Belief Clustering in Inferences about Correlation

Richard B. Anderson (randers@bgsu.edu)
Justin M. Gilkey (jgilkey@bgsu.edu)
Michael E. Doherty (mdoher2@bgsu.edu)
Department of Psychology, 206 Psychology Building, Bowling Green State University
Bowling Green, OH 43403 USA

Abstract

One important characteristic of human society is that individuals have intuitive beliefs about how various aspects of their environment (physical, social, etc.) correlate to other aspects. This paper tests the hypothesis that the mathematical environment can give rise to multiple clusters of beliefs when those beliefs concern the degree of co-relatedness between variables. Simulations were conducted demonstrating that when the sample size is extremely small (i.e., 3), the sampling distribution of correlations is either U-shaped (for distributions of the Pearson *r*) or W-shaped (for distributions of signed r^2). Behavioral data indicated distributions that tended to approximate a W shape. There was also evidence that when people guessed, because they felt they could not extract any useful information from the sample, they were biased to guess that the population correlation was zero. The findings support the hypothesis that a natural, multi-clustering of sample correlations leads to a multi-clustering of beliefs about the population correlation.

Keywords: beliefs; correlation; covariation; inference; judgment; representation; sampling distribution; statistics; polarization.

Introduction

One important characteristic of human society is that individuals have intuitive beliefs about how various aspects of their environment (physical, social, etc.) correlate to other aspects. People often differ with respect to such beliefs, and this sometimes leads to cooperation among like-minded individuals and conflict among those with disparate beliefs. This paper concerns a theory—one not described in any previous report—about how the mathematical environment can give rise to multiple clusters of beliefs when those beliefs concern the degree of co-relatedness between variables.

Building on the work of Kareev (e.g., 1995, 2005), Anderson, Doherty, & Friedrich (2008) examined the effects of sample size on correlation detection—i.e., the discrimination of zero from non-zero population correlations. Simulations results (Anderson et al., 2008), led to the prediction that detection should be better for large than for small samples, except under conditions of extreme decision bias. In that same report, behavioral data indicated that, indeed, detection accuracy was greater for large than for small samples.

But the simulations also produced an unexpected finding that was not directly related to the question of correlation detection. Specifically, when the population correlation was zero and the sample size was 3, the sampling distribution of correlation coefficients (*r*) was U-shaped, with 0 being the least frequent value of *r*. In the present paper, this pattern is interpreted to suggest that there may be a mathematical, computational basis to posit that when people are forced to rely on small, random samples in making correlation-based judgments, their beliefs will not be smoothly distributed over the range of possible beliefs. Rather, due to mathematical factors that are distinct from the qualitative content of the to-be-judged information, the observers' beliefs may tend to cluster systematically around multiple points—points that often correspond to inaccurate inferences about the correlation between variables. For example, suppose each of 100 people undergoes three unrelated surgeries over a 10-year period. Each operation is performed by a different surgeon, and the surgeons differ in age. Suppose further that the post-operative healing times have varied, from substantially shorter to substantially longer than expected. Finally, let us assume that there is no correlation (i.e., $\rho = 0$) between the surgeon's age and healing time for surgeries he or she has performed. Formally, this scenario represents a sampling distribution of 100 correlation coefficients, with each patient experiencing one sample correlation (*r*) defined by three *x*, *y* data pairs (*x* being age, and *y* being healing time). As demonstrated by Anderson et al. (2008), the shape of such a distribution is highly non-intuitive: If the sample size (*n*) were larger—e.g. 15 instead of 3—the distribution would be roughly normal and symmetrical about zero. But in the case of $n = 3$, the distribution of sample correlations is U-shaped, with clustering near $r = -1$ and $r = +1$, and with 0 being the *least* frequent value of *r* (Anderson et al., 2008). Under the assumption that each patient treats the sample *r* as the best estimate of the population correlation (ρ)—that is, the correlation between age and post-operative healing time for all surgeons—clustering of samples translates into a clustering of beliefs about the value of ρ. Moreover, in this particular case (with $\rho = 0$ and $n = 3$) the samples can be characterized as consistent with polarized beliefs: One cluster of patients perceives a strong positive relation between the surgeon's age and post-operative healing time, another perceives a strong negative relation, and relatively few believe there is no relationship.

In the present paper presents new simulation studies, along with the first behavior investigation of the hypothesis that the shape of the distribution of beliefs (about a population correlation) would reflect the shape of the distribution of sample correlations on which the beliefs are based.

Sampling-Distribution Simulations

A series of simulations was conducted to model situations in which each of a large number of individuals encounters a random sample (with replacement) of *x,y* data pairs. In comparison to the simulations in Anderson et al. (2008), the present simulations were manipulated across four levels rather than two, results assessed for two different (though related) measures of correlation, and the distribution shapes were interpreted in the context of a theory about the distribution of beliefs across persons.

Variables *x* and *y* were continuous, and each was normally distributed within the population from which the data pairs were drawn. The number of data pairs per sample and the population correlation were manipulated across simulations. For some simulations, each sample correlation was computed as the Pearson correlation coefficient (*r*), whereas in others it was computed as signed r^2 (*r* multiplied by the absolute value of *r*). Figure 1 shows the simulation results.

Of particular interest are the simulation results for n = 3.

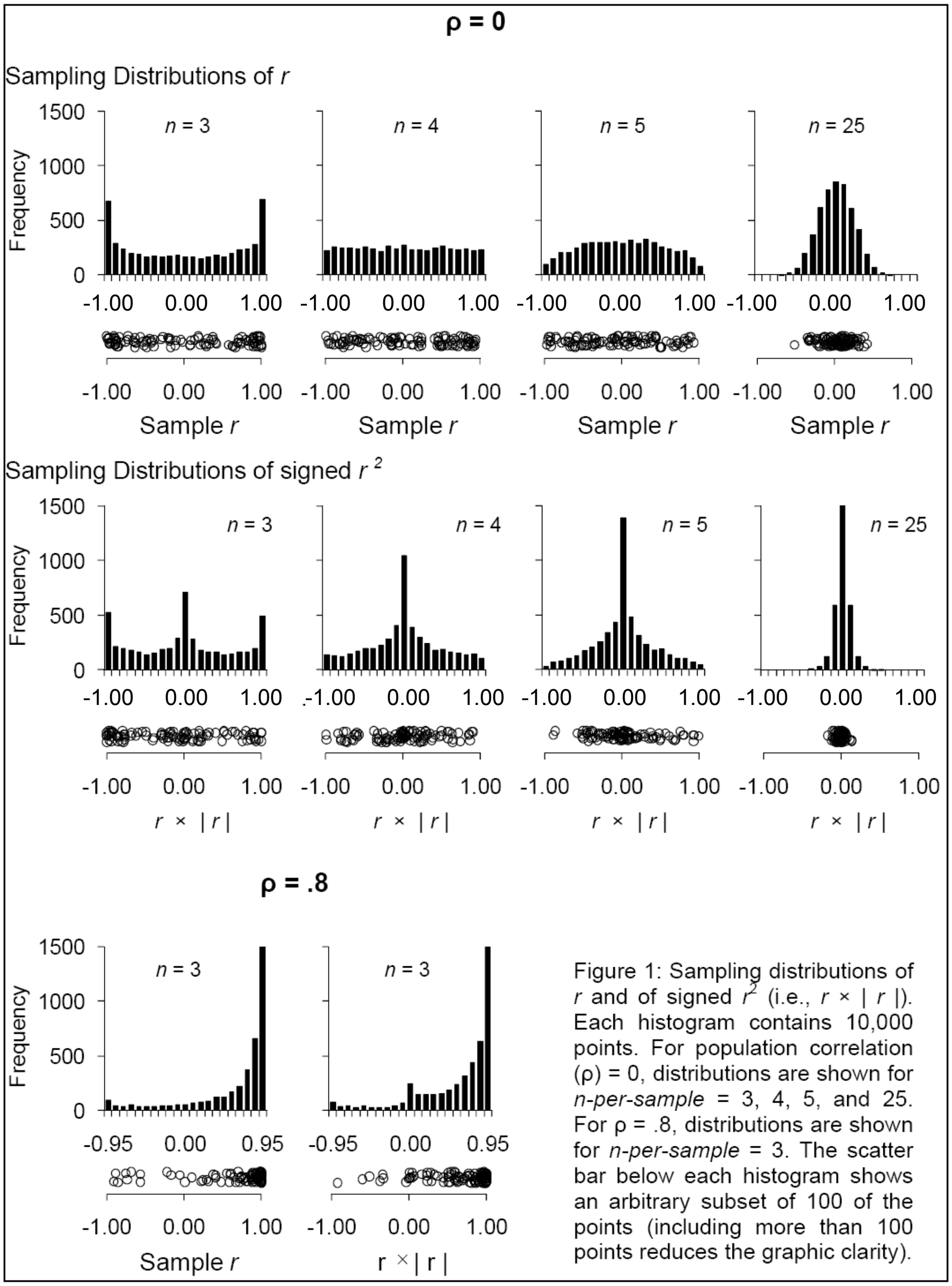

Figure 1: Sampling distributions of *r* and of signed r^2 (i.e., *r* × | *r* |). Each histogram contains 10,000 points. For population correlation (ρ) = 0, distributions are shown for *n-per-sample* = 3, 4, 5, and 25. For ρ = .8, distributions are shown for *n-per-sample* = 3. The scatter bar below each histogram shows an arbitrary subset of 100 of the points (including more than 100 points reduces the graphic clarity).

In such cases the distribution of sample correlations has multiple modes—two modes in the simulation that computes the sample r, and three modes in the simulation that computes signed r^2. Such multi-clustering is especially evident when the population correlation (ρ) is 0, but is also evident when ρ = .8 (for ρ = .8, the cluster at -1 is tiny yet discernable). These findings suggest that when individual perceivers are exposed to extremely small samples, their beliefs of about the correlation in the stimulus population may tend to cluster at the extreme ends of the scale, with a possible third cluster at zero. Moreover, the two-cluster outcome would suggest that people mentally represent correlation in a way that approximates the Pearson *r*, whereas a three-cluster outcome would be consistent with an approximation to signed r^2.

A Behavioral Study

Participants read a description about immigrants from a fictional country. The description included a random sample of three immigrants, each of whom was characterized by two variables: the distance of the immigrant's home town from the national border, and the amount of time the immigrant had to wait before being granted permission to emigrate. Participants then estimated a pair of conditional frequencies for a set of 100 hypothetical immigrants. They were asked: (1) "If you were to meet 100 more immigrants from the Soreltinas Republic, and if all of their home towns were FARTHER-than-average from the eastern border, then about ______ of the 100 would be expected to have a longer-than-average waiting time to emigrate," and (2) "you were to meet 100 more immigrants from the Soreltinas Republic, and if all of their home towns were CLOSER-than-average to the eastern border, then about ______ of the 100 would be expected to have a longer-than-average waiting time emigrate. For each participant, a subjective population correlation was computed from the two frequency estimates. It was predicted that the distribution of subjective correlations, across participants, would exhibit the patterns of clustering observed in the *n = 3* simulations.

Method

Participants and Design The experiment was conducted via the World Wide Web. Participants were randomly assigned to the ρ = 0 condition (128 participants) or the ρ = .8 condition (133 participants). Participants were recruited from psychology courses at Bowling Green State University, and received partial course credit in exchange for their participation. Participants were also solicited via publicly accessible web pages. People were asked to participate only if they were at least 18 years of age.

Procedure The experiment After reading an informed consent form 1, the participant was asked to indicate his or her sex, age, and academic status/affiliation [(a) undergraduate student, (b) graduate student, (c) college/university faculty, administrator, or staff member, (d) other]. Next, the participant read the following scenario, but with a random set of numerical values presented to each participant.

> Imagine that you have met three immigrants from the Soreltinas Republic, and that you have never met anyone else from that country. Each of the three immigrants has a different home town in the Soreltinas Republic. The towns vary with regard to how far they are from the eastern border. In addition, the immigrants had to wait different amounts of time between applying for emigration and being granted permission to emigrate. The table below contains the information about the three immigrants.

	Distance of Home Town from Eastern Border	Waiting Time for migration
Immigrant #1	5.52	373
Immigrant #2	5.38	270
Immigrant #3	4.29	187

For the two stimulus variables, the population means were 6 and 300, respectively, and the population standard deviations were 1 and 60, respectively. There were 500 data pairs in each of the two populations (ρ = 0 and ρ = .8).

Next, participants were asked to estimate the two population frequencies described earlier in this report. Finally, as part of a post-experimental questionnaire, participants were asked whether they tried to give accurate answers, whether they felt they had been able to make use of the table of numbers in making the frequency estimates (options were "yes", or "no, I just guessed"), how many statistics courses they had completed, and whether they were currently taking a statistics course.

Results and Discussion

Twenty participants in the ρ = 0 condition and 19 in the ρ = .8) confessed that they did not attempt to perform the task accurately. Their data were excluded from further analysis.

Demographics Table 1 summarizes the demographics for the participants.

Table 1: Summary demographic statistics indicating the number of males (M.) and females (F.), the percent of undergraduates (Undrgrd.), the mean number of statistics courses completed (#Stats.), and the percentage currently taking a statistics course (Curr. Stats.)

Group	M.	F.	Undrgrd.	#Stats. (mean)	Curr. Stats.
ρ = 0.0	38	70	89%	.6	13%
ρ = 0.8	37	77	96%	.4	11%

Stimulus Distributions Figure 2 shows sampling distributions of the correlations for the random samples shown to the participants. Though the number of data points was small relative to that of the simulations, the distributions of stimulus correlations are a rough match to the distributions produced by the simulations.

Behavioral Findings Figure 3 presents scatter bars and histograms showing the distribution of subjective correlations (range -1 to +1) in the ρ = 0 and in the ρ = .8 condition. On one hand, the results for ρ = 0 are not consistent with the simulations that assumed an internal representation of correlation scaled according to *r*, and that produced a u-shaped distribution. However, the behavioral data (for ρ = 0) do demonstrate small but clearly discernable clusters at -1 and +1 (on the subjective correlation scale) in addition to a large cluster at 0. This resembles the three-cluster pattern predicted by the signed r^2 simulation, though

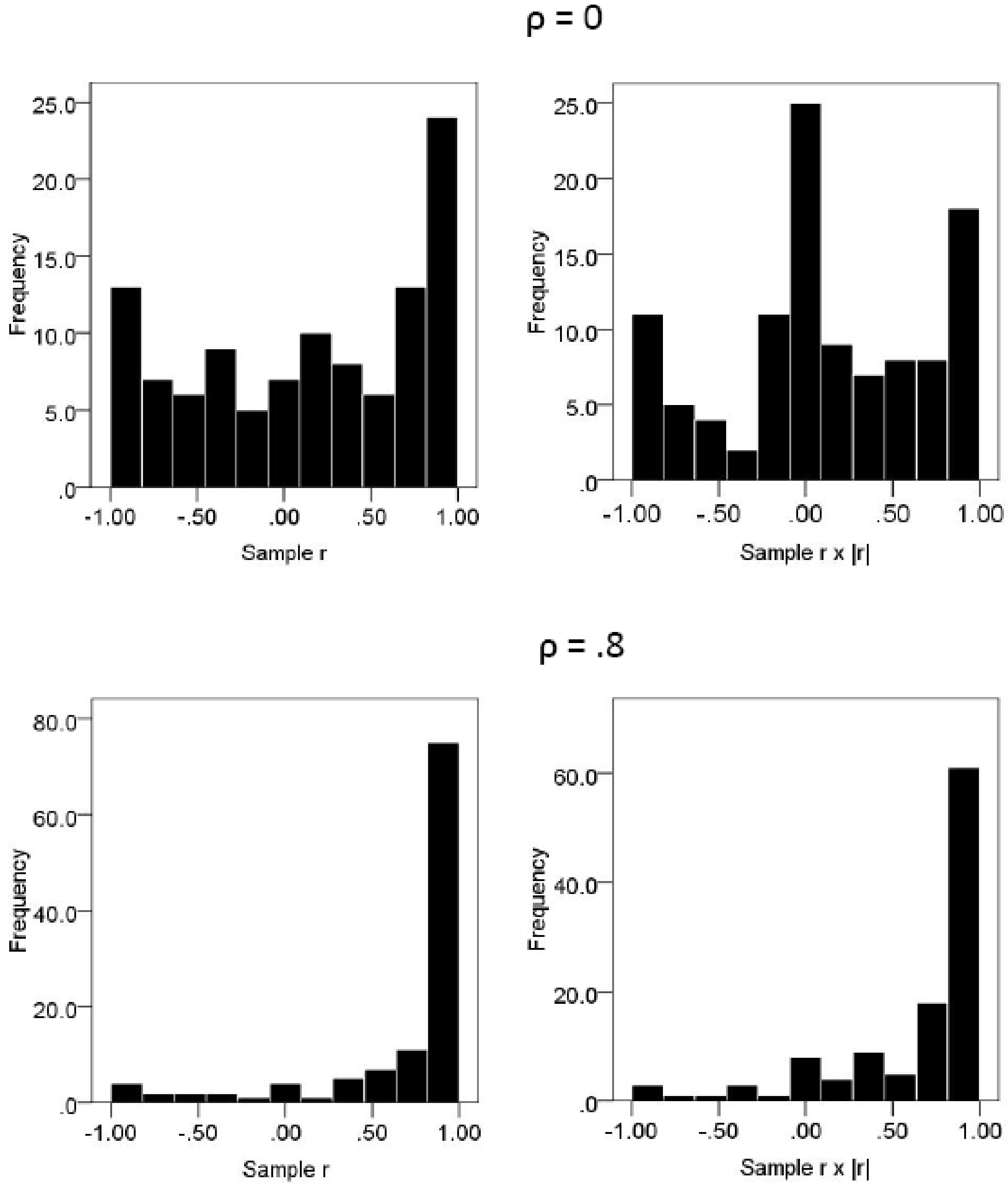

Figure 2: Distributions of sample correlations in the stimulus data.

the clusters at -1 and 1 are much smaller than predicted. This pattern (for ρ = 0) can be interpreted to indicate that, as predicted by the simulations, the output of the stimulus sampling process produced a pair of minority groups whose members inferred *extreme but* wrong beliefs about the correlation between an immigrant's home-town location and his(her) waiting time to emigrate. In addition, because the distribution of beliefs appears to depart from the simulation results (regarding the sizes of the clusters at -1 and +1), it appears that participants' internal representation of correlation may differ both from r and from signed r^2.

For ρ = .8 (Figure 3, bottom graphs), the result is similar to that of the signed r^2 simulation except that the cluster at zero is quite large relative to cluster at +1. This result can be interpreted to suggest that when the population correlation is strongly positive but not perfectly positive, people's beliefs cluster at two *incorrect* values: 0 and +1.

Notably, for ρ = .8, there is little clustering at -1, and the mean of participants' subjective correlations is significantly greater for ρ = .8 (M = .32) than for ρ = 0 (M = .15) [$t(220) = -2.7$, $p = .007$]. Thus, participants' inferences were sensitive to the objective correlation in the stimulus population. Another feature of the results is that participants in the ρ = 0 condition exhibited a positive bias (as in Kareev, 1995b, and Malmi, 1986) in that the mean subjective correlation was significantly greater than 0, $t(107) = 3.46$, $p = .001$.

A second set of analyses was conducted on a subset of the data, based on participants' reported metacognitive awareness of being able to utilize the sample information to form their beliefs about the population. Participants who answered "no" to the question whether they felt they were able to make use of the table of numbers in making the frequency estimates were excluded from this second set of analyses. Consequently, there remained just 60 and 56 participants in the ρ = 0 and the ρ = .8 condition, respectively. Figure 4 shows the results. As in the full data set, the mean subjective correlation was greater for ρ = .8 (M = .38) than for ρ = 0 (M = .11), $t(114) = -2.6$, $p = .01$. Another key finding is that the belief distribution in the ρ = .8 condition (Figure 4, bottom graph) appears to be a better match to the *signed* r^2 prediction (Figure 1, bottom right graph) than does the corresponding distribution from the full set of behavioral data (Figure 3, bottom right graph). The distribution shape is perhaps less clear for the ρ = 0 condition in Figure 4 (top graph), though the clarity might improve with a larger number of participants. It also appears that the number of subjective correlations lying at or near zero, relative to those lying at the extremes, is larger when the data set includes participants who felt that they simply guessed (Figure 3) than when such participants are excluded from the data (Figure 4). This suggests a particular bias in participants' guessing strategy such that, when faced with extreme uncertainty, they guess that the population correlation is zero rather than guessing randomly. Such a bias will of course impact the shape of the observed distribution of beliefs, as is evident in the difference between Figures 3 and 4. Thus it is important, theoretically,

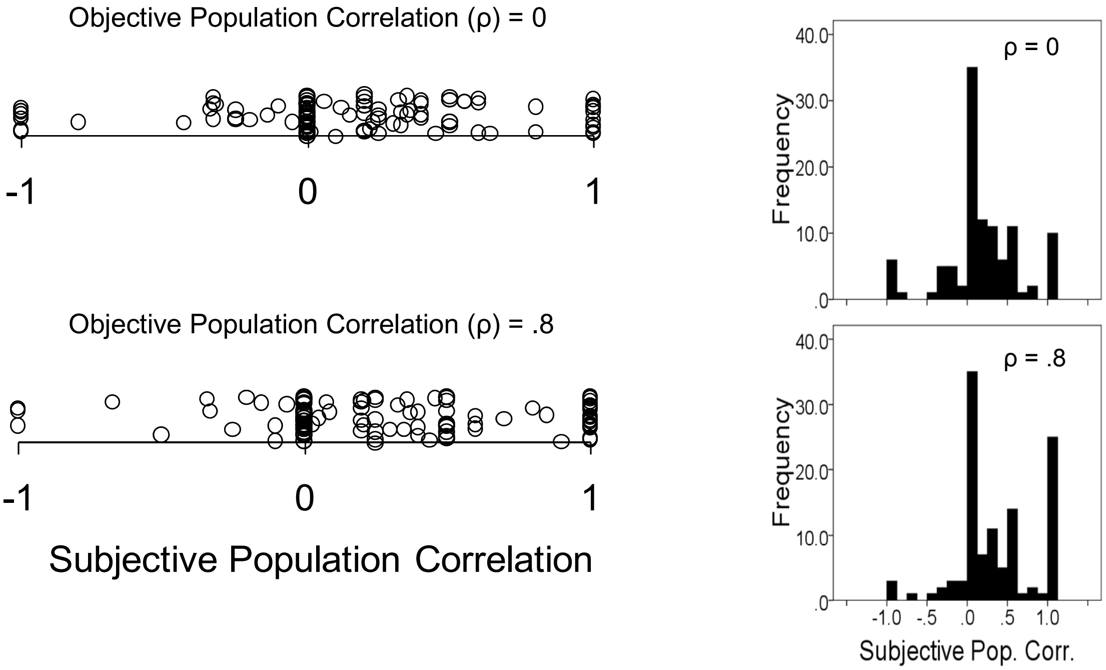

Figure 3: Distributions of subjective correlations. For the scatter bars on the left, the points are randomly displaced on the vertical dimension to minimize instances of complete superposition of points. The histograms on the right show the same behavioral data that are displayed in the scatter bars.

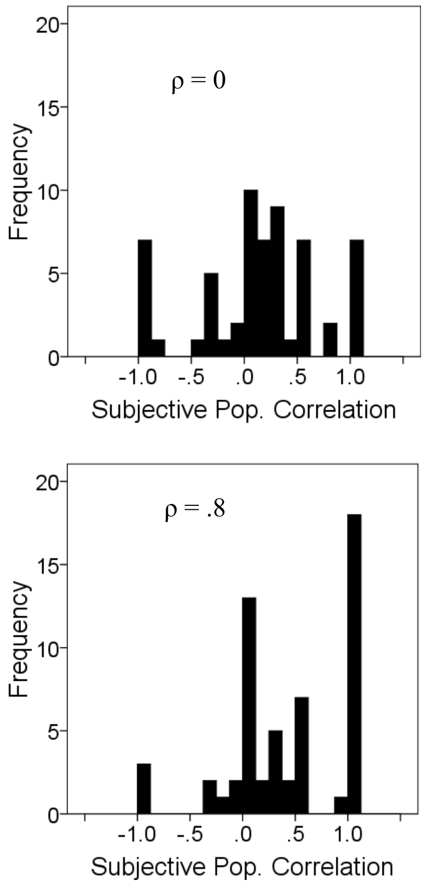

Figure 4: Distributions of subjective correlations for those participants who felt they were able to use the sample data to form a belief about the stimulus population.

to distinguish between decision makers whose metacognition tells them that their expressed belief has a basis in data, and those who feel that they were simply guessing and that their belief consequently has no basis in data.

In summary, the present findings indicate that when decision makers use very small samples to infer correlations, the resulting distribution of beliefs is not a simple scattering of points about the true population correlation. Rather, the belief distribution is multi-clustered in a way that reflects the statistical characteristics of the environment, and that suggests an internal mental representation that matches signed r^2 more than r.

Acknowledgments

We thank Leisha Colyn, Beth Hartzler, Andrew McCracken, Michael Carriaga, and Jennifer Diveto for their contributions to theoretical and methodological discussions related to this project. This research was supported in part by grant SES-0423825 from the National Science Foundation.

References

Anderson, R. B., Doherty, M. E., & Friedrich, J. (2008). Sample size and correlational inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34*, 929–944.

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

Kareev, Y. (1995b). Positive bias in the perception of covariation. *Psychological Review, 102*, 490-502.

Kareev, Y. (2005). And yet the small-sample effect does hold: Reply to Juslin and Olsson (2005) and to Anderson, Doherty, Berg, and Friedrich (2005). *Psychological Review, 112*, 280-285.

Malmi, R. A. (1986). Intuitive covariation estimation. *Memory & Cognition, 14*, 501-508.