Psychonomic Bulletin & Review
2007, 14 (6), 1147-1152

Intuitive *t* tests: Lay use of statistical information

Natalie A. Obrecht, Gretchen B. Chapman, and Rochel Gelman
Rutgers University, Piscataway, New Jersey

Normatively, a statistical pairwise comparison is a function of the mean, standard deviation (*SD*), and sample size of the data. In our experiment, 203 undergraduates compared product pairs and judged their confidence that one product was better than the other. We experimentally manipulated (within subjects) the average product ratings, the number of raters (sample size), and the *SD* of the ratings. Each factor had two levels selected, so that the same change in statistical power resulted from moving from the low to the high level. We also manipulated (between subjects) whether subjects were given only the product rating data as summarized in a statistical format or the summaries plus the raw ratings. Subjects gave the most weight to mean product ratings, less weight to sample size, and very little weight to *SD*. Providing subjects with raw data did not increase their use of sample size and *SD*, as predicted.

Laypeople frequently make comparisons. For example, they compare medical treatments, consumer products, and election candidates. In many cases, these comparisons are informed by information sources, such as doctors, *Consumer Reports*, or political polls. Surgery may be more promising than chemotherapy, a Volvo may be rated better than a Saab, and a candidate may be favored by 52% of those polled.

Sometimes one is advised of the best option and thus does not need to make an inference. But frequently, laypeople have access to statistical data to inform their choices. Online shopping Web sites have long provided potential buyers with the average rating that a product has been given by previous consumers. Recently, many Web sites have also made available the number of consumer ratings on which such an average is based, and a breakdown of those individual ratings, which allow consumers to see the variability of the rating information. In a statistical sense, this allows people who are comparing products to move from a simple comparison of group means to an evaluation that also includes sample size and variability. Given that such statistical data are becoming more widely available to laypeople, it is important to examine the extent to which people are able to use such information. Using sample size and variability information is important when people make comparisons that determine a costly action. For example, a Volvo may have a higher mean consumer rating than a Saab, but the consumer needs to know whether that difference is reliable or not before deciding whether to pay appreciably more for the Volvo.

Previous research has presented a mixed view of human statistical intuition. Much of the judgment and decision-making literature indicates that people fail to use relevant information such as sample size. For example, the majority of Kahneman and Tversky's (1972) subjects reported that a large hospital will have as many or more days in a year as a small hospital on which at least 60% of the babies born are male. Tversky and Kahneman (1974) concluded that this finding shows that people are insensitive to sample size. However, Nisbett, Krantz, Jepson, and Kunda (1983) demonstrated that people do have some understanding about sample size. Their subjects made stronger inferences about characteristics of a population when they were given data from a larger, in comparison with a smaller, sample. Additionally, Nisbett et al. reported that people are more willing to generalize from a sample to a population when the domain is assumed to have little variability (e.g., properties of a chemical element) than when it is thought to have much variability (e.g., obesity in a group of people). For a discussion regarding these divergent findings, see Sedlmeier and Gigerenzer (1997).

In the present research, we focus on lay intuitions about pairwise comparisons. Consumers are often presented with a choice between two products, along with former customer ratings of these. Following this approach, we asked subjects to compare pairs of products and report how confident they were that one product was better than the other. Subjects had three pieces of information about each product: the number of people who had rated the product, the average product rating, and the standard deviation (*SD*) of those ratings. Each of these factors had two levels (high, low), so that changing the level of any one factor resulted in the same change in statistical power as did changing the level of another factor (see Table 1; power calculations from Lenth, 2006).

We compared subjects' responses with the normative benchmark of statistical power. In the present study, power is defined as the probability of rejecting the null hypothesis, given that samples of size *N* are repeatedly drawn from populations with means and *SD*s identical to those of the samples provided to subjects. Power is not being put forth as a descriptive model. That is, we do not expect laypeople to perform power calculations. Nevertheless, by

N. A. Obrecht, natalie@ruccs.rutgers.edu

comparing subjects' responses to this benchmark, we can assess whether the psychological mechanisms underlying lay intuitions incorporate all three relevant factors (mean difference, sample size, and *SD*) and whether each of the three factors is given similar weight. Two other normative standards—*t* statistics, which are closely related to power, and corrected likelihood ratios, a Bayesian alternative (see Glover & Dixon, 2004)—also incorporate these three factors and produce benchmarks highly similar to power (see Table 1). Thus, regardless of which of these normative benchmarks is considered, an ideal subject would integrate not only mean difference, but also sample size and *SD* data into her judgments.

The experiment presented here manipulated how sample size, mean difference, and *SD* were presented. One group of subjects viewed these summary data on number lines, whereas another group saw the same statistical summaries plus the raw rating data, as shown in Figure 1. Gigerenzer (2000) has pointed out that humans and other animals evolved using raw data (e.g., 2 out of 10), not statistical summaries (e.g., 20%) to make decisions. He and his colleagues have demonstrated that a number of judgment biases are reduced or disappear when participants are provided with raw frequency data rather than statistical summaries (Gigerenzer, 2000; Gigerenzer & Edwards, 2003; Gigerenzer & Hoffrage, 1995; Hertwig & Gigerenzer, 1999; Hoffrage & Gigerenzer, 1998; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Sedlmeier & Gigerenzer, 2001). Therefore, in the present study, we hypothesized that subjects provided with the raw rating data would be able to assess the frequencies of each rating value for each product, which may improve encoding of sample size and variance information.

SD is a difficult concept for laypeople. For one thing, it is inversely related to statistical power, so that larger *SD*s translate into lower power. This is in contrast to mean difference and sample size, where larger values translate into greater power. For another, the units in which it is expressed (the square root of the mean squared deviation from the mean) are unfamiliar to most laypeople. Consequently, we provided the numerical presentations on number lines. As shown in Figure 1, the endpoints of the *SD* number lines were labeled *high rater agreement* and *low rater agreement*. Means and sample sizes were presented on comparable number lines labeled from *lowest rating* to *highest rating* and from *no raters* to *many raters*.

In the summary-only condition, we expected that subjects' confidence ratings would be most strongly related to the difference between product ratings, moderately related to the number of raters, and only weakly related to *SD*. We predicted that subjects in the summary + raw data condition would integrate sample size and *SD* information into their confidence ratings more successfully than those receiving only the summary information.

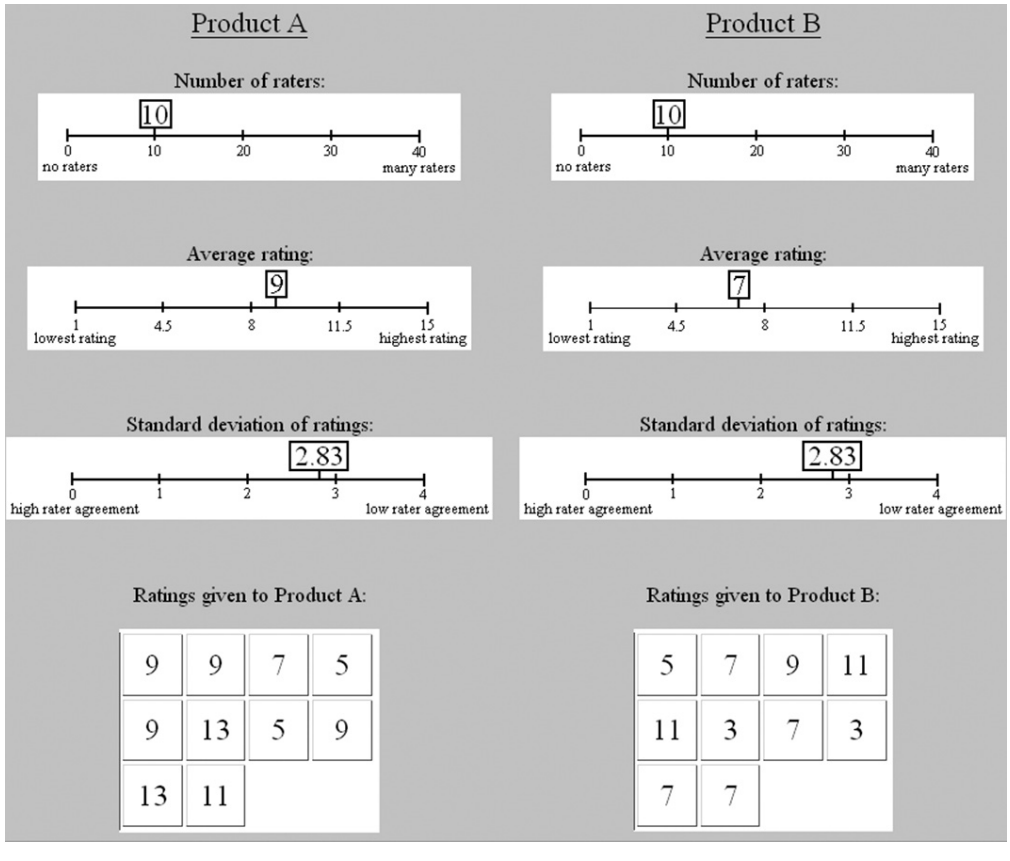

Figure 1. The Web page display used in the summary + raw data condition. The summary-only condition looked identical but without the raw rating data displays.

METHOD

Subjects

College students ($N = 203$) participated for course credit.

Materials

This experiment was Web based. An initial Web page introduced subjects to the idea of comparing two products on the basis of consumer rating information. The concepts of sample size, mean product rating, and *SD* were briefly described. Sample size was defined as the number of individuals who rated a product. The mean rating was described as the average of the product ratings that consumers gave. Finally, the *SD* was described as a measure of the amount of agreement among individual ratings, so that low *SD*s corresponded to high rater agreement. The instructions explained that a given product could be given a rating from 1 to 15. We provided two example products (Product Y and Product Z) and their ratings to demonstrate extremely high and low *SD*s on this scale. Thus, subjects saw that when a product was given the same rating by each consumer, the *SD* of the ratings was zero, but when the ratings were very different, the *SD* was a large value. Understanding of these examples was later assessed using check questions.

As a way of introducing subjects to the presentation format that would be used in the experiment, we presented an example Web page. This Web page displayed the previously introduced products, Products Y and Z. Subjects were given the number of raters, the average rating, and the *SD* of the ratings for each product. Each piece of information was displayed on a labeled number line. Subjects in the summary + raw data condition also saw the individual product ratings displayed. This format for presenting pairs of products was used in the remainder of the experiment (as shown in Figure 1). Subjects were required to answer check questions about Products Y and Z using the information given, and were not able to proceed with the experiment until they had answered correctly. These questions required subjects to read sample size and mean rating values off of the number lines corresponding to the two products. Also, subjects had to compare the *SD*s of the rating data for the two products and indicate which product had an *SD* that indicated higher rater agreement.

After completing the check questions, subjects compared eight different pairs of products for which sample size, mean rating, and *SD* information was given. We manipulated the order of presentation of the eight pairs by using two presentation orders. Each product was simply given an abstract name (e.g., Product A); no further information was provided. The two products within a pair always had the same sample size and *SD*; these factors were manipulated between pairs. Subjects indicated how confident they were that the product with the higher mean rating was better than the comparison product by judging three statements (e.g., *How confident are you that Product X is better than Product Y?*). Each statement was followed by a 5-point scale ranging from *not at all confident* to *extremely confident*. As explained below, a summary score based on these three questions served as our dependent measure. In a previous study (Obrecht & Chapman, 2006), a measure of how much more subjects would be willing to pay for the product with the higher mean rating yielded the same results as the confidence measure; thus, it was not used in this study.

Design

The eight comparison pairs resulted from a 2 (sample size) × 2 (mean difference) × 2 (*SD*) within-subjects design, whereby each variable had a high and low value. The values for these three variables were based on *t*-test statistic and post hoc power calculations, so that only four power values resulted from the eight combinations (see Table 1).

The stimuli used in the present study varied sample size, mean difference, and *SD* orthogonally. When all three of these variables favored statistical power—that is, when the sample size was large ($n = 37$ per group), the difference between the average ratings was high ($\bar{X}_1 - \bar{X}_2 = 2$), and *SD*s were low (1.41)—power was 1.0, or 100%. When the levels of these factors instead detracted from statistical power—that is, when sample size was small ($n = 10$ per group), mean difference was small ($\bar{X}_1 - \bar{X}_2 = 1$), and *SD* was large (2.84)—statistical power dropped to 12%. The three possible combinations, with two variables favoring power, and one variable not favoring power, all yielded equivalent power. For example, when subjects compared two products that had a large difference between their mean ratings, small rating variances, but few raters, the power to find a difference between these products was equal to that when the difference between mean product ratings was small, but *SD* was small and sample size was large. Also, power was equivalent in the three pairs in which only one of the variables was in the direction favoring power and the other two were not. This design allowed us to compare subjects' confidence in product differences to statistical power calculations. Specifically, if subjects' intuitions are in line with statistical power, their confidence ratings should be affected equally by each of the three manipulated statistical variables. That is, the effect sizes for sample size, mean difference, and *SD* should be the same.

As explained previously, presentation format was manipulated so that one group of subjects received only the statistical summary information on number lines, whereas another group viewed the individual rating given to each product in addition to the statistical summaries.

Finally, we collected data about our subjects' numerical abilities. We constructed and used a numeracy measure that required subjects to convert between probabilities, percentages, and frequencies. Also, we used Frederick's Cognitive Reflection Test (CRT; 2005) and asked subjects for their SAT quantitative score, their gender, and whether they had ever taken a statistics class or had lessons in probability theory.

Table 1
Mean and Standard Deviation Confidence Ratings for Each Condition

Manipulated Power			Normative Benchmark			Variables Display Condition: Summary		Variables Display Condition: Summary and Raw Data	
N	M_{diff}	*SD*	Power	*t*	LR	*M*	*SD*	*M*	*SD*
+	+	+	1.00	6.08	535,626	3.45	0.84	3.33	0.79
+	+	−	.85	3.03	9.88	3.28	0.75	3.16	0.76
+	−	+	.85	3.04	10.18	2.59	0.75	2.67	0.88
−	+	+	.85	3.16	9.88	3.26	0.81	3.27	0.76
+	−	−	.32	1.52	0.37	2.41	0.72	2.60	0.78
−	+	−	.32	1.58	0.82	3.14	0.82	3.07	0.82
−	−	+	.32	1.58	0.82	2.56	0.74	2.68	0.77
−	−	−	.12	0.79	0.31	2.43	0.70	2.64	0.79

Note—The "+"s indicate values that increase statistical power (large sample size, large difference between mean product ratings, and low within-groups *SD*s). The "−"s indicate the opposite: values that detract from power. The following are the high and low power values, respectively: sample size per group ($n = 37$, $n = 10$), mean difference ($\bar{X}_1 - \bar{X}_2 = 2$, $\bar{X}_1 - \bar{X}_2 = 1$), and *SD* (1.41, 2.84). The two rightmost columns show subjects' responses. LR, corrected likelihood ratio.

RESULTS

The three confidence rating items were well correlated with one another, as predicted (Cronbach's $\alpha = .72$), so the three were averaged together to form one composite confidence score. The composite score was used as the dependent measure in a mixed ANOVA. Presentation format and order were included as between-subjects factors, and mean difference, sample size, and *SD* were included as within-subjects factors (see Table 2).

The means and *SD*s for each of the eight comparison conditions are given in Table 1 and the marginal means for each of the three manipulated factors are plotted in Figure 2. Notice that if subjects respond to these variables normatively, the slopes of the graphs should be equal and positive. This figure reveals that confidence ratings were strongly related to mean difference, moderately related to sample size, and weakly related to *SD*. These impressions were confirmed by an ANOVA (see Table 2) that revealed a large main effect of mean difference [$F(1,199) = 281.67, p < .0001$], a moderate main effect of sample size [$F(1,199) = 17.47, p < .0001$], and a small effect of *SD* [$F(1,199) = 8.21, p = .0046$]. Condition interacted with mean difference [$F(1,199) = 7.27, p = .0076$] such that subjects in the summary-only condition gave the difference between product means greater weight than did those in the summary + raw data condition. Mean difference and *SD* interacted [$F(1,199) = 10.03, p = .0018$] such that the effect of mean difference was greater when the *SD* was large than when it was small.

Sensitivity Measures

We created measures to assess subjects' sensitivity to changes in sample size, mean difference, and *SD*. The measure was based on each subject's average change in confidence resulting from a change in level for each variable. This was computed by subtracting the mean of the four confidence scores when one of the variables was at the lower power level from the mean of the four confidence scores when that variable was at the higher power level. If, for example, a subject was sensitive to sample size, we would expect her confidence in a product difference to be smaller in comparisons involving small sample sizes than in comparisons involving large sample sizes. Overall, if subjects were adjusting their confidence in the normative direction, then their change in average confidence for each variable should be positive.

These sensitivity measures were used to make pairwise comparisons to test whether subjects' sensitivity for sample size, mean difference, and *SD* were significantly different from one another. Our hypothesis that subjects would be more sensitive to mean difference than to sample size was confirmed [$t(111) = 23.86, p < .0001$; $t(90) = 16.95, p < .0001$] for subjects in both the summary-only and summary + raw data conditions, respectively. Also as predicted, subjects in the summary-only condition were more sensitive to sample size than to *SD* [$t(111) = 3.24, p = .0016$], whereas those in the summary + raw data condition weighted these factors equally [$t(90) = 0.44, p = .6610$]. Cohen's *d* was used to calculate effect sizes for mean difference, sample size, and *SD* for both our summary-only and summary + raw groups (shown in Figure 2). For the summary-only group, the effect size of mean difference was nearly 5 times that of sample size and 11 times as large as the effect of *SD*. In the summary + raw data group, the mean difference effect was nearly 8 times larger than that of both sample size and *SD*. Consistent with this, 87% of individual subjects weighted mean difference in the normative direction, whereas only 56% and 45% of subjects weighted sample size and *SD*, respectively, in the correct direction.

Individual differences. We computed the correlations between each of our three sensitivity measures and our numeracy scale, the CRT, statistics and probability exposure, SAT quantitative scores, and gender (36 correlations total; $\alpha = .05/36 = .0014$). Numeracy was positively correlated with sensitivity to mean difference ($r = .344, p = .0001$), but not sample size or *SD*. Both Frederick's CRT and SAT quantitative score were positively related to sensitivity to sample size ($r = .234, p = .0008$; $r = .301, p < .0001$) and mean difference ($r = .321, p < .0001$; $r = .331, p < .0001$), respectively. Exposure to statistics was correlated with attention to sample size data ($r = .235, p = .0007$). Notably, instruction in probability was not related to the use of mean difference, sample size, or *SD*, nor was gender.

Table 2
Main Effects and Interactions of Interest From the ANOVA

Source	$F(1,199)$	MS_e
Between-Subjects Effects		
Condition (C)	0.21	2.84
Order (O)	0.03	2.84
Within-Subjects Effects		
Sample size (*N*)	17.47***	0.29
Mean difference (M_{diff})	281.67***	0.64
Standard deviation (*SD*)	8.21**	0.30
$N \times C$	3.07	0.29
$M_{diff} \times C$	7.27	0.64
$SD \times C$	0.00	0.30
$N \times M_{diff}$	0.12	0.18
$N \times SD$	0.23	0.21
$M_{diff} \times SD$	10.03**	0.32

Note—The ANOVA included all main effects and interactions. Only the interactions of interest are shown here. The only significant higher order interactions were $SD \times O$,* $N \times SD \times O$,* $N \times SD \times C \times O$,** and $N \times M_{diff} \times SD \times O$.*** *$p < .05$. **$p < .01$. ***$p < .0001$.

DISCUSSION

We find that laypeople have a modest intuitive understanding of how statistical concepts should affect pairwise comparisons. Subjects showed some degree of understanding of how sample size, mean rating, and *SD* information should affect one's confidence in such an evaluation. On average, subjects were more confident that one product was better than another when the rating data were based on many rather then few consumers' opinions, when the difference between the average product ratings was large rather than small, and when there was high, rather than low agreement among raters. However, although sample size and *SD* were shown to significantly affect subjects'

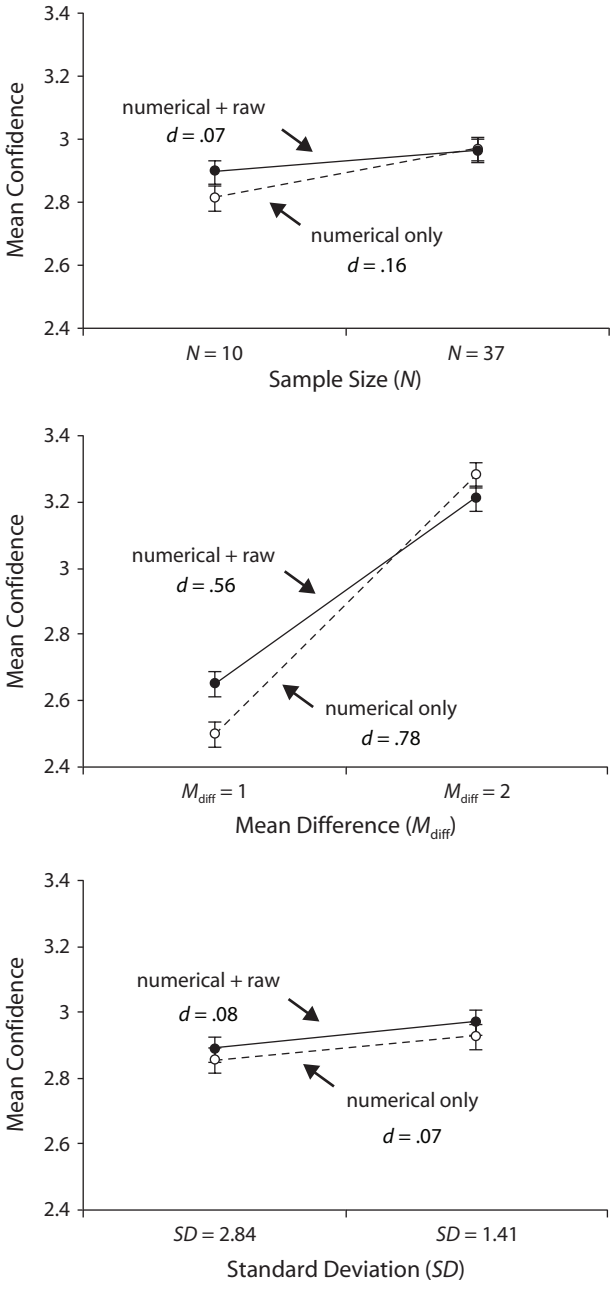

Figure 2. Mean confidence ratings (with standard error bars) for low and high levels of each of the manipulated power factors. Results are shown separately for the summary + raw data and summary-only conditions. The top panel shows the mean confidence for the two levels of sample size, the middle panel shows the mean confidence for the two levels of mean difference, and the bottom panel shows the mean confidence for the two levels of *SD*. The positive slopes indicate that each factor is treated in the normative direction. Effect sizes (Cohen's *d*) are shown.

judgments in the normative direction, their effect sizes were exceedingly small. When we examined individual subjects' sensitivity to sample size and *SD*, we found that only about half of subjects weighted these factors in the normative direction, with the rest showing zero or non-normative weighting. Thus, subjects gave very little consideration to sample size and *SD*. Instead, their judgments were primarily driven by the mean differences.

This result can be understood in light of Gallistel and Gelman's (2005) work on nonverbal numerical representation. According to this dual representation view, numerical quantities can be represented either nonverbally or arithmetically. The nonverbal representation is sufficient for comparing two mean ratings, with larger differences detected faster than smaller differences. In contrast, incorporating sample size and *SD* requires the arithmetic representation, because interpreting mean difference in light of sample size and *SD* entails computations on rational numbers (computing products and ratios). A heavy reliance on the nonverbal representation of discrete quantities could explain why laypeople attend to mean difference almost to the exclusion of sample size and *SD*. Previous research has indicated that individuals have considerable difficulty understanding the arithmetic principles governing rational numbers (Hartnett & Gelman, 1998).

Subjects in the summary + raw data condition saw both the individual ratings given to each product, and also summaries of these data on labeled number lines. In contrast to our prediction, these participants did not attend to sample size or *SD* more than those subjects who only viewed summary representations of the raw data. This general pattern of results is consistent with findings from a previous study in which subjects were shown arrays of circles whose sizes represented individual product ratings. As in the present experiment, subjects' confidence ratings were much more strongly related to mean difference data than to sample size or *SD*. In addition, providing raw data did not make subjects respond in a more normative manner in comparison with subjects receiving only summary representations (Obrecht & Chapman, 2006). Thus, in contrast to Gigerenzer's (2000) view, we found no evidence that providing raw data to subjects prompts them to compare options in a more normative manner. Some recent research indicates that when compared with statistical summaries, raw data trigger different, but not necessarily more normative, decision processes (Gottlieb, Weiss, & Chapman, 2007; Hertwig, Barron, Weber, & Erev, 2004). Our findings appear to be consistent with this work.

Finally, we examined the relationship between the use of the statistical information and individual difference measures. We found no significant relationships between individual difference measures and use of *SD*. However, subjects who performed better on the CRT and reported higher SAT quantitative scores were more sensitive to sample size and mean difference. The CRT is composed of three items, each of which brings to mind an intuitive, but incorrect, answer. Those who bypass the intuitive response to arrive at the correct answers are thought to be less impulsive and more likely to engage in slow, effortful, reflective thinking (Frederick, 2005). The present study also found that subjects higher in numeracy were more sensitive to the difference between means than were less numerate people. One possible account for this difference could be that more numerate subjects are using a numerical notational system to interpret the mean differences, whereas less numerate subjects may rely on their underly-

ing nonverbal discrete numerical system (see Gallistel & Gelman, 2005). One of the hallmarks of the Gallistel and Gelman model of nonverbal representations of discrete values is that they are noisy. Consequently, the mean differences in the present study may appear less salient to less numerate subjects than to their more numerate counterparts. This numeracy finding is in line with some of our recent work showing that numeracy is related to other types of statistical reasoning—namely, how people judge the likelihood of an event occurring (Obrecht, Chapman, & Gelman, 2006). Finally, in the present study, subjects who reported previous exposure to statistics tended to use sample size information more than did those without. Thus, it seems that more numerate people who are more likely to engage in effortful thinking and who have had exposure to statistics are better able to use statistical information to make informed pairwise comparisons.

Overall, the results of this experiment show that people have a limited intuitive sense of how statistical information can be used to make paired comparisons. Although they were more confident in a pairwise difference when mean difference and sample size were large rather than small and when variance was low rather than high, subjects gave these factors far from the equal weight that would be expected if their intuitions were in line with statistical power. Our results indicate that people have a limited ability to make normative inferences on the basis of statistical summary data and that providing raw data may not improve how such information is used. An interesting question for future research is whether performance would be improved by explicit instruction about the principles that govern the interpretation of statistical summary data.

AUTHOR NOTE

This project was supported by NSF Grant SES-03-25080 to the second author, NSF Grant REC-9720410 to the third author, and a Rutgers University excellence fellowship awarded to the first author. We thank Jenny Cooper and Jacob Feldman for their comments and suggestions. Correspondence concerning this article should be addressed to N. A. Obrecht, Rutgers University, Psychology Department, 152 Frelinghuysen Road, Piscataway, NJ 08854-8020 (e-mail: natalie@ruccs.rutgers.edu).

REFERENCES

FREDERICK, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, **19**, 25-42.

GALLISTEL, C. R., & GELMAN, R. (2005). Mathematical cognition. In K. Holyoak & R. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 559-588). Cambridge: Cambridge University Press.

GIGERENZER, G. (2000). *Adaptive thinking: Rationality in the real world.* New York: Oxford University Press.

GIGERENZER, G., & EDWARDS, A. (2003). Simple tools for understanding risks: From innumeracy to insight. *British Medical Journal*, **327**, 741-744.

GIGERENZER, G., & HOFFRAGE, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, **102**, 684-704.

GLOVER, S., & DIXON, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, **11**, 791-806.

GOTTLIEB, D., WEISS, T., & CHAPMAN, G. B. (2007). The format in which uncertainty information is presented affects decision biases. *Psychological Science*, **18**, 240-246.

HARTNETT, P. M., & GELMAN, R. (1998). Early understandings of numbers: Paths or barriers to the construction of new understandings? *Learning & Instruction*, **8**, 341-374.

HERTWIG, R., BARRON, G., WEBER, E., & EREV, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, **15**, 534-539.

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

HOFFRAGE, U., & GIGERENZER, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, **73**, 538-540.

HOFFRAGE, U., LINDSEY, S., HERTWIG, R., & GIGERENZER, G. (2000). Communicating statistical information. *Science*, **290**, 2261-2262.

KAHNEMAN, D., & TVERSKY, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, **3**, 430-454.

LENTH, R. V. (2006). Java applets for power and sample size [Computer software]. Retrieved January 2006 from www.stat.uiowa.edu/~rlenth/Power.

NISBETT, R. E., KRANTZ, D. H., JEPSON, C., & KUNDA, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review*, **90**, 339-363.

OBRECHT, N. A., & CHAPMAN, G. B. (2006, November). *Intuitive t tests: Lay use of statistical information*. Poster session presented at the Psychonomic Society Annual Meeting, Houston, TX.

OBRECHT, N. A., CHAPMAN, G. B., & GELMAN, R. (2006, November). *Statistical reasoning is influenced by serial presentation of information.* Paper presented at the annual meeting of the Society for Judgment and Decision Making, Houston.

SEDLMEIER, P., & GIGERENZER, G. (1997). Intuitions about sample size: The empirical law of large numbers. *Journal of Behavioral Decision Making*, **10**, 33-51.

SEDLMEIER, P., & GIGERENZER, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology*, **130**, 380-400.

TVERSKY, A., & KAHNEMAN, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, **185**, 1124-1131.

(Manuscript received August 18, 2006;
revision accepted for publication February 26, 2007.)