dependent variables with predictor variables would provide additional support for this distinction. The correlations of Table 4 indicate that numeracy significantly and positively correlated with conjunction scores for diagnostic estimation and choice but not for predictive estimation and choice. The log transformation of time spent on the task showed this same pattern as well, only somewhat stronger. In contrast, need for cognition was significantly and negatively correlated with conjunction scores for predictive estimation and choice but not for diagnostic estimation and choice. Finally, class inclusion errors significantly and positively correlated with conjunction scores in predictive choice only. These differential patterns of correlations between predictor variables and criterion variables provide some support for differential processing in predictive and diagnostic reasoning. However, it should be noted that these correlations with conjunction scores constituted small effect sizes (r's between 0.10 and 0.30), suggesting that that none of these individual difference measures provides much help in predicting how participants estimate conjunction probabilities in diagnosis and prediction. Furthermore, if one applies the Bonferroni correction as a conservative criterion for significance testing, the correlations for numeracy and need for cognition with the conjunction scores do not achieve statistical significance. Thus, caution should be exercised in evaluating these relationships.

Because correlations do not take into account the unique and shared contributions of predictor variables in explaining variance in the criterion variables, regression, and path analyses were also conducted. The analytic strategy adopted here was to first run separate regressions that tested for curvilinear and interactive relationships by including squared terms and product terms and then to model the full set of relationships using path analyses, including any additional terms that came out of the regression analyses. Of the four sets of regression analyses, none demonstrated significant curvilinear relationships and only one produced a significant interaction effect. The interaction effect consisted of the cross-product scores of numeracy and log(time) correlating with conjunction scores in predictive choice (i.e., conjunction errors). Figure 3 compares the relationship between these conjunction errors and log(time) for high numerate and low numerate participants (based on upper and lower quartiles). As shown, the relationship between log(time) and conjunction errors is positive for participants low in numeracy and negative for those high in numeracy. Thus, while more time on the task tends to help high numeracy individuals reduce conjunction errors in choice, the opposite relation is true for low numeracy individuals, with more time leading to more conjunction errors in choice.

The strategy in the path modeling was to use need for cognition and numeracy as exogenous variables, log(time) and class inclusion errors as mediating variables, and the four dependent variables as criterion variables of interest. The path analysis was conducted using RAMONA (SYSTAT Version 12.01.02, SYSTAT Software Inc., 2007) and inputting the set of correlations shown in Table 4, supplemented with the correlation of these variables with the numeracy × log(time) variable after first centering the variables.[3] Because the predictors had differential effects on the criterion variables, the criterion variables were included directly in the path model rather than occurring as indicator variables for the two latent factors shown in Figure 2. However, following the inferred structure of the factor analysis, the error scores for these four criterion variables were modeled as being correlated in the pattern described by the factor structure (using two free parameters). Non-significant paths were dropped from the model and paths were constrained to be equal when a theoretical motivation could be demonstrated for doing so.

[3] The cross-product variable combining numeracy and log time on task was created by first subtracting the geometric mean from each variable and multiplying. The geometric mean was used because it minimized the correlation of the cross-product variable with the constituent variables. The correlations of this variable with the variables described in the columns of Table 4, respectively, were as follows: -0.069, -0.005, -0.043, 0.025, -0.166^{**}, 0.036, and -0.015. The correlation with conjunction scores in diagnostic estimation was $r = 0.020$. The only significant correlation was between the cross-product variable and conjunction errors in choice. These correlations are provided for those wishing to conduct parallel analyses on the data.

DOI: 10.1002/bdm

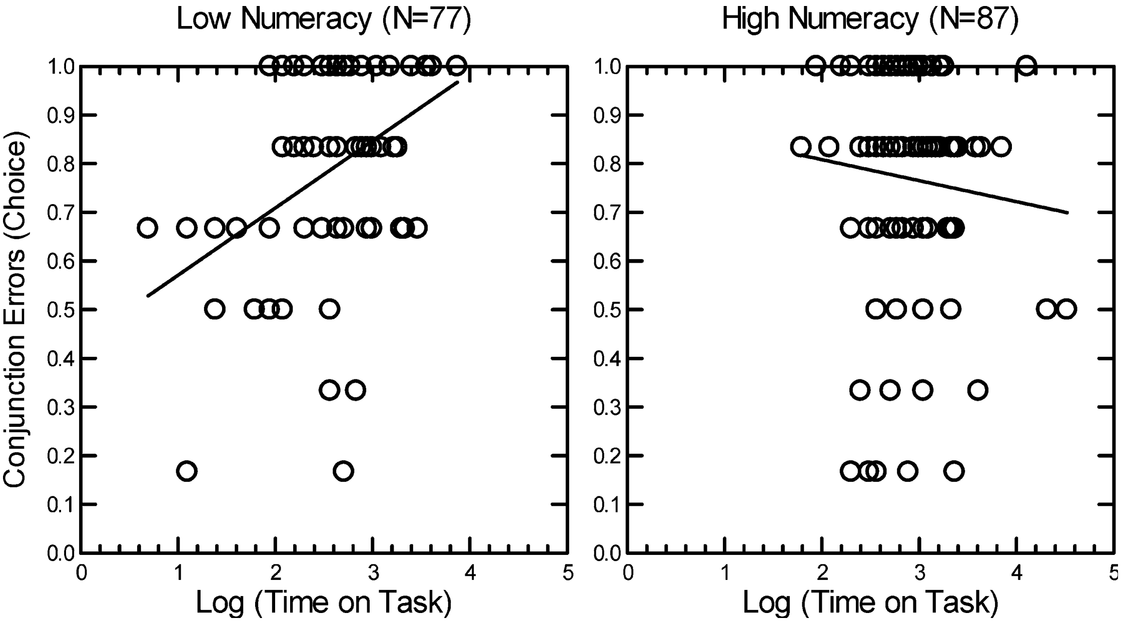

Figure 3. Conjunction errors in choice as a function of log(time) spent on the task for high and low numeracy groups. The interactive effects of numeracy and log(time) are reflected in a positive relationship of time with conjunction errors for the low numeracy group and a negative relationship for the high numeracy group

Figure 4 presents the final path model derived from the process described above. Each path shown is significant at the 0.05 level. The model explained the 36 correlations using 11 fitted parameters resulting in 25 degrees of freedom. The model adequately described the pattern of correlations, as indicated by the non-significant value of χ^2 and the low value of the root mean squared error of approximation (0.008656). Although the negatively signed direct paths between need for cognition and the two prediction tasks were significant, the reduction in conjunction errors attributable to this individual difference attribute appears quite small. The only other variable shown to directly reduce conjunction errors was the numeracy × log(time) variable, reflecting the relationship shown in Figure 3 and applying only to the choice condition. Numeracy was positively related to the log(time) spent on the task, which served as a mediator of several effects. On the positive side, increases in time spent on the task led to increases in conjunction scores for the diagnosis problems and decreases in class inclusion errors. On the other hand, greater time spent on the task reflected modest increases in conjunction errors in choice. Numeracy was also negatively related with class inclusion errors, which is shown as mediating conjunction errors in choice but not estimation. Overall, the non-significant correlations between numeracy and the two predictive tasks appear to reflect some opposing relationships that cancel out. Numeracy had a positive relationship with the two diagnosis tasks, which is modeled as being mediated by increased time spent on the task.

The path analysis described was exploratory in nature and designed to present a concise visual description of one model that accounts for the pattern of effects in the data. In general, there will always be additional path models that fit the data equally well using more free parameters and there often will be several path models that fit the data equally well using the same number of free parameters. To explore additional models that produce correlations that are not significantly different from the observed set and have fewer degrees of freedom, I successively eliminated paths that had the lowest coefficients. Although each time this was done it resulted in a significantly poorer fit, three additional parameters could be fixed at 0.0 and still provide a non-

DOI: 10.1002/bdm

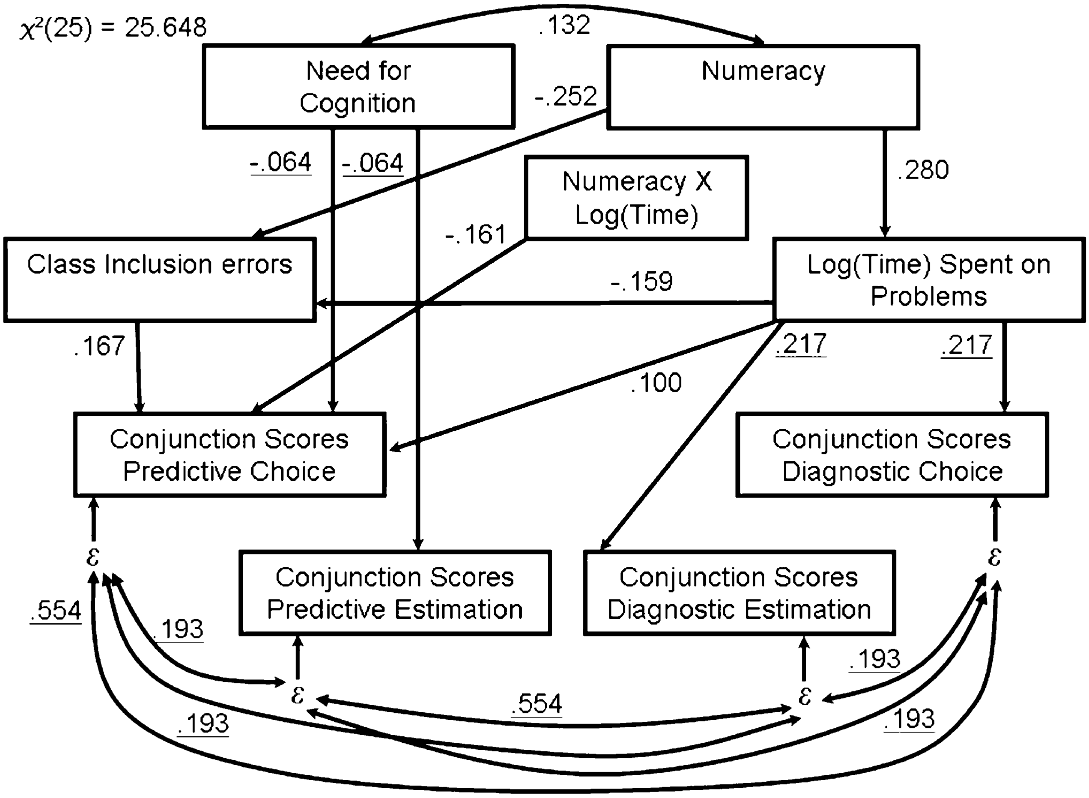
χ²(25) = 25.648
.132
Need for Cognition
Numeracy
-.252
-.064
-.064
.280
Numeracy X Log(Time)
-.161
Class Inclusion errors
-.159
Log(Time) Spent on Problems
.167
.217
.217
.100
Conjunction Scores Predictive Choice
Conjunction Scores Diagnostic Choice
ε
ε
.554
.193
Conjunction Scores Predictive Estimation
Conjunction Scores Diagnostic Estimation
.193
ε
.554
ε
.193
.193

Figure 4. Path model of the relationships among the four predictor variables and the four criterion variables. Underlined values indicate an equality constraint in the path model. All paths are significant at 0.05 level, and the model, which used 11 free parameters to explain 36 correlations, fit the data adequately ($N = 347$)

significant fit. Paths that were eliminated from those shown in Figure 4 are the paths from Need for Cognition (−0.064), the path from Log(Time) Spent on Problems (0.100) and the correlation between Need for Cognition and Numeracy (0.132). The resulting model fit the data, $\chi^2(28) = 40.262$, $p = .061$, and thus presents a somewhat simpler alternative model of the data. Note that no other additional paths could be eliminated without the model significantly deviating from the pattern of correlations.

Supplemental analyses

Table 5 also presents the means and standard deviations for the subset of 54 participants who also participated in the working memory and Raven's tasks. These are very similar to those reported for the full sample. The correlation matrix for this subsample is shown in Table 6. The first seven rows of Table 6 represent a set of correlations parallel to those shown in Table 4 but restricted to the subset of 54 participants. By and large, the subsample demonstrates similar relationships found for the full sample. The criterion variables are intercorrelated in a similar way, reflecting strong response mode dependencies. Numeracy is positively related to diagnosis in the same way and need for cognition is negatively related to prediction in the same way. The correlations of log(time) with diagnosis and numeracy are similar to the full sample, but log(time) now is also positively correlated with conjunction scores in prediction, unlike in the full sample.

The last four rows of Table 6 reflect relationships to performances on the operation span task and the Raven's progressive matrices. The two response time variables are also from the operation span task and reflect the speed of verifying equations and the speed of completing the recall of the words. As one might

DOI: 10.1002/bdm

Table 6. Correlation between individual difference measures and reasoning measures for subsample with additional measures ($N = 54$)

	NUM	NFC	LTIME	Class	CSPC	CSPE	CSDC	CSDE	OSPAN	RAVEN	EQRT
NFC	0.181										
LTIME	$\underline{0.266}$	0.164									
Class	−0.165	−0.177	−0.133								
CSPC	0.113	-0.302^{*}	0.314^{*}	0.129							
CSPE	0.291^{*}	-0.291^{*}	0.349^{*}	0.003	0.371^{**}						
CSDC	$\underline{0.240}$	−0.073	0.401^{**}	0.162	0.683^{***}	0.294^{*}					
CSDE	0.321^{*}	−0.074	0.407^{**}	−0.066	0.222	0.738^{***}	0.314^{*}				
OSPAN	−0.025	0.147	0.075	-0.277^{*}	0.007	−0.030	0.093	0.058			
RAVEN	0.174	0.375^{*}	0.140	$-\underline{0.241}$	−0.041	0.073	0.129	0.221	$\underline{0.252}$		
EQRT	-0.301^{*}	$-\underline{0.231}$	−0.014	$\underline{0.255}$	−0.035	0.105	0.016	−0.074	-0.281^{*}	−0.104	
WRT	-0.318^{*}	$-\underline{0.257}$	−0.098	$\underline{0.235}$	0.006	0.036	−0.024	−0.061	-0.345^{*}	−0.130	0.547^{***}

NUM, Numeracy; NFC, Need for cognition; LTIME, Log of time spent on reasoning task; Class, Class inclusion errors; CSPC, Conjunction Scores in Predictive Choice; CSPE, Conjunction Scores in Predictive Estimation; CSDC, Conjunction Scores in Diagnostic Choice; CSDE, Conjunction Scores in Diagnostic Estimation; OSPAN, Operation Span Score; RAVEN, Score from Ravens Progressive Matrices; EQRT, Equation Response time on Operation Span; WRT, Word recall response time on Operation Span.
Underline $= p < .10$.
$^{*} = p < .05$; $^{**} = p < .01$; $^{***} = p < .001$.

expect, numeracy significantly predicts reduced response times in the equation verification task. It also predicts reduced response times for the word recall component and so may implicate a more general processing efficiency relationship tied to numeracy. Need for cognition also correlates with the response time variables in the same way, but only marginally. Numeracy does not predict either operation span or Raven's scores, but need for cognition is positively related to performance on the Raven's progressive matrices. Class inclusion errors are negatively related to both operation span and Raven's scores. None of the eight correlations of conjunction scores with operation span scores or with Raven's scores was statistically significant. Thus, there is no evidence from this subsample that performance on either type of reasoning task is related to the measures of fluid intelligence and working memory capacity included in this study.

Finally, it is instructive to consider more directly how participants differentiated between the two tasks. To do so, difference scores were calculated in each response mode by subtracting the conjunction score in prediction from the conjunction score in diagnosis. Figure 5 presents the distributions of these scores for estimation and choice. Note that the modal score in choice is 0 and the modal score in estimation is 1. A score of 0 is expected if participants treat the two reasoning tasks in the same way. Normatively, as long as the added event has some diagnostic value, the difference score should be 6. This would reflect a conjunction score of 0 in the prediction task and a conjunction score of 6 in the diagnosis task. None of the 347 participants show this normative pattern. For both choice and estimation, the median difference score is 1. The means are 0.63 in choice and 0.82 in estimation, which do not differ significantly, $t(346) = -1.92$, $p > .05$. However, both means do differ significantly from 0.0, indicating some slight but significant differentiation of prediction from estimation for the current set of problems.

DISCUSSION

One aim of this research was to determine whether people consider probability for conjunctions differently in predictive and diagnostic tasks. In prediction, the likelihood of the events is considered given a particular hypothesis is true. As such, the probability of the conjunction of events cannot exceed that of a constituent event. In diagnosis, the likelihood of a hypothesis is being considered given the observed events. In this case, the conjunction should generally lead to higher probability judgments whenever the added event is diagnostic of the hypothesis. Analyses of means indicated that conjunction scores were significantly higher for diagnosis

DOI: 10.1002/bdm

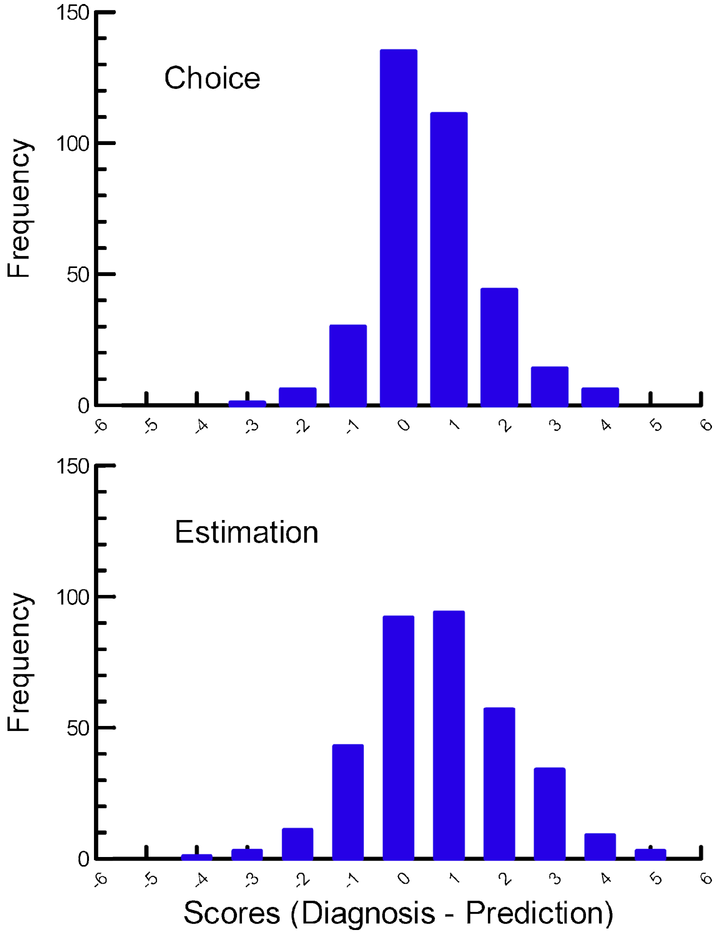

Figure 5. Distributions of difference scores for choice and prediction modes. Normatively a difference score of six is warranted if added information is diagnostic and no conjunction errors are committed. A score of zero is expected if responses in diagnosis and prediction do not differ

than for prediction, indicating that participants were making some distinction between these two types of problems. Differential patterns of correlations of diagnostic and prediction tasks to need for cognition, numeracy, and log(time) spent on the task implicate additional differences in processing for these two types of reasoning tasks, although these relationships were generally very small.

On the other hand, the difference in conjunction score percentages for the two tasks averaged only 12.1%, with the majority in both tasks endorsing the conjunction as the more likely alternative. Furthermore, the factor analysis illustrated in Figure 3 demonstrates that the correlations between the four criterion variables could be adequately explained without reference to the prediction-diagnosis distinction. Finally, the difference scores for these tasks illustrated in Figure 5 revealed that no participant showed the normative pattern and the vast majority showed either no difference or only a difference on one out of six problems when reasoning in these two modes. Thus, overall the evidence points to only subtle differences between people predicting conjunctions of events and diagnosing hypotheses from conjunctions of these same events. In essence, these results support the more general hypothesis that participants misapply diagnostic reasoning processes to the predictive reasoning task, thereby contributing to the production of conjunction errors.

DOI: 10.1002/bdm

A misapplication of diagnosis to prediction may be explained by at least two factors. First, it may occur because it is likely that diagnosis is the more common task faced by individuals in their daily life. Consider the terrorist scenarios used in the current experiment. How often does one need to predict the probability of a conjunction of events resulting from an imminent terrorist attack? While policy makers and prognosticators may need to engage in such thinking, a security guard does not. Instead, the security guard is working with the backward conditional reasoning problem found in diagnosis, namely, does this configuration of traits or behaviors implicate this person as a potential terrorist? To this end, a similarity-based heuristic may provide a good first approximation to the Bayesian algorithm, although often it appears to lead to a failure to integrate base rate information (Kahneman & Tversky, 1973).

Second, it may be that people simply fail to see that prediction and diagnosis tasks are based on different conditional probabilities. It is common in other reasoning tasks that people fail to see the asymmetry of logical relationships. Thus, people often treat an implication as a double implication (Evans, Newstead, & Byrne, 1993), and they may often treat conditional probabilities in much the same way (Eddy, 1982). Indeed, Dawes, Mirels, Gold, and Donahue (1993) demonstrated convincingly that people often give very similar estimates for a conditional probability and its inverse, ignoring the distinction between predictive and diagnostic judgments. Thus, the present results are consistent with the idea that people intuitively apply the logic of diagnosis and then mistakenly assume a symmetric relationship with prediction. This failure to apply the wrong conditional form of reasoning may be exacerbated by the differences in the relative frequencies of predictive and diagnostic tasks encountered in daily life.

As noted in the introduction, not all conjunction errors are based on conditional judgments of a conjunction of events given a hypothesized model. While the M→A paradigm follows this conditional form, as exemplified in the Linda problem, the A→B paradigm may not always follow the conditional form (Tversky & Kahneman, 1983). Thus, the idea that the predictive task is confused with the diagnostic task does not apply to those situations in which the conjunction of events does not lend itself to a diagnostic framework (Sides et al., 2002). Given the difference in the paradigms generating these conjunction errors, it seems reasonable to consider the possibility that they are due to different mechanisms.

The current experiment included a manipulation of response mode: Choosing among three options or making estimates for each of the three options. Current results replicate past findings that conjunction errors are reduced in estimation versus choice (Hertwig & Chase, 1998; Sloman et al., 2003; Wedell & Moro, 2008). By including this manipulation within the diagnosis task, the present study helps to answer the question of whether estimates lead to better judgments overall. If so, then one should observe a greater difference in conjunction scores between diagnosis and prediction conditions in estimation than in choice. However, this was not the case. For both types of problems, estimation led to a similar reduction in conjunction scores. For the combined tasks, estimation did not lead to a more normative pattern than choice. Indeed, the factor pattern of Figure 3 assumes equal loadings for the two estimation tasks on one factor and the two choice tasks on the other. This pattern reflects a clear difference between these response modes but does not implicate a difference between predictive and diagnostic reasoning as a function of response mode.

One may question the degree to which the completely within-subjects design employed in the current study affected the results. For example, could the tendency to treat diagnosis and prediction tasks very similarly be greater when both types of problems appear together? Although this might be the case, often the debate over between- and within-subjects designs argues just the opposite: That is, the transparency of the within-subjects design should enhance debiasing because errors become more apparent in this design (Mellers, Hertwig, & Kahneman, 2001). In previous research, Wedell and Moro (2008) used specific blocked orders and tested for transfer effects but found none. This was surprising in that the same individuals who avoided the conjunction error in estimation would turn around and commit the error in choice. The Scandinavian problem was also used in both the present study and by Wedell and Moro. Note that the conjunction error results for choice (63%) and estimation (29%) in the present study were similar to that reported in the previous study (63% and 18%, respectively). Hence it seems unlikely that the presence of the

DOI: 10.1002/bdm

diagnostic task was altering responses to the prediction task. One might also argue that the presentation of problems in the current experiment led to misinterpretations. However, the use of the three options that include the base event and the complement of the added event has been shown to strongly reduce possible confusion (Tentori et al., 2004; Wedell & Moro, 2008). Furthermore, the use of specific persons or groups (A, B, and C) only in the diagnostic task was designed to help differentiate the two cases. Therefore, it seems unlikely that these results are simply the result of the specific way in which problems were manipulated and presented in the current study.

An important aspect of the current study was the measurement of several individual difference variables. The within-subject manipulation of response mode and reasoning task made it possible to examine how individual difference variables relate to these manipulations. Class inclusion problems were included to examine the link between performance on these and conjunction errors. More than half the sample made at least one error in the set of four class inclusion problems, with the vast majority of these errors occurring in the choice mode. Correlation and path analyses indicate that those who made more class inclusion errors made significantly more conjunction errors in choice, although the effect size was rather small. This relationship is consistent with the choice mode promoting a more qualitative or heuristic based assessment of the problem. However, the pattern of results is nearly the same for the subsample that showed no class inclusion errors as for the full sample, so it appears that the high level of conjunction errors is not simply due to an obvious failure to apply the class inclusion principle (for a further discussion of the relationship of class inclusion and conjunction errors, see Reyna, 1991).

Both numeracy and need for cognition have been demonstrated to impact decision making (Simon et al., 2004). Peters et al. (2006) have provided evidence that more numerate individuals may make better decisions because they are better able to extract and use the appropriate numerical principles. Although numeracy has been shown to reduce framing effects, the current study found no evidence that numeracy reduced conjunction errors. Thus, numeracy apparently does not help one extract the proper set relations underlying the prediction task. Numeracy was positively correlated with conjunction scores in the diagnosis task, although again the effect size was quite small. Path modeling supported the idea that this relationship was mediated by high numeracy individuals spending more time on the task, perhaps because they found the numerical task less aversive than the low numerate individuals. Need for cognition did not correlate with the diagnosis task but did correlate negatively with conjunction errors in the prediction task, although once again the effect size was small. Path analyses supported the idea that this relationship was due to a direct effect, perhaps with high need for cognition individuals more willing to carefully examine the prediction problem and thus make slightly fewer errors. However, given the very small effect size, the results are generally supportive of the robust nature of conjunction errors for different groups of individuals (Tversky & Kahneman, 1983).

Data from the subsample of 54 participants who were also measured on working memory and Raven's progressive matrices did not provide additional insights into probabilistic reasoning about conjunctions. Given the significant negative correlation of need for cognition with conjunction error measures, one might expect these measures to also negatively correlate with conjunction errors. But this was not the case, even though need for cognition positively correlated with the Raven's measure. Stanovich and West (1998) reported a significant correlation between SAT scores and conjunction errors on the classic Linda problem originally presented by Tversky and Kahneman (1983), with those not making conjunction errors scoring 80 points higher on combined SAT than those committing the error (a correlation of $r = -.28$). The fact that SAT scores predict conjunction errors but working memory and Raven's scores do not suggests that differences in crystallized knowledge may be more responsible for this reduction than differences in fluid intelligence. However, such a conclusion would require corroboration from a study that examined both types of intelligence measures.

More generally, the weak relationships between reasoning in these tasks and individual differences in numeracy, cognitive style, and fluid intelligence measures observed here are supportive of Tversky and Kahneman's (1983) original contention that conjunction errors may be likened to cognitive illusions. Like perceptual illusions, these cognitive illusions may arise out of an efficient and automatic process that provides

DOI: 10.1002/bdm

adaptive response tendencies within the organism's natural environment. Within the two-system processing framework adopted by many researchers in the field (Kahneman, 2003; Sloman, 1996; Stanovich & West, 2000), such an automated, associative process would be considered part of System 1 processing, or intuition, as opposed to System 2 processing, or deliberative reasoning. In this regard, Stanovich and West (1998) have suggested that the impact of individual differences related to intellect on decision making tasks will depend in large extent on how strongly System 1 or System 2 processing is cued by the task and whether these cues are in opposition to one another. The hypothesized confusion of diagnosis with prediction maps on to this distinction in two ways. First, if diagnosis is mistakenly assumed rather than prediction, then responses linked to System 1 and System 2 will tend to be consistent and hence errors will not be detected. Second, if diagnosis constitutes the predominant focus in the natural environment, then it may be adaptive that the System 1 utilizes heuristics consistent with this approach.

In conclusion, the present study provided a large-sample parametric examination of how people reason about conjunctions in predictive and diagnostic situations using choice or estimation responses. While small differences were demonstrated in responses to diagnostic and prediction tasks, they fell far short of differences demanded by the normative standards. The main reason for this was that conjunction errors were very robust. Individual difference measures only weakly related to conjunction errors, as has been the case in the past. While estimation led to a considerable decline in conjunction errors, estimation affected diagnosis in much the same way so that the two tasks were not differentiated more in estimation than in choice. Overall, the pattern of results is very much in line with the idea that people intuitively apply diagnostic reasoning strategies to the prediction task, resulting in conjunction errors. This behavior may be a reasonable adaptation to an environment that places the organism in situations demanding diagnosis much more often than in situations demanding prediction.

ACKNOWLEDGEMENTS

The author acknowledge the help of Filip Szymanski and Jessica Anderson who helped in the programming of the web experiment and the collating of the data. The authors thanks Sylvia Fitting for her helpful comments on an earlier version of this manuscript.

REFERENCES

Anderson, N. H., & Shanteau, J. C. (1970). Information integration in risky decision making. *Journal of Experimental Psychology*, *84*, 441–451.

Bar-Hillel, M., & Neter, E. (1993). How alike is it versus how likely is it: A disjunction fallacy in probability judgments. *Journal of Personality and Social Psychology*, *65*(6), 1119–1131.

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology*, *74*, 271–280.

Dawes, R. M., Mirels, H. L., Gold, E., & Donahue, E. (1993). Equating inverse probabilities in implicit peronsality judgments. *Psychological Science*, *4*, 396–400.

Eddy, D. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge, England: Cambridge University Press.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Evans, J. St. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). *Human reasoning: The psychology of deduction*. Hove, UK: Erlbaum.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review*, *90*, 239–260.

Hertwig, R., & Chase, V. M. (1998). Many reasons or just one: How response mode affects reasoning in the conjunction problem. *Thinking and Reasoning*, *4*, 319–352.

DOI: 10.1002/bdm

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

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist*, *58*, 697–720.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*, 237–251.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269–275.

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

Peters, E., Vastfjall, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*, 407–413.

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.

Reyna, V. F. (1991). Class inclusion, the conjunction fallacy, and other cognitive illusions. *Developmental Review*, *11*, 317–336.

Raven, J. C., Court, J. H., & Raven, J. (1977). *Raven’s progressive matrices and vocabulary scales*. New York: Psychological Corporation.

Simon, A. F., Fagley, N. S., & Halleran, J. G. (2004). Decision framing: Moderating effects of individual differences and cognitive processing. *Journal of Behavioral Decision Making*, *17*, 77–93.

Sides, A., Osherson, D., Bonini, N., & Viale, R. (2002). On the reality of the conjunction fallacy. *Memory & Cognition*, *30*, 191–198.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*, 3–22.

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*, 296–309.

Stanovich, K. E., & West, R. F. (1998). Individual differences in framing and conjunction effects. *Thinking and Reasoning*, *4*, 289–317.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, *23*, 645–665.

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

Tetlock, P. E., & Boettger, R. (1989). Accountability: A social magnifier of the dilution effect. *Journal of Personality and Social Psychology*, *57*, 388–398.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, *28*, 127–154.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.

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

Wells, G. L. (1985). The conjunction error and the representativeness heuristic. *Social Cognition*, *3*, 266–279.

White, P. A. (2003). Effects of wording and stimulus format on the use of contingency information in causal judgment. *Memory & Cognition*, *31*, 231–242.

Winer, G. A. (1980). Class-inclusion reasoning in children: A review of the empirical literature. *Child Development*, *51*, 309–328.

Wolford, G., Taylor, H. A., & Beck, J. R. (1990). The conjunction fallacy? *Memory & Cognition*, *18*, 47–53.

Author’s biography:

Douglas Wedell is a Professor of Psychology at the University of South Carolina. His research interests focus primarily on contextual models of judgment and choice and understanding the nature of bias in decision making.

Author’s address:

Douglas Wedell, Department of Psychology, University of South Carolina, Columbia, SC 29208, USA.

DOI: 10.1002/bdm