

# Dilution and confirmation of probability judgments based on nondiagnostic evidence

CARLA LABELLA and DEREK J. KOEHLER  
University of Waterloo, Ontario, Canada

Previous research has shown that probability judgments based on a mix of diagnostic and nondiagnostic information are less extreme than judgments based on the diagnostic information alone. Results of the present experiments suggest that this *dilution effect* holds only under a limited set of conditions. When judgments based on a mix of diagnostic and nondiagnostic information are compared with separately elicited judgments based on the diagnostic information alone, the dilution effect is consistently observed. When judgments based on the diagnostic evidence are revised in light of additional, nondiagnostic evidence, by contrast, the dilution effect is eliminated or even reversed (yielding a *confirmation effect*) depending on the type of nondiagnostic evidence under evaluation.

In evaluating a pair of competing hypotheses, any evidence that is equally likely to be observed under either hypothesis is said to be *nondiagnostic*. Normatively, the probability assigned to one of the hypotheses should be determined exclusively on the basis of diagnostic evidence and should be uninfluenced by nondiagnostic evidence. In other words, the judged probability of the hypothesis based on the full body of diagnostic and nondiagnostic evidence should not differ from that based on the diagnostic evidence alone (assuming conditional independence of the two subsets of evidence). Descriptively, however, a number of studies have demonstrated that people's probability judgments are systematically influenced by nondiagnostic evidence (Nisbett, Zukier, & Lemley, 1981; Shanteau, 1975; Troutman & Shanteau, 1977; Zukier, 1982; Zukier & Jennings, 1983).

Specifically, probability judgments based on a combination of diagnostic and nondiagnostic evidence tend to be less extreme than judgments based on the diagnostic evidence alone, a phenomenon that Nisbett et al. (1981) called the *dilution effect*. In one of their studies, graduate students in social work were presented with client descriptions and given the task of rating the likelihood that the client was a child abuser. Each of the client descriptions included one or more pieces of diagnostic information (e.g., the client was sexually assaulted by his stepfather), crossed with none, two, four, or eight pieces of nondiagnostic information (e.g., he manages a hardware store; has an IQ of 110). The resulting descriptions were evaluated in a random order with interspersed filler items. When evaluated on its own in a pretest study, each piece

of nondiagnostic information was indeed rated as nondiagnostic. Despite this, however, participants rated the client as less likely to be a child abuser when the description included a mix of diagnostic and nondiagnostic information than when the description included only diagnostic information.

What Nisbett et al. (1981) called the dilution effect had been observed previously by Troutman and Shanteau (1977; see also Shanteau, 1975) in a standard "beads and jars" belief-revision task. Participants were shown two jars filled with differently colored beads in known proportions and were provided with a sample of beads drawn at random from one of the two jars. Their task was to judge the probability that the beads had been drawn from a designated jar on the basis of an initial sample of beads and then to revise their judgment as further bead samples were drawn from the same jar. Following an initial, diagnostic, sample of beads, the judged probability assigned to the implicated jar was well above 50%, but then was adjusted to be less extreme following an additional, nondiagnostic, sample. When evaluated in isolation, though, the nondiagnostic bead sample was indeed viewed as nondiagnostic, in the sense that probability judgments for the target jar were very close to 50%. In both the Nisbett et al. and Troutman and Shanteau studies, then, evidence rated as nondiagnostic on its own nonetheless produced a dilution effect when combined with diagnostic evidence, in comparison with judgments based on the diagnostic portion of the evidence alone.

The dilution effect is a widely cited and influential result. Collectively, the Nisbett et al. (1981) and Troutman and Shanteau (1977) articles have been cited over 150 times since their publication. The consequences of the dilution effect have been examined in areas as diverse as the law (Smith, Stasson, & Hawkes, 1999), accounting (Hackenbrack, 1992; Shelton, 1999), and consumer behavior (Meyvis & Janiszewski, 2002). In this paper, we investigate the influence of different types of nondiag-

---

Portions of this article are based on C.L.'s master's thesis, completed at the University of Waterloo. The reported research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to D.J.K. Correspondence should be addressed to D. J. Koehler, Department of Psychology, University of Waterloo, Waterloo, ON, N2L 3G1 Canada (e-mail: dkoehler@uwaterloo.ca).

nostic evidence and different judgment tasks, and their interaction, on the magnitude of the dilution effect as a means of better understanding its causes. Although the dilution effect is a well-established and widely accepted phenomenon in the research literature on judgment under uncertainty, our investigation identifies conditions under which the dilution effect is eliminated or even reversed, yielding what we refer to as a *confirmation effect*. While our results are only preliminary at this point, the identification of possible boundary conditions on the occurrence of the dilution effect may shed light on the judgmental mechanisms that produce it. In the General Discussion section, for example, we show that an averaging model of information integration, which readily reproduces the basic dilution effect (Troutman & Shanteau, 1977), is unable to accommodate our results without additional assumptions regarding changes in how individual pieces of evidence are interpreted as determined by previously encountered evidence.

In our investigation, we use the beads and jars task employed in the Troutman and Shanteau (1977) study. One advantage of this task is that it gives nondiagnostic evidence a clear, objective definition—namely, as any sample of beads that is equally likely to be drawn from either jar. Arguably, such a task leaves less room for variance among individuals in their perceptions of the implications of the nondiagnostic evidence than might be found in perceptions of qualitative evidence of the kind used, for example, by Nisbett et al. (1981). Relatedly, the beads and jars task is not subject to the “conversational norm” critique (e.g., Slugoski & Wilson, 1998; Tetlock, Lerner, & Boettger, 1996) that has been applied to studies conducted in more naturalistic, nonstochastic settings (such as that of Nisbett et al.). According to this critique, participants feel obliged to adjust their judgments in response to nondiagnostic evidence because they assume that the experimenter would not have presented it if it were not somehow relevant. This critique holds less force when applied to the beads and jars task, in which the evidence is generated by a random device rather than by a human being.

A final advantage of the beads and jars task is that it allows evaluation of the impact of different types of nondiagnostic evidence. Specifically, we can distinguish two types of nondiagnostic evidence: neutral and mixed. The Troutman and Shanteau (1977) study investigated the impact of both types of nondiagnostic evidence, though they used different terms. This distinction is most easily illustrated by reference to one of the pairs of jars (A and B) used in our studies (Table 1). Each jar contains a total of 1,000 beads, each of which is red (R), green (G), blue (B), or yellow (Y). Beads are always drawn in pairs; that is, each bead sample consists of a pair of beads. A sample consisting of a pair of blue beads (BB) is clearly nondiagnostic with respect to determining whether the sample was drawn from Jar A or Jar B, because each jar contains an equal proportion of blue beads. We refer to the BB sample as an example of *neutral* nondiagnostic

**Table 1**  
**Depiction of Jar Cases AB and CD**

	Case AB (Moderate Discriminability)		Case CD (High Discriminability)	
	Jar A	Jar B	Jar C	Jar D
Red	270	120	540	60
Green	120	270	60	540
Blue	200	200	200	200
Yellow	410	410	200	200
Total	1,000	1,000	1,000	1,000

evidence, because it consists of two beads, each of which is nondiagnostic on its own and hence neutral in its implications. A sample consisting of one red and one green bead (RG) is also nondiagnostic, because the red and green beads occupy complementary proportions of the two jars and hence drawing exactly one red and one green bead is equally likely for either jar. We refer to the RG sample as an example of *mixed* nondiagnostic evidence because each bead is diagnostic on its own but with opposite implications that effectively “cancel” each other. Identification of mixed evidence as nondiagnostic might be more difficult because it requires integrating the conflicting implications of individually diagnostic pieces of information, which is not required for neutral evidence. Troutman and Shanteau observed comparable degrees of dilution for both types of nondiagnostic evidence in their Experiment 1, but may not have had sufficient statistical power to detect differences between the two. In the present study, we directly compare the amount of dilution produced by the two types of evidence and investigate how any observed difference between the two varies across judgment tasks.

One important difference between the Troutman and Shanteau (1977) and the Nisbett et al. (1981) studies involves the judgment task presented to their participants. As described above, Troutman and Shanteau used a *belief-revision* procedure in which an initial judgment based on diagnostic evidence is compared with a subsequent revision of that judgment when additional, nondiagnostic, evidence is provided. To illustrate, consider our experimental materials, described in Table 1. In the belief-revision task, the participant makes a first probability judgment based on an initial, diagnostic, bead sample (such as RR) and then makes a second, revised, judgment based on an additional, nondiagnostic, bead sample (such as BB or RG) drawn from the same jar. The dilution effect is observed if the second judgment is less extreme than the first.

By contrast, the Nisbett et al. (1981) study compared separately elicited, nonsequential judgments based either on diagnostic evidence alone or else on an aggregate of diagnostic and nondiagnostic evidence. As applied to our experimental materials, this approach would involve a comparison of judgments based on diagnostic evidence only (such as the bead sample RR), with separate judgments based on a combination of diagnostic and

nondiagnostic evidence (such as the bead samples RRBB or RRRG). We will refer to this as an *aggregate* procedure, in which judgments based on two separately evaluated bodies of evidence are compared. Troutman and Shanteau (1977) included two aggregate judgment trials in their Experiment 2, results from which did not show a dilution effect (in contrast to the later results of Nisbett et al.), providing an initial hint that the magnitude of the dilution effect may depend on whether the judgment task takes the form of a belief-revision or aggregate procedure. Troutman and Shanteau's Experiment 2 used only mixed nondiagnostic evidence, so it remains an open issue whether similar results would obtain with neutral evidence.

In the present research, we compared the influence of nondiagnostic evidence in a belief-revision and an aggregate judgment procedure. If we find that the same nondiagnostic evidence has a different influence on judgments in the two procedures, it could help to shed light on the way in which nondiagnostic evidence exerts its influence. Compared with the aggregate procedure, the belief-revision procedure seems more likely to prompt explicit evaluation of the nondiagnostic subset of the evidence on its own, because participants must evaluate whether and by how much to revise their original judgment (based on the diagnostic subset of the evidence alone) in light of the additional, nondiagnostic evidence. In this sense, the belief-revision procedure segregates the nondiagnostic subset of the evidence in a way that the aggregate procedure does not. If the dilution effect is at least partly due to a failure to spontaneously segregate the nondiagnostic subset of the evidence from the remaining, diagnostic evidence, then it should be less pronounced in the belief-revision procedure (where the nondiagnostic evidence is segregated from the diagnostic evidence) than in the aggregate procedure (where it is not). Alternatively, the belief-revision task may encourage some kind of adjustment from the initial judgment, in which case a dilution effect might be more likely for the belief-revision than for the aggregate task (consistent with Troutman & Shanteau's, 1977, results).

Finally, in the present research we can determine whether our two critical independent variables (evidence type and task type) exert interacting influences on the magnitude of the dilution effect, a possibility that has not been tested in previous research. That is, we can examine whether judgments based on neutral and mixed nondiagnostic evidence are similarly or differently influenced by whether the judgment task involves a belief-revision or aggregate procedure. For example, mixed evidence might be more readily identified as nondiagnostic when it is segregated from the diagnostic evidence (in the belief-revision procedure) than when it is not (in the aggregate procedure), while neutral evidence might be readily identified as nondiagnostic even when it is not segregated from the diagnostic evidence. As another example, a working hypothesis established during the initial step of the belief-revision procedure (based on the

diagnostic evidence), which is not established in the aggregate procedure, might influence how the participant evaluates subsequent evidence, particularly if the implications of that evidence are relatively difficult to assess (as we suggested might be the case for mixed evidence).

The Troutman and Shanteau (1977) studies focused on judgments of the hypothesis implicated by the initial, diagnostic, evidence and showed that such judgments (which are relatively high when based on the diagnostic evidence alone) *decrease* when nondiagnostic evidence is introduced. The interpretation of this result as a dilution effect implies that judgments of the nonimplicated hypothesis (which is made less likely by the initial diagnostic evidence) should *increase* following the addition of nondiagnostic evidence. That is, a dilution effect should lead to the implicated hypothesis being judged as less likely and the nonimplicated hypothesis being judged as more likely following the introduction of nondiagnostic evidence. Somewhat surprisingly, however, the latter result was not directly tested. (Nisbett et al., 1981, did examine whether judgments of a nonimplicated hypothesis increased with the addition of nondiagnostic information, but found mixed results across their studies, with the judgments increasing in some experiments but remaining unchanged in others.) Without eliciting judgments of the nonimplicated hypothesis, it is not possible to determine whether the introduction of nondiagnostic evidence leads to less extreme judgments (as implied by calling the pattern one of dilution) or instead to generally lower judgments. In our studies, we also include judgments of the jar not implicated by the initial, diagnostic, information to provide a complete test of the presumed dilution pattern.

In short, the key empirical contribution of our studies is to provide a broader examination of the dilution effect and the conditions under which it is observed than has been conducted in previous research. The manner in which probability judgments are influenced by nondiagnostic information is of theoretical significance because it can help to distinguish alternative models of how people evaluate and integrate information in making inferential judgments. One fundamental issue in the study of information integration, for example, is whether the process is best captured by an additive or an averaging model (e.g., Hogarth & Einhorn, 1992). The observation of the dilution effect was taken as evidence favoring the averaging model (Troutman & Shanteau, 1977), since it can more readily reproduce the effect. (See the Appendix for an application of additive and averaging models to the data from our experiments.) Testing the generality of the dilution effect and identifying its boundary conditions places additional constraints on the kind of model required to capture intuitive inferential judgments.

## Overview of Experiments

Experiment 1 used an aggregate judgment procedure, and Experiment 2 used a belief-revision procedure. The two experiments involved identical beads and jars prob-

lems, described here, such that judgments were elicited contingent on the same evidence (diagnostic only, nondiagnostic only, or a combination of diagnostic and nondiagnostic) in both experiments.

Information regarding the two critical problems presented in each experiment is shown in Table 1. Each jar contains 1,000 beads. Jars A and B are pitted against each other in Case AB; Jars C and D are pitted against each other in Case CD. In both cases, the two jars being compared differ in the number of red (R) and green (G) beads that they contain. This difference is more pronounced in Case CD (9:1 ratio of R:G and G:R beads for Jars C and D, respectively) than in Case AB (7:3 ratio of R:G and G:R beads for Jars C and D, respectively), meaning that drawing a red (green) bead provides stronger evidence for Jar C (D) in Case CD than it does for Jar A (B) in Case AB. This allows us to establish the generalizability of our results over varying levels of diagnosticity associated with the diagnostic portion of the evidence. In both cases, there is an equal number of blue (B) beads in each jar, meaning that drawing one or more blue beads is nondiagnostic with respect to discriminating between either pair of jars. Each jar also contained yellow (Y) beads, which occurred in equal numbers within a case but were more prevalent in Case AB than in Case CD.

For a given case, participants were told that one of the two jars would be selected at random and samples of beads would be drawn from that jar. On the basis of the bead samples, participants were asked to judge the probability that a designated jar was the one from which the samples were being drawn. The "target" jar thus designated was determined on a randomized basis on each trial. Participants judged the probability that the bead sample had been drawn from the target jar rather than from the alternative, nontarget jar on a 0% to 100% probability scale.

Beads were always drawn in pairs. In either Case AB or Case CD, drawing an RG bead pair is nondiagnostic with respect to discriminating between the two jars because red and green beads occupy complementary proportions in both; we refer to the RG bead pair as *mixed* evidence. Drawing a BB bead pair is also nondiagnostic; we refer to the BB pair as *neutral* evidence. (A pair of yellow beads would also be considered neutral evidence, but no such pairs were presented as evidence in our experiments.)

In addition to these two nondiagnostic bead pairs, participants also encountered four different diagnostic bead pairs: RR, GG, RY, and GY. Judgments were elicited contingent on each bead pair (diagnostic or nondiagnostic) on its own, and also for each possible combination of a diagnostic bead pair and a nondiagnostic bead pair. (The order in which the two beads within a pair were presented was also counterbalanced.) There are eight possible ways of combining the four diagnostic and two nondiagnostic bead pairs. Each combination can be presented with the diagnostic pair first (Order 1, e.g., RRRG) or with the nondiagnostic pair first (Order 2, e.g., RGRR), yielding 16 four-bead samples: RRRG, GGRG, RYRG,

GYRG, RRBB GGBB, RYBB,GYBB, RGRR, RGGG, RGRY, RGGY, BBRR, BBGG, BBRY, BBGY.

For each of these four-bead samples, two judgments were elicited, one contingent on the first pair of beads and the second contingent on both pairs taken together. The timing with which these two judgments were elicited is the fundamental procedural distinction between Experiments 1 and 2. Using the first four-bead sample (RRRG) as an example, the aggregate procedure (Experiment 1) would involve eliciting separate nonconsecutive judgments contingent on the first pair (RR) and on all four beads (RRRG). The belief-revision procedure (Experiment 2) would involve eliciting a judgment contingent on the first pair (RR) followed by a revised judgment contingent on the addition of the second pair (RG). In terms of experimental design, trials were blocked by number of beads in the sample (two or four) in the aggregate procedure of Experiment 1, and by four-bead sample in the belief-revision procedure of Experiment 2.

## EXPERIMENT 1 Aggregate Procedure

Several studies demonstrating the dilution effect have compared judgments based on diagnostic evidence only with completely separate judgments based on a mix of diagnostic and nondiagnostic evidence (Nisbett et al., 1981; Zukier, 1982; Zukier & Jennings, 1983). In this experiment, the aggregate samples (e.g., RRRG) were divided into two blocks so that judgments would be based on the first two beads in one block (e.g., RR) and on four beads (e.g., RRRG) in the other. The question we investigated was whether the two types of nondiagnostic evidence, mixed (RG) and neutral (BB), would produce a comparable degree of dilution. We suspected that the size of the effect might generally be larger for the RG pair because the nondiagnosticity of this pair is intuitively less obvious than is the case for the BB pair.

### Method

**Participants.** Seventy-one students enrolled in an introductory psychology class participated in this study for partial fulfillment of their course requirements. Data from 7 participants were excluded from the analyses for the following reasons: (1) accuracy analyses revealed that 3 participants were outliers (on average, judgments deviated by more than .25 from the corresponding Bayesian calculation); (2) 2 participants failed to finish the task; (3) 1 participant entered the same four numbers (1–4) for most of the trials; and (4) 1 participant was excluded because she reported that she entered the same response for all of the trials in the second half of the experiment.

**Stimuli.** For the practice phase, a turntable and two jars with varying proportions of red, green, blue, and yellow beads were used. For the test phase, the bead samples and the competing jars were presented on an IBM-compatible computer.

**Practice phase.** The experiment began with a physical demonstration of the bead-sampling procedure that would subsequently be simulated in the computerized task. Two clear jars, with varying proportions of colored beads, were used. The ratio of R:G:B:Y (red:green:blue:yellow) beads in the two jars was as follows: Jar 1, 400:100:200:300; and Jar 2, 100:400:200:300. There were a total of 1,000 beads per jar. The surface of a turntable was divided into two

equal sections and the two jars were placed on each side of this dividing line. The composition of these jars was illustrated on a screen concealing the turntable. Behind the screen, the experimenter proceeded to spin the table. Simultaneously, the participant chose a slip of paper, with an illustration of the target jar (to which the participant would be asked to assign a probability). The table ceased spinning with an arrow pointing to the selected jar. The experimenter then drew two beads from the selected container. After seeing the drawn beads, the participant was asked to make a probability judgment, between 0% and 100%, that the beads were in fact drawn from the target jar indicated by the slip of paper, rather than from the remaining, nontarget, jar. The same process was then repeated with a new jar being selected, but this time, four beads were drawn from the selected jar. This was followed by an additional two- and four-bead practice trial. No feedback was provided after each trial.

**Test phase.** Following this practice phase, participants were told that the same general procedure would be simulated on a computer for the rest of the experiment. Instructions on completing the computer-based task were then provided. Each participant made judgments for two jar cases (Cases AB and CD), shown in Table 1. Participants either received the two-bead block first followed by the four-bead block, or vice versa. Within each block, two subblock orders were used: AB, CD; or CD, AB. The same subblock order was used for each block. For each subblock, the composition of the two jars under consideration was shown on the screen and the participants were presented with samples of either two or four beads. For the four-bead block, 16 aggregate bead samples were presented (RRRG, GGRG, RYRG, GYRG, RRBB, GGBB, RYBB, GYBB, RGRR, RGGG, RGRY, RGGY, BBRR, BBGG, BBRy, BByG). For the two-bead block, participants were presented with only the first pair of beads from the four-bead strings so that for half of the trials, participants received a diagnostic pair on its own (e.g., RR, GG, RY, and GY) and on the remaining half of the trials, they received a nondiagnostic pair only (e.g., RG and BB). Each jar under consideration served as the target jar for each sample at some point during each subblock, in a randomly determined fashion so that the target jar varied from trial to trial. Participants judged the probability that the sample had been drawn from the target jar (which was designated by a box drawn around it) rather than from the remaining nontarget jar. In total, 128 judgments were obtained from each participant.

## Results

Bead samples including yellow beads were presented to maintain the plausibility of our cover story to participants that the samples were drawn randomly from the jars. Our analysis, however, focused on samples involving only red, green, and blue beads because these samples were the closest to the ones analyzed in the original Troutman and Shanteau (1977) study. Thus, we concentrated our analysis on eight sample pairs (RR–RRRG, GG–GGRG, RR–RRBB, GG–GGBB, RG–RGRR, RG–RGGG, BB–BBRR, and BB–BBGG).

**Mean judgments.** We analyzed the first four sample pairs, RR–RRRG, GG–GGRG, RR–RRBB, and GG–GGBB (Order 1 samples, where the diagnostic portion of the evidence was presented first in the aggregate sample) to determine whether participants' judgments based on the aggregate samples (e.g., RRRG) were diluted compared with those based on samples containing the diagnostic information alone (e.g., RR).

Table 2 reports the means and the results from a simple effects analysis. For this analysis, the data were col-

**Table 2**  
Mean Probability Judgment Assigned to the Target Jar (Implicated or Nonimplicated) Based on Diagnostic Evidence Alone or on Diagnostic and Nondiagnostic Evidence, Along With Their Difference ( $\pm 95\%$  Confidence Interval) and Associated Simple Effects Test, for Mixed (RG) and Neutral (BB) Nondiagnostic Evidence in Experiment 1, Presented Separately for Each Case (AB and CD)

Case	ND Sample	Jar	Judged Probability		
			D only	D+ND	Difference
AB	RG	implicated	72.0	68.6	$-3.3 \pm 3.3^*$
		nonimplicated	29.1	28.1	$-0.9 \pm 3.1$
	BB	implicated	71.1	66.5	$-4.7 \pm 2.4^{**}$
		nonimplicated	30.3	36.1	$5.8 \pm 3.2^{**}$
CD	RG	implicated	83.4	78.3	$-5.1 \pm 3.9^{**}$
		nonimplicated	17.4	22.9	$5.5 \pm 4.9^*$
	BB	implicated	83.1	74.0	$-9.2 \pm 3.1^{**}$
		nonimplicated	18.9	28.1	$9.2 \pm 4.7^{**}$

Note—D, diagnostic bead sample; ND, nondiagnostic bead sample; R, red; G, green; B, blue.  $*p < .05$ .  $**p < .01$ .

lapsed across the diagnostic portion (RR and GG) of the evidence. Taking the AB case as an example, the first cell ( $72.0 \pm 1.7$ ) represents participants' mean responses (and 95% confidence interval) to (1) RR when evaluating Jar A and (2) GG when evaluating Jar B, since Jars A and B, respectively, are the ones that are implicated by the corresponding diagnostic evidence. The next cell ( $68.6 \pm 2.6$ ) contains the mean (and 95% confidence interval) for the corresponding judgment based on the four-bead RG sample (i.e., RRRG and GGRG). Likewise, the third cell ( $29.1 \pm 1.9$ ) represents participants' responses to (1) RR when evaluating Jar B and (2) GG when evaluating Jar A, since Jars B and A, respectively, are the ones that are not implicated by the corresponding diagnostic evidence. Again, the following cell ( $28.1 \pm 2.0$ ) contains the mean for the corresponding judgment based on the four-bead RG sample (i.e., RRRG and GGRG).

Table 2 indicates that both the mixed RG and neutral BB types of nondiagnostic evidence tended to produce a dilution pattern, in which the implicated jar's judged probability decreased and the nonimplicated jar's judged probability increased when RG and BB appeared in the aggregate samples. For each of the four critical sample pairs (RR–RRRG, GG–GGRG, RR–RRBB, and GG–GGBB), the difference between the judgment based on diagnostic information only and the corresponding judgment based on a combination of diagnostic and nondiagnostic information was subjected to a 2 (diagnostic evidence: RR vs. GG)  $\times$  2 (nondiagnostic evidence: RG vs. BB)  $\times$  2 (target jar: implicated vs. nonimplicated)  $\times$  2 (case: AB vs. CD) repeated measures analysis of variance (ANOVA). In addition to a main effect of jar produced by the dilution effect [ $F(1,63) = 31.16$ ,  $MS_e = 900$ ,  $p < .01$ ], there were two statistically significant interactions. First, a relatively stronger dilution effect was obtained for the neutral (BB) than for the mixed (RG) nondiagnostic evidence, as reflected by the nondiagnostic evidence  $\times$  target jar interaction [ $F(1,63) = 10.69$ ,  $MS_e = 375$ ,  $p <$

.01]. Second, relatively stronger dilution effects were observed when RG and BB followed more highly diagnostic data (i.e., in the CD case as compared with the AB case), as reflected by the case  $\times$  target jar interaction [ $F(1,63) = 7.08, MS_e = 591, p < .05$ ]. The stronger dilution effect for the neutral than for the mixed nondiagnostic evidence was also evident when each case was analyzed separately for the AB case [ $F(1,63) = 10.46, MS_e = 201, p < .01$ ] and for the CD case [ $F(1,63) = 4.86, MS_e = 393, p < .05$ ]. No other effects were statistically significant in this analysis.

To ensure that the RG and BB samples alone were indeed viewed as equally supportive of both hypotheses, we examined judgments from all of the trials in which a RG and BB pair appeared on its own. As expected, for each case, each of the competing hypotheses was judged as approximately equally likely ( $M = 50.1\%$  for the four critical Order 2 samples).

**Ordinal analysis.** The variance in judged probability both within and between participants can be quite high, and consequently the pattern of means depicted in Table 2 may not provide a completely accurate depiction of the typical judgment pattern of the typical individual. One way of reducing such variance while still examining the key questions of interest is to focus on the direction rather than the magnitude of differences between judgments based on diagnostic evidence alone versus a combination of diagnostic and nondiagnostic evidence. Thus, in addition to examining the means, we also looked at the frequency with which participants gave less extreme judgments for the four-bead samples (e.g., RRRG and RRBB) than for the diagnostic samples alone (e.g., RR). For each of the four critical sample pairs (RR–RRRG, GG–GGRG, RR–RRBB, and GG–GGBB), we counted the number of times a participant's judgment based on the aggregate sample was higher, lower, or the same as his/her judgment based on the diagnostic sample alone. The results, collapsed across the diagnostic portion RR and GG, are reported in Table 3. Items in boldface indicate the number of trials in which the direction of the dif-

ference is consistent with the general dilution pattern of results shown by the mean data. Generally speaking, for both nondiagnostic pairs (RG and BB), judgments of the jar implicated by the evidence decrease toward 50% and judgments of the jar not implicated by the evidence increase toward 50% for the four-bead aggregate samples relative to the diagnostic pair-only counterparts. This analysis shows that the trends revealed in the pattern of mean judgments are not the result of a few participants giving extreme judgments (or of a number of participants each giving a few extreme judgments) but instead characterize most of the judgments of most participants in the experiment.

## Discussion

In a comparison of the judgments based on the diagnostic portion only (e.g., RR) with separate judgments based on the mixed sample of diagnostic and nondiagnostic evidence (e.g., RRRG, RRBB), the dilution pattern was observed. In this aggregate procedure, the presence of RG or BB in the four-bead samples reduced the impact of the diagnostic portion on the extremity of people's judgments.

One surprising aspect of the results is that the BB sample produced a stronger dilution effect than did the RG sample. This finding is surprising because, intuitively, it should have been easier to give normative judgments for the aggregate samples including the BB pair. Presumably, when individuals are presented with a sample of beads, they segregate the sample according to color. Thus, when BB is included in the four-bead samples, participants could segregate the beads into two groups. For example, RRBB was most likely perceived as two red beads and two blue beads, making salient the nondiagnostic BB pair in the sample. In contrast, when RG was the nondiagnostic pair, the samples were likely to have been segregated differently. For example, RRRG was likely to have been perceived as three red beads and one green bead, not as a diagnostic pair RR and a nondiagnostic pair RG. Segregated this way, the nondiagnosticity of the RG pair is not highlighted. The observed dilution for the BB pair is important because it suggests that even though the BB pair is perceptually segregated from the diagnostic pair in the four-bead sample, individuals were still influenced by this evidence.

## EXPERIMENT 2 Belief-Revision Procedure

Perhaps the normative requirement to ignore nondiagnostic evidence—particularly the obviously neutral BB sample—is more easily put into practice when the judgment task prompts evaluation of its implications in isolation from that of the diagnostic portion of the evidence. This possibility can be tested through the use of a belief-revision procedure. Instead of separating the sample pairs (e.g., RR–RRRG) into two blocks, we investigated whether participants would be more inclined to ignore

**Table 3**  
Frequency With Which Probability Judgment Assigned to Target Jar (Implicated or Nonimplicated) Increased, Decreased, or Remained Unchanged in Response to Nondiagnostic (Mixed RG or Neutral BB) Evidence in Experiment 1, Listed Separately for Each Case (AB and CD)

Case	ND Sample	Jar	Change in Judged Probability		
			Increase	Decrease	Same
AB	RG	implicated	44	<b>64</b>	20
		nonimplicated	<b>53</b>	58	17
	BB	implicated	29	<b>72</b>	27
		nonimplicated	<b>84</b>	31	13
CD	RG	implicated	35	<b>65</b>	28
		nonimplicated	<b>63</b>	31	34
	BB	implicated	14	<b>87</b>	27
		nonimplicated	<b>88</b>	18	22

Note—ND, nondiagnostic; R, red; G, green; B, blue. Bold entries reflect the predominant pattern of results.

the neutral (BB) and mixed (RG) types of nondiagnostic evidence if it directly followed an assessment of the diagnostic portion of the evidence. Thus, using a belief-revision procedure in Experiment 2, we examined whether individuals would alter their initial judgments (based on a piece of diagnostic information alone, e.g., RR) when presented with additional mixed (RG) or neutral (BB) nondiagnostic evidence. Troutman and Shanteau (1977) reported comparable degrees of dilution for both types of evidence but, as noted in the introduction, they did not elicit judgments of the nonimplicated hypothesis and also may not have had sufficient statistical power to detect differences between judgments based on the two types of evidence.

## Method

**Participants.** Thirty-five undergraduates enrolled in introductory psychology participated in this study for partial fulfillment of their course requirements. Data from 2 participants were excluded from the analyses for entering the same probability value for the majority of the trials in the experiment. Data from a 3rd participant were omitted because he/she was revealed to be an outlier in the accuracy analyses (on average, judgments deviated by more than .25 from the Bayesian calculation).

**Stimuli.** The stimuli were the same as in Experiment 1.

**Practice phase.** This phase of the experiment was almost identical to that in Experiment 1, except for the change from an aggregate procedure to a belief-revision procedure. As before, the experimenter drew two beads from the selected jar. After seeing the drawn beads, the participant was asked to make a probability judgment, between 0% and 100%, that the two beads were in fact drawn from the target jar indicated by the slip of paper, rather than the remaining, nontarget, jar. Two additional beads were then drawn from the same selected jar and the participant was asked to provide a revised probability judgment based on all four beads. The same sequence of events was repeated so that by the end of the practice phase, the participant had completed 2 two-step practice trials.

**Test phase.** Participants made 64 pairs of judgments corresponding exactly to the 128 aggregate judgments obtained in Experiment 1. The only difference was that each sample was now presented in two steps. As in the practice phase, after the participant provided a judgment based on the first pair of beads (e.g., RR), a second pair of beads (e.g., RG) was presented so that the participant's final judgment was based on all four beads (e.g., RRRG). Of interest were the Order 1 samples, where a diagnostic piece of information (e.g., RR) was followed by a nondiagnostic piece of information (e.g., RG or BB). We were interested in whether individuals would alter their initial judgments (based on a piece of diagnostic information alone, e.g., RR) when presented with mixed (RG) or neutral (BB) nondiagnostic evidence at Step 2. Two block orders were used: AB, CD; and CD, AB. For each block, the jars under consideration were shown on the screen and the participants were presented with two-step sequences of bead pairs, in a random order. Each jar under consideration served as the target jar for each sample at some point during each subblock, in a randomly determined fashion so that the target jar varied from trial to trial. Participants judged the probability that the sample had been drawn from the target jar (which was designated by a box drawn around it), rather than from the other, nontarget, jar. In total, 64 pairs of judgments were obtained from each participant.

## Results

**Mean judgments.** We concentrated our analysis on the same eight sample pairs as in Experiment 1 (RR–RRRG,

**Table 4**  
Mean Probability Judgment Assigned to the Target Jar (Implicated or Nonimplicated) Based on Diagnostic Evidence Alone or on Diagnostic and Nondiagnostic Evidence, Along With Their Difference ( $\pm 95\%$  Confidence Interval) and Associated Simple Effects Test, for Mixed (RG) and Neutral (BB) Nondiagnostic Evidence in Experiment 2, Presented Separately for Each Case (AB and CD)

Case	ND Sample	Jar	Judged Probability		
			D only	D+ND	Difference
AB	RG	implicated	67.0	69.4	2.4 $\pm$ 3.9
		nonimplicated	32.8	26.4	-6.4 $\pm$ 2.7**
	BB	implicated	65.1	65.2	0.1 $\pm$ 2.4
		nonimplicated	31.9	32.6	0.8 $\pm$ 2.1
CD	RG	implicated	78.2	82.3	4.1 $\pm$ 4.6
		nonimplicated	25.2	19.4	-5.8 $\pm$ 3.5**
	BB	implicated	77.6	75.2	-2.4 $\pm$ 2.6
		nonimplicated	24.9	26.5	1.7 $\pm$ 2.3

Note—D, diagnostic bead sample; ND, nondiagnostic bead sample; R, red; G, green; B, blue. \* $p < .05$ . \*\* $p < .01$ .

GG–GGRG, RR–RRBB, GG–GGBB, RG–RGRR, RG–RGGG, BB–BBRR, and BB–BBGG). With the belief-revision procedure instead of the aggregate procedure, Troutman and Shanteau's (1977) dilution effect was not replicated for either nondiagnostic pair of beads (RG and BB). Surprisingly, as shown in Table 4, we obtained a consistent pattern of results across both cases: The probability judgment for the jar that was initially favored by the diagnostic pair of beads actually increased after the nondiagnostic mixed pair RG was presented, while the probability judgments for the remaining, nonimplicated, jar decreased. We will refer to this pattern of results as *confirmation* because the initial judgments became more extreme with the introduction of the mixed evidence. In contrast, the neutral BB pair of beads did not have any impact on the original judgment for the majority of samples. We interpreted this result as indicating that the BB pair of beads was perceived as being nondiagnostic and hence was ignored.

For each of the four critical sample pairs (RR–RRRG, GG–GGRG, RR–RRBB, and GG–GGBB), the change between the initial judgment, based on diagnostic information only, and the subsequent judgment, based on a combination of diagnostic and nondiagnostic information, was subjected to a 2 (diagnostic evidence: RR vs. GG)  $\times$  2 (nondiagnostic evidence: RG vs. BB)  $\times$  2 (target jar: implicated vs. nonimplicated)  $\times$  2 (case: AB vs. CD) repeated measures ANOVA. There was a statistically significant nondiagnostic evidence  $\times$  target jar interaction [ $F(1,31) = 16.83$ ,  $MS_e = 260$ ,  $p < .01$ ], reflecting the tendency for mixed nondiagnostic evidence but not neutral evidence to produce a confirmation effect. This interaction was also statistically significant when each case was analyzed separately for the AB case [ $F(1,31) = 12.26$ ,  $MS_e = 116$ ,  $p < .01$ ] and for the CD case [ $F(1,31) = 15.36$ ,  $MS_e = 202$ ,  $p < .01$ ]. The confirmation effect for mixed evidence was sufficiently strong to produce an overall main effect of target jar [ $F(1,31) =$

9.91,  $MS_e = 159$ ,  $p < .01$ ]. No other effects were statistically significant in this analysis.

The pattern of mean judgments (and, to a lesser extent, the ordinal analysis reported next) reveals an apparent asymmetric contribution to the confirmation effect by judgments made for the nonimplicated versus the implicated jar, with the probability assigned to the nonimplicated jar decreasing by a larger amount than the probability assigned to the implicated jar increases following the introduction of nondiagnostic evidence. While we have no specific explanation for this asymmetry, it may help to explain why past research—which, as noted in the introduction, generally did not involve judgments of the nonimplicated hypothesis—has not previously documented the confirmation effect reported here.

To ensure that the RG and BB samples alone were indeed perceived as nondiagnostic, we looked at all of the trials in which a RG or BB pair appeared first. Under these circumstances, each of the competing hypotheses was judged as approximately equally likely on Step 1 ( $M = 49.5\%$  for the four critical Order 2 samples).

**Ordinal analysis.** In addition to examining the means, we also looked at the frequency with which participants altered their judgments in an upward or downward direction at Step 2, regardless of the magnitude of the revision. For each intertrial (Judgment 1 to Judgment 2) comparison for the Order 1 samples (in which the diagnostic portion of the evidence was presented first), we counted the number of times the revised judgment (relative to the initial response) increased, decreased, or remained unchanged with the introduction of the nondiagnostic RG or BB pair at Step 2. The results, again collapsed across the diagnostic portion RR and GG, are reported in Table 5. Items in boldface indicate the number of trials in which the direction of change is consistent with the general pattern of results shown by the mean data. For the nondiagnostic pair RG, judgments of the jar implicated by the evidence increase away from 50% and judgments of the nonimplicated jar decrease away from 50%. For the nondiagnostic pair BB, the majority of the judgments remain unchanged. This analysis shows that the trends revealed

in the pattern of mean judgments characterize most of the judgments of most participants in the experiment.

## Discussion

The results from Experiment 2 show that individuals' probability judgments became more extreme after the mixed (RG) evidence was presented and remained unchanged with the introduction of the neutral (BB) evidence. The mixed evidence result was in the opposite direction of that reported by Troutman and Shanteau (1977). One possible interpretation of this finding is that our participants tentatively generated a working hypothesis about which jar was more likely, given the evidence on Step 1, and then interpreted the RG pair in a biased manner that exaggerated its compatibility with the working hypothesis. Russo, Meloy, and Medvec (1998) documented a similar effect in the domain of preference formation, in which initial information supporting a particular choice option leads to distorted impressions of later information biased in favor of the initially supported option. Although our experiment does not directly correspond to the studies reported in the confirmation bias literature (for a review, see Nickerson, 1998), the standard interpretation of such studies is consistent with our finding for the mixed (RG) pair: People are more sensitive to, or give greater weight to, evidence that confirms their currently held beliefs. When evaluating mixed evidence, on this account, participants first generated a hypothesis about which jar was most likely given the initial evidence, and then they interpreted new evidence in a biased, confirmatory manner. The process underlying this effect is not clear. One possibility is that participants evaluated each bead separately. Thus, the one bead (e.g., R) in the RG pair that on its own favors the most likely jar (e.g., the predominantly red jar, in this case) was given greater weight, relative to its counterpart (e.g., G) when RG followed a RR pair than when it followed a GG pair. Further studies are required in order to identify more clearly the process underlying this effect.

Apparently, the belief-revision procedure used in Experiment 2 blocked the dilution effect otherwise obtained using the aggregate procedure in Experiment 1. This suggests that dilution occurs because individuals do not spontaneously segregate the evidence (e.g., RRRG) into its diagnostic and nondiagnostic components when it is presented in an aggregate format. When the evidence is segregated in a manner that requires separate evaluation of the diagnostic and nondiagnostic portions of the evidence (using a belief-revision procedure), individuals can readily identify the neutral BB sample as being nondiagnostic, thus leaving their initial judgment unchanged. In contrast, it may be less intuitively obvious, even with this segregation, that the mixed RG sample can be safely ignored in the revised judgment; instead, the mixed evidence apparently increases individuals' confidence in their working hypothesis under such circumstances, producing a pattern opposite of the dilution effect, which we have called a confirmation effect.

Table 5

Frequency With Which Probability Judgment Assigned to Target Jar (Implicated or Nonimplicated) Increased, Decreased, or Remained Unchanged in Response to Nondiagnostic (Mixed RG or Neutral BB) Evidence in Experiment 2, Listed Separately for Each Case (AB and CD)

Case	ND Sample	Jar	Change in Judged Probability		
			Increase	Decrease	Same
AB	RG	implicated	<b>35</b>	17	12
		nonimplicated	6	<b>45</b>	13
	BB	implicated	11	9	<b>44</b>
		nonimplicated	11	6	<b>47</b>
CD	RG	implicated	<b>36</b>	12	16
		nonimplicated	9	<b>39</b>	16
	BB	implicated	9	15	<b>40</b>
		nonimplicated	13	9	<b>42</b>

Note—ND, nondiagnostic; R, red; G, green; B, blue. Bold entries reflect the predominant pattern of results.

## FOLLOW-UP STUDIES

We conducted two follow-up studies that established the generalizability of the results from Experiment 2 over various changes in how the evidence is presented to participants, how the response scale is represented, and the anchor value for the response at Step 2 of the belief-revision procedure. In light of the inconsistency between our results in Experiment 2 and the results of Troutman and Shanteau's (1977) original study, despite their very similar experimental design, we designed the follow-up studies to investigate three seemingly minor ways in which our experimental procedure differed from theirs.

First, in the Troutman and Shanteau (1977) study, participants recorded on a sheet of paper (using letters to represent colors) the colors of the beads drawn in the first sample, which they referred to in revising their judgment when the second sample was drawn; by contrast, in our study, colored circles presented on the computer screen were used to represent the results of both draws. It is conceivable that the letter-based representation of the first drawn led participants in Troutman and Shanteau to place greater weight on the more salient second bead sample when making their revised judgments (see Shanteau, 1975). To test this, in our first follow-up study ( $N = 64$ ), one group of participants was presented with a belief-revision task in which the results of the first bead sample were represented by letters rather than colored circles when the results of the second draw were presented at Step 2; another group was presented with the same representation (colored circles) used in Experiment 2. Results indicate that this factor exerted no systematic influence on people's judgments [ $F(1,60) = 0.60$ ,  $MS_e = 509$ ].

Second, in the Troutman and Shanteau (1977) study, the sliding response marker used to elicit judgments was reset to the midpoint of the response scale prior to eliciting the revised judgment in Step 2; by contrast, in our experiment, the initial judgment at Step 1 served as the starting point or anchor for the revised judgment elicited at Step 2. It is possible that Troutman and Shanteau's procedure induced an anchoring effect drawing Step 2 judgments toward the center of the scale (but see Shanteau, 1975, Experiment 2) and in this manner produced or enhanced the dilution effect. Our first follow-up study also investigated this possibility, by comparing Step 2 judgments anchored at the response scale midpoint (i.e., 50%) with Step 2 judgments anchored at the initial Step 1 judgment. Results indicate that this factor also exerted no systematic influence on people's judgments [ $F(1,60) = 0.65$ ,  $MS_e = 509$ ].

Third, Troutman and Shanteau (1977) used what might be called a "complementary" response scale representation in which the 100% end of the scale was labeled as "definitely from the (predominantly) white jar" and 0% was labeled as "definitely from the (predominantly) red jar." By contrast, our experiments used a noncomplementary scale representation in which 100% was labeled

as "definitely from the target jar" and 0% was labeled as "definitely not from the target jar." Scale representation could conceivably influence how the evidence is interpreted (e.g., McKenzie, 1998). To test this possibility, in a second follow-up study ( $N = 32$ ), we compared judgments made on the same labeled response scale used in Experiments 1 and 2 with judgments made on a "complementary" response scale in which each end of the scale was associated with one of the competing hypotheses (i.e., jars; the 100% end of the scale was labeled with a picture of the target jar, and the 0% end of the scale was labeled with a picture of the alternative jar). Results indicate that this factor did not exert a systematic influence on people's judgments [ $F(1,30) = 0.47$ ,  $MS_e = 330$ ].

Because the results from the follow-up studies exhibited a similar pattern of judgments across the various task manipulations we conducted, we can collapse the results across these manipulations and assess whether the general pattern of results obtained in Experiment 2 are replicated. The ordinal analysis (Table 6) of the direction of changes in judgments following the introduction of nondiagnostic evidence shows a pattern of results similar to that of Experiment 2. For the mixed evidence pair RG, responses to the jar implicated by the evidence increases away from 50% and those to the jar not implicated by the evidence decreases away from 50%, replicating the confirmation effect observed in the previous experiment. For the neutral evidence pair BB, the majority of the responses remained unchanged from those based on the diagnostic portion of the evidence alone.

As with the data from the previous experiment, the difference between the initial judgment based on diagnostic information and the revised judgment based on additional nondiagnostic information was subjected to a 2 (diagnostic evidence: RR vs. GG)  $\times$  2 (nondiagnostic evidence: RG vs. BB)  $\times$  2 (target jar: implicated vs. non-implicated)  $\times$  2 (case: AB vs. CD) repeated measures ANOVA. Once again, there was a statistically significant nondiagnostic evidence  $\times$  target jar interaction [ $F(1,95) =$

**Table 6**  
**Frequency With Which Probability Judgment Assigned to Target Jar (Implicated or Nonimplicated) Increased, Decreased, or Remained Unchanged in Response to Nondiagnostic (Mixed RG or Neutral BB) Evidence in Collapsed Data From the Two Follow-up Studies, Listed Separately for Each Case (AB and CD)**

Case	ND Sample	Jar	Change in Judged Probability		
			Increase	Decrease	Same
AB	RG	implicated	<b>97</b>	53	42
		nonimplicated	47	<b>107</b>	38
	BB	implicated	32	47	<b>113</b>
		nonimplicated	59	21	<b>112</b>
CD	RG	implicated	<b>91</b>	55	46
		nonimplicated	64	<b>93</b>	35
	BB	implicated	27	61	<b>104</b>
		nonimplicated	62	32	<b>98</b>

Note—ND, nondiagnostic; R, red; G, green; B, blue. Bold entries reflect the predominant pattern of results.

28.67,  $MS_e = 175$ ,  $p < .01$ ], reflecting the tendency for mixed nondiagnostic evidence but not neutral evidence to produce a confirmation effect. This interaction was also statistically significant when each case was analyzed separately for the AB case [ $F(1,95) = 28.52$ ,  $MS_e = 128$ ,  $p < .01$ ] and for the CD case [ $F(1,95) = 15.42$ ,  $MS_e = 212$ ,  $p < .01$ ]. Examination of the mean differences again reveals a confirmation effect induced by exposure to the mixed (RG) evidence, though the magnitude of this effect was smaller than that observed in Experiment 2. On average, exposure to the neutral (BB) evidence produced a trend toward dilution, even though the ordinal analysis revealed that the majority of the judgments remained unchanged following its introduction; of the minority of judgments that were changed in response to the BB pair, however, about two thirds were in the direction of dilution rather than confirmation, thus pulling the means in this direction as well (Table 6).

## GENERAL DISCUSSION

Results of the present studies indicate that the influence of nondiagnostic evidence depends on the type of judgment task in which it is encountered. We investigated the impact of two types of nondiagnostic bodies of evidence: (1) neutral evidence, in which each individual component fails to discriminate among the competing hypotheses; and (2) mixed evidence, consisting of individually diagnostic components that, taken together, fail to discriminate among the competing hypotheses. Both types of nondiagnostic evidence produced dilution in the aggregate judgment procedure of Experiment 1. By contrast, they produced divergent effects in the belief-revision judgment procedure used in the second experiment, with neither type of nondiagnostic evidence producing dilution. Instead, the neutral evidence was effectively ignored, as required by probability theory, while the mixed evidence produced an apparent confirmation effect so that judgments became more extreme (rather than less extreme, as in the dilution effect) with the introduction of the mixed nondiagnostic evidence. Results from two follow-up studies largely replicated the results of the second experiment and ruled out several procedural variables as the cause of our inability to replicate the results of Troutman and Shanteau (1977). Although further research is clearly required, we conclude with a few speculations regarding the judgmental processes underlying the pattern of results observed in our studies.

The key distinction between the aggregate and belief-revision procedures is that the nondiagnostic portion of the evidence is segregated from the diagnostic portion of the evidence in the latter but not in the former. As a result, only in the belief-revision procedure is there an explicit prompt to evaluate separately the implications of the nondiagnostic portion of the evidence in terms of how it should modify one's conclusions based on the previously introduced diagnostic portion of the evidence.

When segregated in this manner, neutral and mixed nondiagnostic evidence appear to be interpreted as having quite different implications, even though both produce dilution in the aggregate judgment procedure.

When evaluated in isolation, both the mixed bead sample RG and the neutral bead sample BB were rated by the majority of our participants as providing equal support for the competing hypotheses (i.e., jars), in that both were assigned approximately equal probabilities of very close to 50%. This observation, however, does not necessarily imply recognition of the general statistical principle that either nondiagnostic sample can be safely ignored in rendering a judgment. Indeed, if people recognized and applied that principle in a consistent fashion, then nondiagnostic evidence would not have any systematic influence on their judgments. We suggest that the judge will attempt to apply this principle only when the nondiagnostic portion of the evidence is (1) segregated from the diagnostic portion of the evidence and then (2) classified as information that can be ignored. Even under those conditions, the judge will be able to render judgments that are invariant over the presence or absence of the nondiagnostic portion of the evidence only if he/she is actually able to ignore the nondiagnostic evidence—that is, avoid having it influence his/her judgment, which other research suggests is often a difficult goal to achieve (e.g., Wilson & Brekke, 1994).

Our previous discussion implies that condition (1) is more likely to be met in the belief-revision procedure than in the aggregate judgment procedure. Likewise, on the basis of our earlier discussion of the relative ease with which the nondiagnosticity of neutral evidence can be detected compared with that of mixed evidence, it seems that condition (2) is more likely to be met in the case of neutral evidence than in the case of mixed evidence. Taken together, then, we might expect nondiagnostic evidence to be successfully ignored when neutral evidence is evaluated in a belief-revision task, which is indeed the only set of circumstances under which nondiagnostic evidence was observed to have no systematic impact on judgments in our investigation. Once the judge decides to ignore the neutral evidence, actually doing so is straightforward in the belief-revision procedure because the previous judgment made on the basis of the diagnostic portion of the evidence alone is readily available.

What happens when condition (1) is not met—that is, when the nondiagnostic portion of the evidence is not segregated from the diagnostic portion of the evidence? If no attempt is made to selectively disregard or ignore certain components of a body of evidence, then the judge must integrate the implications of each into a final assessment of the extent to which the body of evidence as a whole supports a particular hypothesis relative to its competitors. A number of researchers have developed models to describe the integration process, many of which adopt an averaging method of integration (e.g., Anderson, 1981; Hogarth & Einhorn, 1992). If we assume that

each component in the neutral body of evidence is assigned a value of zero (e.g., 0 for each B in Bead Sample BB), and the conflicting components in the mixed body of evidence are assigned values with opposite sign (e.g., +1 for R and -1 for G in Bead Sample RG), then both types of nondiagnostic evidence would produce dilution under an averaging integration method. In short, as we show in more detail in the Appendix, an averaging integration process can produce equivalent results (namely dilution) for neutral and mixed nondiagnostic evidence even when the implications of their individual components are encoded quite differently from each other, leaving open the possibility that they may not exert equivalent effects in some other judgment task. In the belief-revision task, for example, even when the nondiagnostic evidence is segregated, the judge may classify only zero-valued evidence components as ignorable.

What happens when condition (2) above is not met—that is, when (as in the case of mixed evidence) the nondiagnostic portion of the evidence is not classified as safe to ignore even though it has been segregated from the diagnostic portion of the evidence? In the belief-revision task, the nondiagnostic portion of the evidence is segregated via elicitation of an initial judgment based exclusively on the diagnostic portion of the evidence. We have suggested that the initial judgment may establish a working hypothesis that influences how subsequently encountered evidence is interpreted (assuming that the new evidence is not deemed to be ignorable). Thus mixed evidence that is seen as being neutral in its implications when evaluated in isolation, and that exerts a diluting influence when integrated with diagnostic evidence in an aggregate judgment procedure, may be seen as having quite different implications when evaluated in light of an established working hypothesis in the belief-revision procedure. Under these conditions, the judge may be subject to a confirmatory bias (e.g., Nickerson, 1998; cf. Russo et al., 1998) in which the evidence is evaluated in a biased manner that favors an established working hypothesis. Thus the mixed RG bead sample might be viewed as favoring the predominantly red jar if previously encountered evidence has already implicated that jar.

We show in the Appendix that this result is not easily captured using a standard information integration model (Hogarth & Einhorn's influential 1992 belief-adjustment model) of either the averaging or additive variety. One possible modification of such models, consistent with

our interpretation of the observed confirmation effect, is to allow the subjective evaluation of the implications of a piece of evidence to depend on what evidence has already been encountered. Further research into the determinants and boundary conditions of the dilution effect can help to shed additional light onto the processes underlying evidence-based judgment.

## REFERENCES

- ANDERSON, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.
- HACKENBRACK, K. (1992). Implications of seemingly irrelevant evidence in audit judgment. *Journal of Accounting Research*, **30**, 126-136.
- HOGARTH, R., & EINHORN, H. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, **24**, 1-55.
- MCKENZIE, C. R. M. (1998). Taking into account the strength of an alternative hypothesis. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 771-792.
- MEYVIS, T., & JANISZEWSKI, C. (2002). Consumers' beliefs about product benefits: The effect of obviously irrelevant product information. *Journal of Consumer Research*, **28**, 618-635.
- NICKERSON, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, **2**, 175-220.
- NISBETT, R., ZUKIER, H., & LEMLEY, R. (1981). The dilution effect: Nondiagnostic information weakens the implications of diagnostic information. *Cognitive Psychology*, **13**, 248-277.
- RUSO, J. E., MELOY, M. G., & MEDVEC, V. H. (1998). Predecisional distortion of product information. *Journal of Marketing Research*, **35**, 438-452.
- SHANTEAU, J. (1975). Averaging versus multiplying combination rules of inference judgments. *Acta Psychologica*, **39**, 83-89.
- SHELTON, S. W. (1999). The effect of experience on the use of irrelevant evidence in auditor judgment. *Accounting Review*, **74**, 217-224.
- SLUGOSKI, B. R., & WILSON, A. E. (1998). Contribution of conversation skills to the production of judgmental errors. *European Journal of Social Psychology*, **28**, 575-601.
- SMITH, H. D., STASSON, M. F., & HAWKES, W. G. (1999). Dilution in legal decision making: Effect of non-diagnostic information in relation to the amount of diagnostic evidence. *Current Psychology*, **17**, 333-345.
- TETLOCK, P. E., LERNER, J. S., & BOETTGER, R. (1996). The dilution effect: Judgmental bias, conversational convention, or a bit of both? *European Journal of Social Psychology*, **26**, 915-934.
- TROUTMAN, C., & SHANTEAU, J. (1977). Inferences based on nondiagnostic information. *Organizational Behavior & Human Performance*, **19**, 43-55.
- WILSON, T. D., & BREKKE, N. C. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, **116**, 117-142.
- ZUKIER, H. (1982). The dilution effect: The role of the correlation and the dispersion of predictor variables in the use of nondiagnostic information. *Journal of Personality & Social Psychology*, **43**, 1163-1174.
- ZUKIER, H., & JENNINGS, D. (1983). Nondiagnosticity and typicality effects in prediction. *Social Cognition*, **2**, 187-198.

## APPENDIX

Here we attempt to account for the following observations or “stylized facts” from the experiments reported in this paper, using Hogarth and Einhorn’s (1992) belief-adjustment model. See Table 1 for a description of the beads and jars task used in the experiments; note that R = red, G = green, and B = blue bead.

1. RR produces judgments favoring the predominantly red jar. (GG likewise produces judgments favoring the predominantly green jar, but we can focus without loss of generality on RR as the diagnostic bead sample.)
2. Both BB and RG produce 50% judgments for either jar when presented in isolation.
3. Both RRBB and RRRG, in the aggregate procedure, produce less extreme judgments than those based on RR alone (dilution effect).
4. A more pronounced dilution effect is produced in the aggregate procedure when RR has higher diagnosticity (in the CD case vs. the AB case).
5. The RRBB sample produces a more pronounced dilution effect than does the RRRG sample in the aggregate procedure.
6. In the belief-revision procedure, RR then BB produces judgments that are no more or less extreme than those based on RR alone (no dilution or confirmation effect).
7. In the belief-revision procedure, RR then RG produces more extreme judgments than those based on RR alone (confirmation effect).

### Hogarth and Einhorn (1992) Belief-Adjustment Model, General Form

The model specifies how strength of belief in a hypothesis is updated after one encounters new evidence, as follows:

$$S_k = S_{k-1} + w_k [s(x_k) - R],$$

where  $S_k$  = degree of belief in hypothesis after evaluating  $k$  pieces of evidence ( $0 \leq S_k \leq 1$ );  $S_0$  = initial degree of belief;  $s(x_k)$  = subjective evaluation of  $k$ th piece of evidence; and  $R$  = reference point against which impact of  $k$ th piece of evidence is evaluated. The adjustment weight  $w_k$  for  $k$ th piece of evidence ( $0 \leq w_k \leq 1$ ) is defined as follows:

$$w_k = \alpha S_{k-1} \text{ when } s(x_k) \leq R,$$

with  $\alpha$  ( $0 \leq \alpha \leq 1$ ) representing sensitivity to “negative” evidence;

$$w_k = \beta(1 - S_{k-1}) \text{ when } s(x_k) > R,$$

with  $\beta$  ( $0 \leq \beta \leq 1$ ) representing sensitivity to “positive” evidence.

Evidence may be evaluated in either a “step-by-step” (SbS) manner, with belief adjusted sequentially following separate subjective evaluation of each new piece of evidence, or in an “end-of-sequence” (EoS) manner, with belief adjusted following subjective evaluation of an entire sequence of evidence taken as a whole.

The general form of the belief-adjustment model has two alternative implementations, depending on the nature of the judgment task and the type of evidence under evaluation:

$$\text{Estimation (averaging): } R = S_{k-1}; 0 \leq s(x_k) \leq 1.$$

$$\text{Evaluation (additive): } R = 0; -1 \leq s(x_k) \leq 1.$$

Hogarth and Einhorn (1992) suggested that the evaluation form is used when one is expressing belief on a dichotomous (“true–false”) scale using bipolar evidence encoded in terms of whether it is for or against the hypothesis, and that the estimation form is used when one is expressing belief on a continuous (“how much”) scale using unipolar evidence. This description would seem to imply that the probability judgment task used in the present experiments is best implemented using the estimation form of the model. We implement that form here, but have found the same general conclusion follows when the evaluation form is implemented instead.

Note that the model takes the subjective evaluation of a piece of evidence,  $s(x_k)$ , as input. To implement the model, then, we need to specify how this variable is determined from the composition of the jars in our task. We begin by imposing some constraints on this variable and then establish that the qualitative pattern of results produced by the model does not depend on the specific values of  $s(x_k)$  as long as they obey these constraints.

### Initial Assumptions

We use the subscripts  $R$  and  $G$  on the strength of belief and subjective evaluation functions  $S$  and  $s$  to refer to belief in the hypothesis that the beads were drawn from the predominantly red jar and the predominantly green jar, respectively. We make a number of initial assumptions to implement the belief-adjustment model and later explore whether relaxing one or more of these assumptions is necessary for the model to reproduce the effects in question.

## APPENDIX (Continued)

Our initial assumptions are as follows:

1.  $S_{R0} = S_{G0} = 0.5$ ; prior to encountering any bead samples, the two hypotheses (jars) are seen as equally likely.
2. The value of  $s$  for a bead pair (or a four-bead sequence in the aggregate task) is given by the (equal-weighted) average value of  $s$  for its component beads.
3.  $s_R(RR) > s_R(RG) > s_R(GG)$  and  $s_R(RR) > s_R(BB) > s_R(GG)$ , and vice versa for  $s_G$ .
4. Four-bead sequences receive EoS processing in the aggregate judgment task, and Sbs processing (i.e., of sequential bead pairs) in the belief-revision task.
5. Judged probability reflects the (normalized) balance of belief strength for the competing hypotheses; that is,  $p(\text{predominantly red jar}) = S_R / (S_R + S_G)$ .

Assumption 1 is consistent with the observation that nondiagnostic bead pairs, when evaluated in isolation, led to equal probability judgments for the two jars. Assumption 2 implies that each bead contributes independently to the subjective evaluation of the sequence; the equal-weighting assumption reflects the fact that the order of the individual beads within a pair was counterbalanced and the results collapsed over this variable. Assumption 3 is consistent with the general ordering of judged probabilities for each target jar being observed to follow the jar's composition in terms of R, G, and B beads. Given that the aggregate and belief-revision procedures produce different results, some sort of processing difference between them must be captured in the model implementation, and Assumption 4 seems like the most obvious way to do this. Assumption 5 is probably too strong, since it is only one of any number of possible mappings from belief strength to judged probability, but is convenient; the qualitative pattern of judgments produced by the model depends only on the use of a mapping that increases monotonically with strength of belief in the focal hypothesis and decreases monotonically with strength of belief in the alternative hypothesis, with equal weighting.

### Belief-Adjustment Model: Estimation Form

In the estimation (or averaging) form of the belief-adjustment model, subjective evaluations of evidence strength vary from 0 to 1. Given Assumption 1 above, it is necessary to set  $s_R(B) = s_G(B)$  and  $s_R(R) + s_R(G) = s_G(R) + s_G(G)$  to ensure that probabilities of .5 are assigned to both jars when either nondiagnostic bead pair (RG or BB) is evaluated on its own. We begin, more specifically, by setting  $s_R(R) = s_G(G) = 0.7$ ,  $s_R(G) = s_G(R) = 0.3$ , and  $s_R(B) = s_G(B) = 0.5$  (the midpoint of the  $s$  scale in the estimation form of the model). We then compute the values of  $s$  for the bead pairs and the four-bead sequences in the aggregate task as the equal-weighted average value of  $s$  of the component beads.

Having determined the values of  $s$  for the bead samples serving as evidence in the task, we then implement the belief-adjustment model starting with the initial assumptions above, and setting  $\alpha = \beta = 1$ . Under these conditions, the estimation form of the model produces a dilution effect of equal magnitude for both types of nondiagnostic evidence, which is more pronounced for the belief-revision task than it is for the aggregate task. Furthermore, the magnitude of the dilution effect increases with the difference between  $s_R(R)$  and  $s_R(G)$ , consistent with Effect 4 above. This implementation of the model, then, manages to reproduce the first four effects listed above, but fails to capture the observed effects of judgment task (belief revision vs. aggregate) and type of nondiagnostic evidence (mixed RG vs. neutral BB).

We can adjust the initial values of  $s$  associated with each bead in an attempt to reproduce the observation that the dilution effect is more pronounced in the aggregate judgment task for neutral evidence (BB) than for mixed evidence (RG). Specifically, increasing  $s_R(B) = s_G(B)$  to a value greater than .5 yields a more pronounced dilution effect (for both the aggregate and belief revision tasks) for BB than for RG. Under this assumption, the model is able to reproduce the first five observed effects listed above.

The principal difficulty encountered in this model implementation, however, is with regard to the belief-revision task. As noted above, this version of the model produces a consistent dilution effect for the belief-revision task, contrary to our findings. In fact, the model produces a dilution effect that is consistently larger than that produced for the aggregate task, when in fact we found a dilution effect in the aggregate task but not in the belief-revision task. One way to reduce the magnitude of the dilution effect produced by the model for the belief-revision task is to decrease  $\beta$  (sensitivity to negative evidence); as  $\beta$  approaches zero, so does the magnitude of the dilution effect produced for the belief-revision task. The magnitude of the dilution effect produced by the model in the aggregate task also decreases, but not as quickly and not all the way to zero. By setting  $\beta$  to a near-zero value, then, the model can reproduce the first six of the seven observed effects listed above. This version of the model, however, is unable to reproduce the observation that judgments in the belief-revision task became more extreme following the introduction of mixed (RG) evidence, which we have called a confirmation effect. In fact, there is no way for the model to reproduce this effect without relaxing one or more of our initial assumptions. Under these assumptions, we were also unable to produce a confirmation effect using the evaluation form of the belief-adjustment model.

---

**APPENDIX (Continued)**

---

**Conclusion**

Neither form of the belief-adjustment model could reproduce the full set of observed effects given our initial assumptions. The confirmation effect observed for mixed evidence in the belief-revision task, in particular, proved difficult for the model to produce without also producing other effects that were inconsistent with the observed results (e.g., that both RG and BB were judged to be nondiagnostic when evaluated on their own).

How might our initial assumptions be relaxed so that the model could be able to reproduce the full pattern of observed effects? Assumption 2 would seem to be the most obvious candidate. It associates a unique value of  $s$  with each constituent bead in the context of particular jar, and in this sense assumes that the contribution of a particular bead to the value of  $s$  for a bead pair is independent of all other evidence (i.e., beads) that has been encountered. A natural modification, consistent with our interpretation of the observed confirmation effect, is to allow the subjective evaluation of a piece of evidence to depend on what evidence has already been encountered. For example, after encountering RR, which strongly implicates the predominantly red jar, subsequent assessment of RG would treat both R and G as more supportive of the predominantly red jar (and less supportive of the predominantly green jar) than after encountering GG.

---

(Manuscript received March 31, 2003;  
revision accepted for publication February 24, 2004.)