# האוציברסיטה העצברית בירושלים THE HEBREW UNIVERSITY OF JERUSALEM 

# ON THE MISPERCEPTION OF VARIABILITY 

by

## YAAKOV KAREEV, SHARON ARNON and REUT HORWITZ-ZELIGER

Discussion Paper \# 285 February 2002

מרכז לחקר הרציונליות

## CENTER FOR THE STUDY OF RATIONALITY

URL: http://www.ratio.huji.ac.il/

# On the Misperception of Variability 

Yaakov Kareev, Sharon Arnon, \& Reut Horwitz-Zeliger<br>School of Education<br>and Center for Rationality and Interactive Decision Theory<br>The Hebrew University

email: kareev@vms.huji.ac.il
in press, Journal of Experimental Psychology: General


#### Abstract

Ever since the days of Francis Bacon it has been claimed that people perceive the world as less variable and more regular than it actually is. Such misperception, if shown to exist, could explain a host of perplexing behaviors. However, the only evidence supporting the claim is indirect, and there is no explanation of its cause. As a possible cause, we suggest the use of sample variability as an estimate of population variability. This is so since the sampling distribution of sample variance is downward attenuated, the attenuation being substantial for sample sizes that people are likely to consider. The results of five experiments show that people use sample variability, uncorrected for sample size, in tasks in which a correction is normatively called for, and indeed perceive variability as smaller than it actually is.


Even though accurate assessment of variability is necessary to estimate the confidence with which predictions and choices between alternatives can be made, there is reason to suspect that the perception of variability is downward attenuated, that people perceive the world around them as less variable and more orderly than it actually is. As long ago as 1620, Francis Bacon, describing the "idols" - those bad habits of mind that cause people to fall into error - noted that people expect more order in natural phenomena than actually exists: "The human understanding is of its own nature prone to suppose the existence of more order and regularity than it finds" (1620/1905, p. 265). Many findings in psychology may also be interpreted as evidence that people overestimate the degree of regularity in their environment. The fundamental attribution error (Gilbert \& Malone, 1995; Ross, 1977) consists in people perceiving other people's behavior as more consistent than it is, and then ascribing that consistency to stable personality characteristics (Mischel, 1968). The illusion of control (Langer, 1975) is another case in which people exhibit greater confidence than warranted in their ability to predict and control future events. A downward attenuation in the perceived variability of other people's behavior, or in the spread of outcomes around a predicted value, could be a factor contributing to these ubiquitous errors. The confidence with which people predict representative values of a distribution (Kahneman \& Tversky, 1973), and their stronger-than-warranted belief in the accuracy of statistics observed in samples of small size (Tversky \& Kahneman, 1971), could also be caused, at least in part, by perceiving variability around a central value as smaller than it actually is.

To date, there have hardly been any studies with a direct bearing on the accuracy with which variability is perceived. Earlier studies (for reviews, see Peterson \& Beach, 1967; Pollard, 1984) have involved either the detection of change in variability or its estimation on
an arbitrary scale. Similarly, studies of risky choice (see March, 1996, for a review) employ tasks that call for a choice between two options that differ in variability. Thus, even though these earlier studies revealed sensitivity to differences in variability and a strong positive correlation between actual and perceived variability, their methods could not detect any systematic bias in its perception.

In the present paper we try to determine whether or not people's perception of variability is indeed biased. We first propose a theoretical argument to explain why people's perception of variability may be expected to be inaccurate: We point out that, typically, people use sample data to infer population variability. Since sample variance is a biased estimate of population variance, to the extent that people use sample variability as their estimate of population variability, their estimates will be lower, on average, than the true values (unless they apply an additional, statistical correction for sample size). Following the theoretical argument, we present the results of five experiments bearing on this issue. The first two experiments assessed whether the variability of a population seen previously in its entirety, or one from which only a sample was being observed, was perceived to be smaller than that of a comparable population seen in its entirety at the time of judgment. The next two experiments tested the issue of correction for sample size: In Experiment 3 participants judged which of two populations had a higher mean, after having observed a sample out of each. We then determined whether measures employing observed sample variance, or measures employing sample variance corrected for sample size, yielded a better prediction of judgments and confidence in them. In Experiment 4 the behavior of participants with different working-memory capacities (which were therefore likely to differ in the size of the sample they consider) was compared. Finally, Experiment 5 assessed which parts of a sample of a size exceeding working-memory capacity are most likely to be taken into account in
subsequent choices that depend on estimates of variability. The final section explores some implications of the findings.

## The Theoretical Argument

In order to function properly, people must come to know the environment in which they live. However, time constraints (e.g., the need for a speedy decision), memory limitations (e.g., the limit on sample size imposed by working-memory capacity), or simply the unavailability of more information, often force people to use sample data (statistics) to infer population characteristics (parameters). In most cases, that inference is straightforward: Since many statistics are unbiased estimates of their respective parameters, one's best estimate of the parameter of interest is the sample statistic as observed. For estimates of these parameters, sample size is significant only in that it indicates how confident (or cautious) one should be in adopting the observed statistic as an estimate of its respective parameter. For some parameters, however, the expected value of their sample statistic is systematically biased. When the information of interest involves population parameters of the latter type, the use of sample statistics is likely to lead to a distorted view. Furthermore, the degree of bias is inversely related to the sample size; therefore, the bias due to sample size, unless corrected for, is of much greater consequence when inferences involving those parameters are based on small samples.

One such parameter is the correlation between two variables: When two variables are correlated, the sample value of the correlation is, more often than not, more extreme than its value in the population (Hays, 1963; for a recent treatment, see Fiedler, 2000). With respect to correlation, theoretical analyses indicate that the bias thus induced may improve, rather than hinder, overall performance, since it increases the chances for early detection of a correlation when one exists (Kareev, 1995a, 1995b, 2000). Empirical studies (Kareev,

Lieberman, \& Lev, 1997), comparing the performance of people who differed in their working-memory capacities (hence in the sample sizes they could consider), or in the size of the sample presented to them, revealed that people using smaller samples not only perceived the correlation as more extreme, but also detected it more rapidly and were more accurate in their predictions.

The variability of a population is another parameter for which sample statistics provide a systematically biased value. For the most widely used measure of variability, variance, the expected value in a sample of size $\underline{n}$ is smaller by a factor of $\underline{(n-1) / n}$ than that in the population (Hays, 1963). To illustrate: For $\underline{n}=7$ - the most likely upper bound on the number of items considered simultaneously by a typical adult (Miller, 1956) - variance is expected to be downward attenuated by approximately $14 \%$. For continuous variables, the probability that the variance of a sample of size $\underline{n}$ drawn from a normal population will be smaller than that in the population is equal to the probability that $\chi^{2}(n-1)>n$; for $\underline{n}=7$ this probability is .679 . For binary variables, where the degree of bias depends not only on $\underline{n}$ but also on the proportion of each value, sample variance, calculated over the complete range of proportions, is smaller than that in the population in .641 of samples of size 7 . There is no reason to believe that people use variance - which is, of course, a statistical construct - as their estimate of variability. Nonetheless, all statistical measures of variability are highly correlated with each other, and people's estimates of variability are also known to be highly correlated with variance (see Pollard, 1984, for a review). Thus, if people use sample variability as their estimate of population variability, without correcting for sample size, their estimate will be on average smaller than the true value. ${ }^{1}$

In what follows, we describe five experiments designed to ascertain if people indeed regard variability as smaller than it is, and whether they use sample data, without correction
for sample size, in tasks calling for the assessment of variability.

## Experiment 1

This experiment was designed to determine whether the perception of the variability of a population, observed earlier in its entirety but out of sight at the time of judgment, is biased or not. Earlier studies revealed a high degree of correspondence between the actual variance of distributions and people's ratings of it (Beach \& Peterson, 1967; Pollard, 1984); however, these studies do not bear on the question whether perceived variability is systematically biased or not. This is so because the arbitrary measures used in those studies could only indicate correspondence but not reveal bias.

To avoid this difficulty, we employed measures that require comparison of variabilities. Thus we could detect the presence of a bias (though we could not measure its strength, if found to exist). Participants first viewed one group of items (a population) and then had to indicate which of two other groups, in full sight at the time of comparison, better resembled the original one (which was then out of sight). The items were either paper cylinders, colored up to a certain height, or plastic discs, marked with one of two letters. The participants were not aware that one of the two comparison populations was identical to the original, while the other was either less or more variable.

The study allowed for a stringent test of our question: If people accurately perceive and remember the variance of a population - either because they use the whole population to estimate it, or because they use sample variability but correct for the bias due to sample size they should be as good at choosing the identical comparison population over both the less and the more variable one. However, if people are sensitive to variability, but in trying to infer it for the out-of-view population they do so on the basis of a sample retrieved from memory without correcting for sample size, their estimates should be downward attenuated. If such an
attenuation indeed occurs, people should find it more difficult to choose correctly between the identical population and the one of lower variability than between the identical and the more variable one. Obviously, if people are insensitive to variability, they should be unable to distinguish between the two comparison populations, irrespective of their variability.

## Method

Task and Design. Participants were presented with one population and then had to indicate which of two comparison populations was more similar to the original one. As mentioned above, one of the two comparison populations was identical to the original, item for item. The other comparison population was different in variability from the original, its variance being either $6 / 7$ or $7 / 6$ that of the latter. Thus, the ratio between the variances of the two comparison populations was always $6 / 7$. To enhance generality, Type of Variable continuous or discrete (and binary) - was manipulated. Type of Variable was manipulated within participants; thus, each participant saw two original populations: one of them consisted of paper cylinders colored up to a certain height (continuous variable), whereas another consisted of a mixture of small discs marked with one of two letters (discrete, binary variable). The variability of the nonidentical comparison population was manipulated independently of Type of Variable, and all combinations of the variability of the nonidentical population vis-à-vis those of the original occurred with equal frequency.

Participants. Ninety-six volunteers studying at the Hebrew University Mount Scopus campus, who were paid 5 IS (about $\$ 1.25$ ) for participation.

Materials. The materials used in this experiment, as well as in the others, were deliberately chosen so as to eliminate the effects of any prior notions on the perception of variability. For the continuous variable, a population consisted of 28 paper cylinders, each 12
cm long and colored up to a certain height; color-heights were normally distributed, with a mean of 6 cm . In the original population, the standard deviation was 1.955 cm ; in that with the smaller variance, it was 1.811 , and in that with the larger variance, 2.112 . The binaryvalued population consisted of 28 white plastic discs, 2.54 cm in diameter. Each disc had either the letter T or the letter L printed on both sides. The original population consisted of 19 L's and 9 T's. For the nonidentical comparison populations, the one with the lower variance had 21 L's and 7 T's, and that with the larger variance, 15 L's and 13 T's. All items were stored in identical-looking, covered boxes.

Procedure. Participants were tested individually. At the onset of the session they were informed that they would see two different groups of items, each stored in a separate box, and would have 10 seconds to watch each. One box would contain paper cylinders, each colored up to a certain height, whereas the other would contain discs bearing the letters T or L . They were also told that, after seeing the original two groups, they would be presented with two pairs of boxes, one pair for each of the original groups, and would have to judge the contents of which better resembled that of the original box. Following the instructions the experimenter lifted the cover of the box containing one of the two original groups, and let the participant watch it for 10 secs. The cover was then closed and the procedure repeated with the original box containing the other type of material. Presentation of the original boxes was followed by the comparison stage. The order of presentation during comparisons corresponded to that in the first phase, with the two comparison boxes for each type of variable being presented simultaneously.

Performance during comparison was self-paced and typically very fast, with the first pair of boxes being judged about 20 to 30 seconds after the participants had seen the first original box, and the second pair of boxes being judged 30 to 45 seconds after the participants
had seen the second original box.
The order of presentation was counterbalanced, with the binary-valued population presented first in half of the cases. For every comparison, the variability of the nonidentical comparison population was equally often smaller or larger than that of the identical one, and the identical comparison box was positioned equally often to the left or to the right of the original. All boxes were shaken horizontally before being opened for inspection; this not only assured a different arrangement of the items, but also implied that the spatial arrangement of the items was immaterial for the judgment. At the same time, the nature of the items in each box ensured that they all remained in full view, rather than staggered on top of each other.

## Results

For the main analysis, each choice was scored +1 or -1 . The score was +1 when the comparison box with smaller variability was judged as being more similar to the original, and -1 when that with larger variability was so judged. Thus, the scoring of choices reflected not their accuracy but rather the tendency to judge the less variable one as being more similar to the original. This, in turn, reflects the degree of support for the prediction that people perceive variability as smaller than it actually is. Since participants had to judge two different populations, their score could be $-2,0$, or +2 . The number of participants having each of these scores was 16,48 , and 32 , respectively. The mean of .33 differed significantly from zero $(\mathrm{t}(95)=2.364, \mathrm{p}=.020)$. It is important to note that Type of Material (continuous vs. discrete) did not have a significant effect on the choices $(\underline{F}(1,95)<1)$. This is of interest since, for the discrete variable, the change in variability also involved a change in the central value (the proportion). The similar effect observed for both types of variable rules out an explanation based on sensitivity to a change in proportion, rather than to a change in variability.

A breakdown of the participants' choices revealed that the identical comparison box
was judged as more similar to the original box in only .49 of the cases when compared with that of the smaller variability, but in fully . 66 of the cases when the comparison box had the larger variability. As reasoned above, such a result could not be obtained if people assess variability correctly, or do not notice it. These findings are compatible, on the contrary, with our claim that people, when retrieving a population from memory, indeed conceive of its variability as smaller than it actually is. Another possibility, one that cannot be ruled out on the basis of our data, is that participants were particularly sensitive to the most extreme items, noticing, for the non-identical distribution with the larger variability, items they had not seen before. ${ }^{2}$ As will be seen, this alternative explanation cannot be applied to account for the results of Experiments 2 through 5.

## Experiment 2

Experiment 2, like Experiment 1, was designed to determine if estimates of variability are downward attenuated. However, the participants in this experiment performed two new tasks, each calling for a comparison of two populations (which, unbeknownst to the participants, were of equal variance). For one pair of populations, they answered a direct question as to which of the two was more variable; for another pair, they chose which should be used in a subsequent drawing, in which chances of success were greater for the population of lower variability. Both tasks were performed under one of two viewing conditions: In one, both populations were first seen in their entirety, but at the time of comparison one was removed, while the other was in view. In the other, the participants saw one population in its entirety, but only a sample of seven items from the other (both sample and population were in view at the time of decision). In light of our earlier finding, we predicted that when both populations had been seen in their entirety, the one out of view would be judged to be of smaller variability. In the condition in which the variability of one population was inferred
from a sample, it was predicted that the variability inferred from a sample would be smaller. Note that, if people's estimates of variability were corrected for sample size, these predictions would be false.

Method
Task and Design. Every participant performed two tasks. For the first, comparison task, participants had to respond to a direct question, asking which of a pair of populations was more variable. For the second, choice task, participants had to choose one of two populations for a subsequent drawing of two items, in which the reward depended on the similarity between the two drawn items (i.e., the chances for a reward were inversely related to the variability of the population). Type of Variable - whether continuous or binary - was manipulated in a manner similar to that in Experiment 1 (see Materials, below). Task (comparison or choice) and Type of Variable were within-participant factors, every participant performing one task with the continuous-variable pair of populations and another with the binary-variable pair of populations. A third variable, Viewing Condition, was a between-participants factor, having two values. In the Out-of-sight/In-sight condition, both populations were first viewed in their entirety, but only one was in view at the time of the decision. In the Sample/Population condition participants drew a sample of seven items out of one population, but saw the other in its entirety. In this condition, both sample and population remained in view at the time of decision.

Participants. The participants were 144 Hebrew University or Ben Gurion University students. They were paid a fixed amount of 8 IS (about $\$ 2$ ) for participation and an additional bonus if successful in their choice.

Materials. For the continuous variable, both populations consisted of 28 paper cylinders, identical to those employed in the original distribution of Experiment 1. The two
populations differed only in their colors. For the binary variable, the populations consisted of 28 pieces of uniformly colored wood (actually, halved clothespins). Eighteen of the pieces were of one color, 10 of another; both populations had the same $18 / 10$ split, but with different colors.

Procedure. The participants were tested individually, in a quiet room. Upon entering, they were informed that they would perform two tasks, each involving either a comparison of two populations, or a choice between them. The participants in the Out-of-sight/In-sight viewing condition were informed that for each task they would see the contents of two boxes - one exposed for 10 seconds and then covered, the other exposed and remaining in view. The participants in the Sample/Population viewing condition were informed that for each task they would be presented with two boxes, one of which would be open, so that they could see all its contents, while from the other they would draw a sample of 7 items. They were further informed that both the open box and the sample would remain in view at the time of the decision.

The phrasing of the direct question depended on Type of Variable. For the continuous variable, the participants were asked in which of the two populations the distribution of values was more homogeneous; for the binary variable, they were asked which population had a split of colors closer to $.5 / .5$. Note that a simple response bias to choose one of the two populations over the other would result in opposite effects for the two types of variable. For the choice task, the participants were promised a reward of 4 IS if a pair of items drawn from the population they chose fell within a range of 1.6 cm (for the continuous variable) or if the wooden pieces turned out to be of the same color (for the binary variable). Since the chances of winning the reward in the choice task were greater, the smaller the variability of the population from which items were to be drawn, we took the participants' choice as an
indication of which population they believed had smaller variability.
All combinations of Task and Type of Variable were equally frequent, the order of presentation being fully counterbalanced within each viewing condition.

## Results and Discussion

Both decisions were scored either ' 0 ' or ' 1 '. For the Out-of-sight/In-sight viewing condition, an answer was scored ' 0 ' if the population in full view was judged to be less variable or was chosen for the drawing, and ' 1 ' if the population out of sight was so judged or chosen. The scores of ' 0 ' and ' 1 ' were similarly assigned for decisions in the Sample/Population viewing condition, with choices of the population from which only a sample had been seen corresponding to those of the out-of-sight population. Thus, mean scores represent the proportion of answers in line with our hypothesis that, when people have to infer variability from memory or from a sample, the variability of a distribution seems smaller than it actually is. The mean results for each condition are presented in Table 1. An analysis of variance of these data revealed that the only significant effect was the deviation of the overall mean from the value of .5 expected by chance $(\underline{F}(1,140)=6.34, \underline{\mathrm{MSE}}=1.06, \underline{p}=$ .013). Thus, the results indicate that, if the variability of a population is assessed when it is out of sight, or if only a sample of it is available, the variability indeed seems smaller than that of an identical population in view at the time of judgment. Since in most life situations the complete population is unavailable for variability assessment, these results further indicate that variability is normally perceived as smaller than it actually is.

It is worth noting that the overall effect was identical under both viewing conditions. This further suggests that, when a population is out of sight, its parameters are assessed by recalling a sample. Another interesting point is that, of all other effects, the only one with an $\underline{\mathrm{F}}$ value exceeding 1 was that of $\operatorname{Task}(\underline{\mathrm{F}}(1,140)=2.16, \underline{\mathrm{MSE}}=.93, \underline{p}=.144)$, reflecting the
fact that the downward attenuation in the perception of variability was more pronounced in the Direct Evaluation than in the Choice task (. 618 vs. . 535 , respectively). Since choice of the out-of-sight or the sample-only population entails the use of more ambiguous information, the difference, if real, may indicate that when a monetary reward rather than an inconsequential choice is involved, aversion of ambiguity (Frisch \& Baron, 1988; Heath \& Tversky, 1991) may play a role. Finally, it should be noted that the overall proportion of cases in which the variability of the population in view was judged as larger was .577 ; this value is significantly greater than chance, but lower than the .65 or so expected if people retrieve samples of size 7 from memory for a population out of view, but correctly judge the variability of a population in full view. One possible explanation of the difference could be that people partially correct for the bias induced by small sample data. Another possibility is that, even for populations in full view, variability is assessed on the basis of sampling. This would lead to biased perception of variability, and hence to a smaller difference than predicted by theory. To anticipate, the results of Experiment 4 seem to favor the latter explanation.

The first two experiments demonstrated that, as suspected, variability estimated from memory or from a limited sample is downward attenuated. As suggested above, the cause of such attenuation could be the use of sample variability, uncorrected for sample size, as the estimate of population variability. Experiments 3, 4 and 5 were designed to find out if people indeed use sample data when estimating variability, and whether or not they correct these estimates for sample size. All three experiments required a prediction to be made - of the difference in value between items to be drawn out of two populations (Experiment 3); of the makeup of a sample to be drawn out of a population (Experiment 4); or of the values of single items to be drawn out of one of two populations (Experiment 5). Thus, in none of these experiments were participants asked to assess variability. However, the tasks were such that
performance would improve, should participants take note of variability. As explained below, we could use participants' choices to infer whether they took notice of variability and, if they did, whether or not they corrected their estimates of variability for the size of the sample on which they were based.

## Experiment 3

In this experiment, the participants were presented with a big box containing a hundred matchboxes. Fifty of the matchboxes were of one color and 50 of another. Each matchbox contained a number of matches; the number of matches was variable, but the mean number of matches in the matchboxes of one color differed from that in the other. After having observed the number of matches in a sample of matchboxes of each color, the participant had to decide which color was that of the matchboxes containing a larger number of matches. The participant then also placed a bid on the outcome of a subsequent draw of one matchbox from both populations, whose reward was the product of the bid and the difference between the number of matches in the two boxes. For analysis, we correlated measures derived from the samples actually observed with participants' choices and bids, and compared the predictive power of various measures, to determine which of them best predicted behavior. Normatively, confidence in the choice and hence the size of the bid should reflect not only the estimated difference between the means, but also the variability of the sampling distributions of that difference (which depends, of course, on the variability of each distribution). To ascertain whether this was the case, we included measures that took sample variability into account and measures that did not. More important, we compared the predictive power of measures that included sample variability, as observed, as one of their terms, with the predictive power of analogical measures which use an estimate of variability corrected for sample size. We reasoned that those measures that turned out to be the best
predictors of behavior in our tasks would indicate what measure of variability, if any, plays a role in people's choices and their confidence in those choices.

## Method

Participants. The participants were 80 Hebrew University students, each paid a fixed amount of 12 IS (about \$3) for participation, plus a reward calculated as described above.

Materials. As mentioned above, the task was performed with a pair of populations, stored together in a big box. Each population consisted of 50 matchboxes containing a number of matches. The two populations differed in their mean number of matches but had either the same or very similar variance. The color of the matchboxes of one population differed from that of the other.

Since our analysis was to relate participants' performance to characteristics of the samples they actually drew from a pair of populations, it was necessary to ensure a wide range of sample values (and, more importantly, a wide range of values based on differences between the samples, such as differences between sample means, estimates of effect size, or measures involving inferential statistics and their significance). To that end we have used eight pairs of populations in all. One of the two populations in every big box always had the same characteristics (that population is henceforth referred to as the standard population). The mean number of matches of the other population in a pair differed from that of the standard by $1.5,2,2.5$, or 3 matches. Thus, of the eight pairs of populations, there were two pairs for each of the four possible differences in means. In addition we manipulated the size of the sample participants drew from each population; sample sizes were always different for the two pairs of populations with the same difference between their means. It should be noted that changes in sample size affect the expected values of inferential statistics and of their significance, but not estimates of differences between means or of effect sizes. Table 2
presents the standard population and the eight populations paired with it in the eight big boxes. We also present the mean and standard deviation of all populations, and the effect size and expected significance of the $\underline{t}$-test of the difference between each population and the standard.

Procedure. The participants were tested individually, first receiving a description of the task, then performing it three times, every time comparing a different pair of populations. For each pair, the participants first sampled the specified number of boxes with replacement, opening and noting the number of matches in each of the sampled boxes. Sampling was selfpaced and the order in which boxes were sampled was at the participant's discretion. When the predetermined sample size was reached for the matchboxes of each color, the participant was told that two additional boxes, one of each color, would be drawn later; he or she was then asked to bet on which type would contain more matches, placing a bid (limited to 2.5 IS) on the difference in number of matches between them. To calculate the reward (positive or negative), the bid was to be multiplied by the difference in the number of matches between the two boxes actually drawn. To eliminate the effects of success (or failure) on subsequent choices and bids, the reward for each task was determined only after all three choices and bids had been made.

Each participant performed the comparison task three times. Of the eight populations that could be paired with the standard, only the following subsets were used: (I1, I2, I3), (I1, I2, I4), (I1, I3, I4), (I2, I3, I4), (II1, II2, II3), (II1, II2, II4), (II1, II3, II4), (II2, II3, II4). As can be verified from Table 2, within each triad (e.g., I1, I2, I3) one of the three comparisons resembled another in effect size (e.g., I1 and I2), and the third in expected significance (e.g., I1 and I3). The comparison that thus resembled the other two (I1 in the I1, I2, I3 triad) was always presented second. Order of the other two comparisons (first or third) was
counterbalanced across the 10 participants who performed each triad.

## Results and Discussion

Correlations were calculated to determine which statistics of the actually observed samples best accounted for participants' choices and bids. Choices were scored 1 when the population with the higher mean was chosen, and 0 when not. ${ }^{3}$

The following statistics were derived from the samples actually observed for each pair of populations, and correlated with choices and bids made for the same pair: a) difference between sample means; b) variances (actual sample variance; and estimated population variance based on sample variance, corrected for sample size); c) ratio between the variances of the samples; d) sample sizes; e) point-biserial correlation and point-biserial correlation squared, both serving as estimates of the effect size of the difference between the two populations (each of these was calculated twice, once with the actual sample variance and once with sample variance corrected for sample size); f) $\underline{Z}$ and $\underline{t}$ inferential statistics for the difference between the two means (i.e., a statistic involving the difference between the sample means, divided by the standard deviation of the sampling distribution of that difference, with the value for $\underline{Z}$ using actual sample variances divided by sample sizes, and that for $\underline{t}$ using estimates, employing a correction for sample sizes); g) significance of the $\underline{Z}$ and $\underline{t}$ statistics. Thus, the analysis juxtaposed the predictive power of measures that use actual sample variance and measures that use estimates of population variance derived from sample variance corrected for sample size.

As noted above, our main object was to determine which of the sample statistics was the best predictor of behavior, and to determine a) whether it was one having a measure of variance as one of its terms, and, if it did, b) whether that term reflected sample variance, as observed, or sample variance corrected for sample size. Since many of the predictive
measures we have compared had common components, the correlation between them were very high; thus there was no point in testing for the significance of the difference between correlation coefficients. Instead, the measure of interest was the rank order of those correlations. Table 3a presents the correlations between the best six predictors and the criteria we used; Table 3b presents the correlations between these criteria. As it turned out, the best single predictor of the participants' choices was the significance of the $\underline{Z}$-test - a measure which is affected not only by the difference between the two means, but also by their variances and by the sample size: Its correlation with choice was $\underline{r}=.522(\underline{F}(1,238)=89.31, \underline{p}$ <.001). The best single predictor of the bids was the uncorrected estimate of effect size - a measure that takes into account the difference between the two means and the variability in the combined samples, but without correcting it for sample size. Its correlation with bid size was $\underline{\mathrm{r}}=.302(\underline{\mathrm{~F}}(1,238)=24.45, \underline{\mathrm{p}}<.001)$. When individual differences in bid-size were taken into account, by subtracting each individual's mean bid from each of his or her bids, that same measure remained the best single predictor, with prediction even improving ( $\underline{r}=.387$, $\underline{\mathrm{F}}(1,238)=44.50, \mathrm{p}<.001)$.

Most important from the present perspective are the findings that, for choices and bids alike, a) variability was a component of the best predictors of behavior (i.e., people were taking variability into account), and b) the best predictors included a measure of variability that was based on actual sample variance, rather than on estimates of variance corrected for sample size. It is also of interest that the best predictor of choices (the significance of $\underline{Z}$ ) also reflects the sizes of the samples considered. We do not contend, of course, that in making choices or bids people calculate the significance of a Z-test or effect size. Rather, we see the current findings as an indication that people, in making decisions, use some composite measure which takes variability into account, but in which the variability of the sample, as
observed, is used, rather than an estimate of population variability corrected for sample size. As pointed out in the introduction, the sample variability is, on average, downward attenuated, and the more so, the smaller the sample. Such attenuation does not affect the choice itself, which depends on the difference between the means, but increases the confidence with which the choice is made.

## Experiment 4

Experiment 4, like Experiment 3, was designed to determine whether people correct their estimates of variability for the bias introduced by the use of sample data. Unlike Experiment 3, however, where the participants saw a sample whose size we ourselves manipulated, in Experiment 4 we compared the behavior of participants who were known to have different working-memory capacities. We assumed that these participants also differed in the maximum sample size they could consider when presented with a population whose size exceeded their working-memory capacity, and hence they would behave differently in tasks involving the perception of variability (unless those estimates were corrected for sample size). The task in this experiment was to predict the makeup of a sample to be drawn from a population. Assuming that people would expect the makeup of the sample to resemble that of the population, we used the variability of the predicted sample as a measure of the perceived variability of the population. We reasoned that if people accurately perceive variability (either because they use whole population data, when available, or because they correct for sample size when using sample data), then individual differences in memory capacity should not be related to the variability of the predicted sample. If, however, such a relationship is found, that is, people of lower capacity expect lower variability, this would strongly suggest that people rely on samples, that the size of these samples is positively correlated with their capacity, and that they do not correct for the bias induced by sample size.

## Method

Task and Design. The participants in this experiment first saw a population consisting of 28 items, and then had to predict the makeup of a sample of 7 items to be drawn without replacement from that population. Having made the prediction, participants drew their own sample and were rewarded according to the degree of correspondence between their prediction and the actually drawn sample. Each participant performed the task four times twice for populations that were fully in view at the time of prediction, and twice for populations seen a minute before prediction (out of sight). In each viewing condition, one population involved a continuous variable, and the other, a binary-valued variable. The participants' working-memory capacity was estimated by their performance in a standard digit-span task. Thus, the experiment had a three-way factorial design, the factors being Memory Capacity as a between-participants factor, Viewing Condition and Type of Variable as within-participants factors.

Participants. The participants were 59 Hebrew University undergraduate students, participating for course credit (plus the monetary reward, if they earned any).

Materials. The materials employed in the study were the same paper cylinders and halved clothespins used in Experiment 2. Different colors were used for the two continuous and the two binary populations.

Procedure. The task was performed individually, in a quiet room. The items were stored in opaque boxes that were vigorously shaken, once before exposure and again before drawing. In the out-of-sight condition, the lid of the box was raised for 10 secs and then closed; for the full-view condition, the lid was raised, the participants looked at the box for 10 secs and the lid remained open during prediction. To strengthen reliance on memory, the out-of-sight boxes (one for the binary, the other for the continuous variable) were always
presented before the full-view boxes, with order of presentation counterbalanced between types of variable.

Predictions for all four boxes were made after all of them had been seen, but before any drawing began. For the binary-valued box, the participants predicted the number of items of each color to be found in the sample of seven items to be drawn. For the continuous variable, they predicted the number of items, out of the seven comprising the sample that would be colored above or below a reference height of 5.3 cm , marked on a cylinder held by the experimenter (that height was chosen as to create the same $18 / 10$ split of the 28 items as in the case of the binary variable). After predictions had been made for all four boxes, the boxes were closed and the participants drew a sample of seven items without replacement from each box. The makeup of the sample was compared with the prediction. The participants were rewarded 4 IS (about \$1) when their prediction exactly matched the makeup of the sample, and 2 IS when it deviated by 1 . The digit-span task was administered last.

## Results and Discussion

An extreme group design, comparing participants who had a digit span of less than 6 $(\underline{N}=21)$ with those whose digit span exceeded $6(\underline{N}=24)$, revealed that sample makeups predicted by participants with smaller capacity were significantly less variable than those made by participants with larger capacity $(\underline{\mathrm{F}}(1,43)=4.45, \underline{\mathrm{MSE}}=.32, \mathrm{p}=.041)$. Capacity also interacted with Viewing Condition $(\underline{F}(1,43)=6.29, \underline{\mathrm{MSE}}=.34, \underline{p}=.016)$ : The two groups did not differ in the out-of-sight condition, but did in the full-view condition. Finally, predicted sample makeups were less variable for the binary than for the continuous variable $(\underline{\mathrm{F}}(1,43)=6.61, \underline{\mathrm{MSE}}=.34, \underline{p}=.014)$. See Table 4 for mean number of items of the larger of the two groups of items (a larger number implies lower variance).

The main effect of Capacity indicates that people of different working-memory
capacities behave differently in a task in which performance reflects perceived variability: The products of people with smaller working-memory capacity exhibit smaller variability than those of people with larger capacity. As reasoned above, such a difference could occur if the participants relied on samples whose size corresponded to their working-memory capacity, without correcting their estimates of variability for sample size.

While the main effect of Capacity was in line with our predictions, the significant interaction between Capacity and Viewing Condition was unexpected, as we had expected the difference between the two groups to be larger in the out-of-sight rather than in the full-view condition. Furthermore, the results of Experiment 2 led us to expect a main effect of Viewing Condition, but none was obtained.

In light of the unexpected interaction between capacity and viewing condition, we looked for another way to test the hypotheses that people, in making predictions, rely on limited samples, commensurate with their working-memory capacity, and more so in the out-of-sight than in the full-view condition. The setup of Experiment 4 provided us with such another way of testing both hypotheses. If people with smaller capacity indeed use smaller samples to estimate variability, the overall distribution of their answers should be more variable. This is because the variability of a sampling distribution is larger, the smaller the sample on which it is based. Similarly, if the use of sample data is more prevalent in the out-of-sight condition, one expects larger variability in answers there than in the full-view condition. To find out if this was the case, we compared the within-group variance of the predictions made by members of the low-capacity group to that of members in the highcapacity group. Across all four tasks, the variance of the low-capacity group was 1.059, whereas that of the high-capacity group was 0.776 . This difference in variances is significant $(\underline{F}(83,95)=1.365, \underline{p}=.036$, one-tailed $)-$ another indication that people use sample data in their
assessment of variability.
The same type of analysis could also be employed to compare performance in the two viewing conditions. If people use sample data, sample size (or number of samples considered) would presumably be smaller when the population is out of sight than when it is in full view. Indeed, the variance of the answers given under the out-of-sight condition was larger than that of those given under the full-view condition ( 1.032 vs. 0.782 , respectively; $\underline{\mathrm{F}}(89,89)=1.319$, $\mathrm{p}=.048$, one-tailed).

These two findings, taken together with the main effect of Capacity reported above, support the claim that people use samples to infer population variability. Furthermore, though the measures we used could not establish the exact correspondence between digit-span and sample size, it is clear that people with smaller capacity use smaller samples than people with larger capacity.

Finally, it is of interest to note that the results of Experiment 4 also help to address an issue left unsettled in Experiment 2. There, we recall, the overall proportion of cases in which the out-of-sight or the sampled population was judged to be of smaller variability was .577 significantly greater than chance, but less than the .65 or so expected if no sampling had occurred for the population in full view (and samples of size seven were used for the others). In the discussion of Experiment 2 we commented that such a discrepancy could have resulted either from partial correction for sample size, or from the use of samples even for the population in full view (in which case the number of samples or the size of the sample was larger). The data observed in the present experiment seem to favor the latter explanation: The difference in the overall variance of the answers provided by the low-capacity and highcapacity groups persisted, even when only cases in full view were considered (the variances were 1.071 and 0.510 for the low- and high-capacity groups; $\underline{\mathrm{F}}(41,47)=2.100, \underline{p}=.0037$, one-
tailed). Thus, it seems that the use of small-sample data to infer population statistics is prevalent, even when the whole population, or a large sample of it, is available when the inference is drawn.

## Experiment 5

The results of the first four experiments helped to establish that people are sensitive to the variability of populations, base their assessment of the variability of a population on samples drawn from it, and do not correct these assessments for the bias induced by the size of the samples used. As a result, they perceive variability as smaller than it actually is. Experiment 5 further explored the manner in which people employ samples available to them to determine variability. Unlike Experiment 3, in which we studied how people estimate variability in a situation involving a choice based on the difference between two means, Experiment 5 was designed to explore the behavior of people in a situation in which the variance within a single population is of consequence. Within this context, we intended Experiment 5 to answer the following questions: 1) Do people regard estimates based on larger samples as more reliable? 2) If people view a sample whose size exceeds their working-memory capacity, do they integrate all the available data or do they use only a subsample of it?

In this experiment, after seeing a sample from each of two populations (which, unbeknownst to the participants, were identical), the participants chose one of them to be used in a subsequent prediction task. Since the expected reward in tasks involving the prediction of a value from a single population is greater, the smaller the variability of that population, we expected participants to choose for the prediction task that population that seemed the less variable of the two. As the samples actually drawn from the two populations during the first phase always differed in size and almost always in variability (the latter owing
to sampling error and difference in sample size), we could use multiple regression analysis to assess the degree to which characteristics of the actually drawn samples (or parts thereof) accounted for the participants' choices.

Task and Design. The task consisted of three phases: The participants first drew samples from two populations, and then chose one of them to be used in the subsequent prediction task. Finally, they predicted - for a reward - the value of 20 items, drawn one at a time with replacement from the population of their choice. The study had a $2 \times 2$, between participants design, the factors being Type of Variable and Total Number of Sampling Trials. First, the pairs of populations differed in the items they contained: One pair consisted of items with one of two values (binary populations), and the other, of items with one of five values (multi-valued populations). Second, the total number of sampling trials (from both populations combined) was either 13 or 19. These numbers ensured that samples drawn from the two populations were of unequal size. Moreover, the total sample size was announced in advance, rendering the eventual difference in sample size highly prominent.

Participants. The participants were 160 Hebrew University undergraduate students. Some of them took part in the experiment as partial fulfillment of a course requirement, the others, for payment. All the participants also received a reward reflecting their success in the prediction phase.

Materials. Each population was stored in a tall, opaque urn, with an opening wide enough to insert a hand but not to see the contents. Each urn contained 36 items. For the binary-valued populations, each urn contained 36 beads -24 of one color and 12 of another. The beads in one urn were red and green, those in the other, blue and yellow. For the multivalued populations, the items were playing cards with face values ranging from 6 to10, with frequencies of $4,8,12,8$, and 4 , for the five values, respectively. The cards in one urn were
from the red suits, and in the other, from the black suits.
Procedure. The participants were tested individually, in a quiet room. They were informed of the three phases of the task (sampling, choice and prediction). They were also informed of the contents of the urns (beads or cards), and of the corresponding range of values. Finally, they were informed of the total number of sampling trials, and told that the urn to be sampled at each trial would be determined by rolling a die (the urns were prominently labeled "odd" and "even"). Note that this method could, and indeed did result in a wide range of values with regard to the number of items observed from each urn. Sampling proceeded, with replacement, until the total number of sampling trials was reached. Following the sampling stage, the participant chose which urn would be used in the subsequent prediction phase. In this phase, the participant predicted, for each of 20 trials, the value of an item to be drawn from the chosen urn. For the beads, a correct prediction was rewarded 0.5 IS (about $\$ 0.12$ ). For the cards, a correct prediction was rewarded 0.6 IS, and a prediction deviating by 1 from the value drawn was rewarded 0.3 IS. This scheme resulted in an expected reward of $1 / 3$ IS at each trial, both to participants who predicted, for beads, the more frequent value and to participants who predicted, for cards, the mean value. During prediction, items were also drawn with replacement. Prior to every draw, the urn was vigorously shaken.

## Results and Discussion

The characteristics of the actually drawn samples were entered as independent variables in multiple regression analyses using the stepwise method, to determine how well they accounted for the participants' choices. For all four conditions combined (with sample sizes and variances standardized within condition), the best combination of predictors turned out to involve the difference between sample sizes (whose correlation with choice was $\underline{r}=$
$.341, \mathrm{p}<.001)$ and the difference between the two standard deviations $(\underline{r}=-.169, \mathrm{p}=.033)$; with these, the multiple regression reached a value of $\underline{R}=.418(\underline{F}(2,157)=16.65, p<.001)$. It is of interest that, on average, the urn from which a larger sample had been drawn was the one chosen in .63 of the cases, and that the urn with the smaller sample variance was chosen in .54 of the cases; when one of the urns was both the one from which a larger sample had been drawn and the one with the smaller sample variance, the proportion with which it was chosen over the other was .69. These results indicate that people are aware that predictions based on larger samples are more reliable than those based on smaller samples. They also show that people are sensitive to differences in variability when they are of consequence.

The data obtained in the study were further analyzed to determine if people base their decision on all the data they have seen or only on a sub-sample. Given the large body of literature on primacy and recency effects (e.g., Hogarth \& Einhorn, 1992), we analyzed the quality of prediction using sub-samples taken either from the beginning or from the end of the actual sample. The size of the sub-samples varied from 6 to 13, in steps of 1. It turned out that, for each of the four conditions, it was possible to identify a subset whose data better predicted the participants' choices than did the sample as a whole. That subset tended to include the first (rather than the last) items; its median size was 10. Across all four conditions combined, the use of the first 10 items to predict choices not only did not impair our ability to predict participants' choices correctly, but in fact slightly improved it. The value of $\underline{R}$ was .434 , and the proportion of correct predictions rose from .675 , when all data were used, to .688, when data from only the first 10 items were used.

The results of this analysis indicate that people are indeed restricted in the size of the sample considered when the total number of available items exceeds their capacity. With a total of 10 items considered, this means that an average of five items from each population
were used - a value commensurate with estimates of working-memory capacity when data can be assigned to different categories (Mandler, 1967). The analysis also reveals that the first 10 items encountered are the more likely to be taken into account (i.e., a primacy effect). This is also in line with results reported in other studies comparing the strength of primacy and recency effects (see Hogarth \& Einhorn, 1992).

## Concluding Remarks

As noted in the introduction, it is widely assumed that, in order to function efficiently, people must acquaint themselves with the environment in which they operate. Implicit in the expression "must acquaint themselves" is the notion that the knowledge should be accurate, that perception of the environment should be a veridical reflection of reality. As we have also pointed out, observers of human behavior have long come to suspect that, contrary to this general assumption, the perception of variability is persistently biased, that people perceive the world as less variable and more regular than it actually is. The present set of experiments tested this suspicion by direct comparison of actual and perceived variability and found, for the first time, that this is indeed the case. Various measures - similarity judgments, choices and answers to direct questions - all yielded a coherent set of results: People are sensitive to variability, but perceive it to be smaller than it really is.

Our explanation of that misperception is that people use sample data to assess the degree of variability (with the size of the sample apparently related to their working-memory capacity), and do not correct their estimates for the bias due to sample size. In other words, we suggest that there is a simple relationship between people's misperception of variability and the size of the samples that they can and do employ in their assessment of variability. Since by the very nature of the sampling distribution of measures of variability, the samples that people encounter and consider, even when random, will be on the average downward
attenuated, people really see variability as lower than it actually is (in the population as a whole). It follows that, for the case of variability, the fault for the bias does not necessarily lie with faulty information processing. Rather than inaccurate processing of unbiased data, the reason for misperception of variability may be accurate processing of biased data. No biased processing need be postulated for the bias to emerge: Accurate processing of well-sampled data will inevitably lead to the bias.

Time constraints, memory capacity or lack of available data mandate the use of sample data to estimate population parameters. The only way to avoid the bias inherent in estimating variability from sample data, therefore, is to apply some correction for sample size. Our data indicate that such a correction is not applied: Otherwise our results would have been different. Most telling in this respect is the finding that the variance of the actually observed sample was a better predictor of people's behavior than sample variance corrected for sample size.

It should be noted that the process we propose is not the only one that can result in biased perception of variability. For example, people may have learned to disregard (or at least to discount) extreme values, since such values might belong to other distributions or reflect measurement error. ${ }^{4}$ The model of Huttenlocher and Hedges (1994, see also Huttenlocher, Hedges, \& Vevea, 2000) offers another explanation, as it suggests that recall of instances that belong to a category is biased towards the mean of that category. While further research may be required to determine which process best explains the results, this does not alter the main finding - that variability is perceived as smaller than it really is.

The variability of a population determines the confidence with which one can predict a value from it. The larger the variability, the larger the likely prediction error, hence the lower the confidence. Similarly, when the choice is between two populations, the larger the
variability of the populations, the lower the confidence. Hence, a downward attenuation of estimates of variability is bound to increase confidence. In turn, such an increase in confidence in one's decisions - even if objectively unwarranted - is likely to increase one's sense of control and optimism. Furthermore, a downward attenuation in the perception of variability could play a role in the formation of the fundamental attribution error and the illusion of control, and could explain over-confidence in the accuracy of statistics observed in small samples. Any misperception of the environment carries, of course, a price tag. The fact that people do not correct for their misperception of variability may indicate that its benefits outweigh its cost.

Our studies have established that the perception of variability is downward attenuated, and moreover that a possible cause of that bias is the use of sample data, uncorrected for the degree of bias due to sample size. As to the exact degree of attenuation, future studies will hopefully address this question, adding missing details to the picture whose broad outline we have started to sketch here.

## References

Bacon, F. (1620/1905). Novum Organum (trans. R. L. Ellis \& J. Spedding, edited by J. M. Robertson). London: George Routledge and Sons.

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. Psychological Review, 107, 659-676.

Frisch, D., \& Baron, J. (1988). Ambiguity and rationality. Journal of Behavioral Decision Making, 1, 149-157.

Gilbert, D. T., \& Malone, P. S. (1995). The correspondence bias. Psychological Bulletin, 117, 21-38.

Gilbert, D. T., Pelham, B. W., \& Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. Journal of Personality and Social Psychology, 54, 733-740.

Hays, W. L. (1963). Statistics for psychologists. New York: Holt, Rinehart, \& Winston.

Heath, C., \& Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. Journal of Risk and Uncertainty, 4, 5-28.

Hogarth, R. M., \& Einhorn, H. J. (1992). Order effects in belief updating: The beliefadjustment model. Cognitive Psychology, 24, 1-55.

Huttenlocher, J., \& Hedges, L. V. (1994). Combining graded categories: Membership and typicality. Psychological Review, 101, 157-165.

Huttenlocher, J., Hedges, L. V., \& Vevea, J. (2000). Why do categories affect stimulus judgment? Journal of Experimental Psychology: General, 129, 220-241.

Kahneman, D., \& Tversky, A. (1973). On the psychology of prediction. Psychological

Review, 80, 237-251.
Kareev, Y. (1995a). Positive bias in the perception of covariation. Psychological Review, 102, 490-502.

Kareev, Y. (1995b). Through a narrow window: Working memory capacity and the detection of covariation. Cognition, 56, 263-269.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlation. Psychological Review, 107, 397-402.

Kareev, Y., Lieberman, I., \& Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. Journal of Experimental Psychology: General, 126, 278287.

Langer, E. J. (1975). The illusion of control. Journal of Personality and Social Psychology, 32, 311-328.

Mandler, G. (1967). Organization and Memory. In K. W. Spence \& J. T. Spence (Eds.), The psychology of learning and motivation, Vol. 1, (pp. 328-372). New York: Academic Press.

March, J. G. (1996). Learning to be risk averse. Psychological Review, 103, 309-319.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63, 81-97.

Mischel, W. (1968). Personality and assessment. New York: Wiley.
Peterson, C. R., \& Beach, L. F. (1967). Man as an intuitive statistician. Psychological Bulletin, 68, 29-46.

Pollard, P. (1984). Intuitive judgments of proportions, means, and variances: A review. Current Psychological Research and Review, 3, 5-18.

Ross, L. (1977). The intuitive psychologist and his shortcomings. In L. Berkowitz
(Ed.) Advances in experimental social psychology, 10, 173-220. San Diego: Academic Press.
Shoda, Y., \& Mischel, W. (1993). Cognitive social approach to dispositional inferences: What if the perceiver is a cognitive social theorist? Personality and Social Psychology Bulletin, 19, 574-585.

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


#### Abstract

Authors Note This project was funded by the Israel Science Foundation (Grant No. 712/98). An earlier version of this paper was presented at the $40^{\text {th }}$ annual meeting of the Psychonomic Society, Los Angeles, November, 1999. Yaakov Kareev, Goldie Rotman Center for Cognitive Science and Education, School of Education, and Center for Rationality and Interactive Decision Theory, The Hebrew University, Jerusalem, Israel. Email: kareev@vms.huji.ac.il.


## Footnotes

${ }^{1}$ There is, in fact, indirect evidence that people use sample data to estimate variability, without correcting for sample size. Such evidence is forthcoming from studies that reveal differences in responses related to perceived variability, in cases in which sample size may be assumed different. One relevant line of research is the study of the effect of cognitive load on the attribution of people's behavior (Gilbert, Pelham, \& Krull, 1988); another involves the attribution of behavior by self and by a stranger (Shoda \& Mischel, 1993). In both lines of research, participants likely to have a larger sample at their disposal (unloaded, judging self) perceive variability as larger. Such a pattern of behavior could not have been observed, if people were taking note of sample size and correcting for its effects on sample variability. ${ }^{2}$ We would like to thank D. Stephen Lindsay for pointing out this alternative explanation. ${ }^{3}$ Another possible scoring of choices would be to score a choice as 1 when that population having the higher sample mean was preferred over the other. As one might expect, due to the large differences between population means, the difference in sample means was in the same direction as that between population means in most cases ( 229 out of 240 , with 3 other cases in which sample means were identical). Thus, the two scoring schemes would have yielded identical values in the vast majority of the cases. Indeed, the same measure turned out to be the best predictor for both scoring systems.
${ }^{4}$ We would like to thank Klaus Fiedler for having pointed out this possibility.

## Table 1

Proportion of Answers in which Populations Seen through Sample Only, or Out of Sight,
Rather than that in Full View, Were Judged as having the Smaller Variability.

| Task | Direct Evaluation |  | Choice |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Condition | Binary | Continuous | Binary | Continuous | Mean |
| Sample / <br> Population | .667 | .555 | .528 | .555 | .577 |
| Out of Sight / <br> In Sight | .611 | .639 | .528 | .528 | .577 |
| Mean | .639 | .597 | .528 | .542 | .577 |

Table 2
Characteristics of the Distributions of Number of Matches.

|  |  |  | Population |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| N of | Standard | I1 | I2 | I3 | I4 | II1 | II2 | II3 |  |  |
| matches |  |  |  |  |  |  |  |  |  |  | II4

Table 3a:
The Best Six Predictors of Choices and Bids - Absolute Correlations ( $\mathrm{N}=240$ ).

| Choice |  | Bid |  | Corrected Bid ${ }^{\text {a }}$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Predictor | Correlation ${ }^{\text {b }}$ | Predictor | Correlation ${ }^{\text {b }}$ | Predictor | Correlation ${ }^{\text {b }}$ |
| Signif. of $\underline{Z}$ | . 522 | Effect ( $\mathrm{r}_{\mathrm{pb}}{ }^{2}$ ) | . 302 | Effect ( $\mathrm{r}_{\mathrm{pb}}{ }^{2}$ ) | . 387 |
| Signif. of $\underline{t}$ | . 501 | $\underline{Z}$ | . 291 | t | . 377 |
|  | . 414 | t | . 290 | $\underline{\mathrm{r}}$ pb $^{\text {b }}$ | . 375 |
| $\underline{\mathrm{r}}$ b | . 406 | $\underline{\mathrm{r}}_{\mathrm{pb}}$ | . 282 | $\underline{Z}$ | . 374 |
| Var. t | . 367 | Signif. of t | . 255 | Signif. of t | . 334 |
| Var. $\underline{Z}$ | . 366 | Signif. of $\underline{Z}$ | . 237 | Signif. of $\underline{Z}$ | . 310 |

${ }^{\mathrm{a}}$ Bid corrected for individual differences in bid size.
${ }^{\mathrm{b}}$ For all correlations, $\mathrm{p}<.001$.
${ }^{c}$ The point biserial correlation, with estimate of variance corrected for sample size.

Table 3b
Intercorrelations Between Predictors ( $\mathrm{N}=240$ ).

|  | $\underline{\mathrm{Z}}$ | $\underline{\mathrm{t}}$ | Sig. $\underline{\mathrm{Z}}$ | Sig. $\underline{\mathrm{t}}$ | $\underline{\mathrm{r}}_{\mathrm{pb}}$ | $\underline{\mathrm{r}}_{\mathrm{pb}}(\underline{\mathrm{t}})^{\mathrm{a}}$ | Effect $^{\mathrm{b}}$ | Var. $\underline{\mathrm{Z}}$ | Var. $\underline{\mathrm{t}}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\underline{\mathrm{Z}}$ | - | .996 | .649 | .708 | .870 | .557 | -.917 | .241 | .244 |
| $\underline{\mathrm{t}}$ |  | - | .667 | .733 | .866 | .617 | -.902 | .253 | .249 |
| Sig. $\underline{\mathrm{Z}}$ |  |  | - | .986 | .837 | .730 | -.676 | .490 | .496 |
| Sig. $\underline{\mathrm{t}}$ |  |  |  | - | .845 | .796 | -.710 | .464 | .458 |
| $\underline{\mathrm{r}}_{\mathrm{pb}}$ |  |  |  |  | - | .568 | -.941 | .341 | .338 |
| $\underline{\mathrm{r}_{\mathrm{pb}}(\underline{\mathrm{t}})}$ |  |  |  |  |  | - | -.423 | .307 | .301 |
| Effect |  |  |  |  |  | - | -.189 | -.184 |  |
| Var. $\underline{\mathrm{Z}}$ |  |  |  |  |  |  | - | .999 |  |
| Var. $\underline{\mathrm{t}}$ |  |  |  |  |  |  |  | - |  |

${ }^{\text {a }}$ The point biserial correlation, with estimate of variance corrected for sample size.
${ }^{\mathrm{b}}$ Effect is $\underline{\mathrm{r}}_{\mathrm{pb}}{ }^{2}$.

## Table 4:

Mean Number of Items, out of 7, in the Larger of the Two Groups

|  | Out of View |  | In full View |  |
| :--- | :---: | :---: | :---: | :---: |
|  | Binary | Continuous | Binary | Continuous |
| Span Under 6 | 4.38 | 4.62 | 4.57 | 4.76 |
| Span Over 6 | 4.42 | 4.67 | 4.29 | 4.29 |

