Mem Cogn (2011) 39:107–116
DOI 10.3758/s13421-010-0008-5

The effects of problem content and scientific background on information search and the assessment and valuation of correlations

Shira Soffer · Yaakov Kareev

Published online: 4 November 2010

Abstract The effects of problem contents and one's scientific background on the detection of correlations and the assessment of their strength were studied using a task that required active data search, assessment of the strength of a correlation, and monetary valuation of the correlation's predictive utility. Participants (N = 72) who were trained either in the natural sciences or in the social sciences and humanities explored data sets differing in contents and actual strength of correlation. Data search was consistent across all variables: Participants drew relatively small samples whose relative sizes would favor the detection of a correlation, if one existed. In contrast, the assessment of the correlation strength and the valuation of its predictive utility were strongly related not only to its objective strength, but also to the correspondence between problem contents and one's scientific background: When the two matched, correlations were judged to be stronger and more valuable than when they did not.

Keywords Assessment of correlations · Information search · Predictive value

The ability to detect contingencies and correlations underlies inductive learning, category formation, and the discovery of cause and effect relationships; it is therefore essential for cognitive functioning. Once a relationship between antecedent conditions and subsequent outcomes or between co-occurring values is learned, the learner can use this knowledge to better control the present and predict the future. Given the importance of contingencies and correlations for cognitive functioning, it is no wonder that the way they are perceived has been the focus of extensive research (for review articles, see Allan, 1993; Alloy & Tabachnik, 1984; de Houwer & Beckers, 2002; Shanks, 2004). This research revealed that, by and large, people's assessment of the strength of correlations corresponds quite well to their actual strength, although some factors may bias that assessment.

It is important to note that the assessment of correlations requires the execution of a number of different cognitive activities, such as information search, its storage and retrieval, integration, assessment of the strength of the correlation, and, finally, using the correlation (see Crocker, 1981). However, the psychological study of this assessment has focused almost exclusively on only two of them: recalling of the evidence and integrating it. In a typical study, it is the experimenter who determines what items are to be presented and the strength of the correlation in the data presented. The participants' task is to assess the strength of that correlation (e.g., Shaklee & Tucker, 1980; Wasserman, Elek, Chatlosh, & Baker, 1993). In contrast, behavior in the first two stages of Crocker's model—data search and sampling behavior—has been only rarely studied in connection with the assessment of correlations. Behavior during these early stages has been studied mostly in research involving developmental changes in scientific inquiry skills (Klahr & Dunbar, 1988; Klahr, Fay, & Dunbar, 1993; Kuhn & Dean, 2005; Kuhn & Pease, 2008; Kuhn, Schauble, & Garciamila, 1992; Schauble, 1990, 1996; Schauble, Klopfer, & Raghavan, 1991), but these studies typically have required participants to reach a binary decision—to determine whether two variables are (causally) related—rather than to assess the strength of the relation-

S. Soffer (✉) · Y. Kareev
School of Education, The Hebrew University of Jerusalem,
Jerusalem 91905, Israel
e-mail: kareev@vms.huji.ac.il

ship. Finally, the stage at which the covariation estimate is used to make predictions—a stage that requires distinguishing between the strength of a correlation and its predictive value—has only recently been studied at all (Kareev, Fiedler, & Avrahami, 2009; Vadillo & Matute, 2007; Vadillo, Miller, & Matute, 2005), and that in contexts that did not require information search.

In the present study, we employed an integrated paradigm, starting with a free, self-terminated information search stage, continuing with the assessment of the strength of the correlation, and ending with the valuation of the correlation for prediction. The behaviors observed were used to find out what type of information and how much of it are deemed sufficient to stop sampling and pass judgment and what factors affect the assessed strength and subsequent valuation of the correlation. The participants in the study were students trained in the natural or the social sciences, and the correlations to be detected and judged were between pairs of binary variables that were characteristic of those encountered in the natural sciences, the social sciences, or everyday life.

The reason for systematically studying the joint effects of field of study and problem contents was that earlier research revealed that when the correlation whose strength is to be assessed is presented as one between meaningful variables, beliefs about the strength of that correlation are often the cause of biases that sometimes show up in judgments of correlation strength in the data presented (e.g., Alloy & Tabachnik, 1984; Fugelsang & Thompson, 2000; Jennings, Amabile, & Ross, 1982; Koslowski, Okagaki, Lorenz, & Umbach, 1989). Still, it should be noted that although beliefs may affect the judgment of correlations and even distort it, the effect does not necessarily reflect faulty performance; it may, rather, reflect the plausible integration of newly acquired data and preexisting knowledge (e.g., Chater & Oaksford, 2006; Griffiths & Tenenbaum, 2009; McKenzie & Mikkelsen, 2007; Oaksford & Chater, 1994, 2003; Tenenbaum, Griffiths, & Kemp, 2006).

In view of the earlier findings of the effects of prior knowledge, we expected the correspondence between the subject matter of a data set and the field of study of a person to affect some or all stages of the detection of correlations and the assessment of their strength. To find out, we explored to what extent the correspondence between problem contents and the participants' academic training would affect their data search, assessment, and evaluation. We expected participants to be more familiar with correlations between variables commonly encountered in their own field of study and, hence, to perceive the relationship between them as stronger than that between variables more remote from their own field of study. We also wondered whether content familiarity would be related to the amount of data collected, the strength of correlation when sampling was terminated, or the valuation of predictive power.

It should also be noted that the distinction between one’s field of study and a field of study that one is less familiar with may map onto the well-known distinction between experts and novices. If that is the case, expert–novice differences in recall, although observed in different domains and with more complex stimuli (Chase & Simon, 1979; Simon & Gobet, 2000; Vicente, 1992) could lie behind experience-related interactions, if observed. Expert–novice differences in cognitive accessibility of interpretations (Lau, Smith, & Fiske, 1991) or in the ease with which cognitive representations of scripted activities may be applied in imposing an organization upon sequences of events (Pryor & Merluzzi, 1985) could also be relevant in this respect.

Another variable that has been much studied and has been found to affect the judgment of correlations is outcome density—the frequency with which a focal outcome occurs, relative to its nonoccurrence (for a review, see Shanks, 2004). Previous research indicates that greater prevalence of the focal outcome tends to amplify estimates of the strength of a correlation. This finding, long considered a challenge to association-based theories of contingency learning, has received much attention recently (Allan, Hannah, Crump, & Siegel, 2008; Allan, Siegel, & Tangen, 2005; Crump, Hannah, Allan, & Hord, 2007; Perales, Catena, Shanks, & González, 2005). This research revealed that outcome density affects participants' decision criterion, but not their sensitivity to the strength of the correlation. To follow up on this line of research, we employed throughout the study binary variables that differed in the frequency of their values, with one of the two values being more prevalent (70%) than the other. This preplanned imbalance in the relative frequency of the variables' values enabled us to address issues related both to data sampling and to the valuation of the correlation.

With respect to sampling, a recent analysis (Kareev & Fiedler, 2006) revealed an inherent tension between accurate estimation and the chance of detecting a correlation, when one exists. When one's goal is to obtain an unbiased estimate of the correlation in the population, one should sample data in a way that preserves the makeup of the data set (e.g., random or proportional sampling). When one's goal is to enhance the chances of detecting a correlation if one exists (i.e., increase the power of the test), one would be better off if each subgroup is equally represented in the sample, rather than being represented in line with its relative size in the data set. However, the latter mode of sampling results in systematically biased sample correlations. To see this point, consider the examples in Table 1. Suppose that the true probabilities of two joint binary events are those appearing in Table 1a. The

Table 1 Tables of frequencies and the resulting strengths of correlations with a skewed original distribution (1a) and following the drawing of equal sample for one of the variables (1b)

	C1	C2	Total
1a – φ = .375			
P1	70	10	80
P2	10	10	20
Total	80	20	100
1b – φ = .411			
P1	44	6	50
P2	25	25	50
Total	69	31	100

correlation between the two variables is φ = .375.[1] The probabilities in Table 1a and a correlation of .375 are also the values that would be expected if one were to draw random samples from the population. Suppose, however, that one wants to sample an equal proportion of cases, either by the row or by the column variable. The resulting expected values (a .5/.5 distribution is imposed on the relative frequencies of one of the variables, and the values of the other reflect their original relative frequencies) appear in Table 1b. For these values, the correlation is stronger, at φ = .411. The analysis carried out by Kareev and Fiedler revealed that an amplification of the correlation is to be expected in a large majority (93%) of the sample space if sample relative frequencies are closer to a .5/.5 division than are the corresponding population values.

The sampling behavior in Kareev and Fiedler's (2006) study indicated that participants valued power more than accuracy of estimation. The design of the present study allowed participants to control not only the size, but also the makeup of the sample they observed; we could therefore use our data to find out what sampling techniques were adopted and whether they varied as a function of the participants' field of study and their familiarity with the content of the problem.

Almost all the studies mentioned thus far have focused on the perception of the correlation—the accuracy with which its strength is judged or assessed. It should be kept in mind, however, that the importance of detecting a correlation derives from the way in which it might be utilized: increasing the likelihood of choosing the right action or the better option. Obviously, the strength of a contingency and its utility are closely related, but for binary variables, the relationship is not perfect (see note 1). The implications of the dissociation between the strength of a contingency and its predictive power have only recently become the focus of empirical research (Kareev et al., 2009; Vadillo & Matute, 2007; Vadillo et al., 2005). This research demonstrates that people distinguish between the strength of the correlation and its predictive power and flexibly rely on that source of regularity—whether base rate of outcomes or the contingency—whose use maximizes the accuracy of their predictions. To further our understanding of the distinction, we used the random-pricing mechanism (Becker, DeGroot, & Marschak, 1964; see the Method section) to elicit our participants' evaluation of the monetary value of the correlation for the performance of a for-reward prediction task. The use of the technique added a dimension hitherto lacking from studies of correlations. In addition, the use of data sets consisting of unequal number of items bearing each value enabled us to explore whether the evaluation of the relative utility of base rates and correlations is differentially affected by one's familiarity with the contents of the variables.

Method

The participants' main task was to assess the strength of the correlation between two binary variables. The procedure consisted of three distinct stages: sampling of items, estimating the strength of the correlation, and indicating the monetary value of participating in a prediction task with the same set of items. The procedure was repeated 3 times, once for every one of the content areas employed. As is described below (see the Materials section), in each content area, one of the two variables could be construed as a cause, the other as an effect. To control for this difference (see Waldmann, 2001), the data sets were organized by the "cause" variable for half of the participants and by the "effect" variable for the other half.

Procedure

Participants were tested individually in a quiet room. Upon arrival, they were read the instructions and then performed the three stages of the task. To illustrate, we present the instructions for one of the conditions:

> Researchers studied a certain gene and found that some people have a normal form of it, whereas others have a mutation of it. They also found out that some people have a normal level of cholesterol, whereas for

[1] For continuous variables, the strength of a correlation, r, and its utility, as expressed in the proportion of variance accounted for, r^2, may be viewed as one and the same. In contrast, for binary variables, a correlation may differ from zero and nevertheless be useless for predictions; this situation occurs when the proportion of cases in the more common diagonal is smaller than that in the more common criterion value (Kareev, 1995, 2000; Kareev et al., 2009). For example, in a 2 × 2 table with cell frequencies of $a = 50$, $b = 10$, $c = 30$, and $d = 20$, the correlation is not zero, but a person is better off ignoring the correlation and always predicting the more common outcome.

> others it is above the normal level. The research was conducted with 100 people.
>
> We shall present you with two groups: One group consisting of people whose cholesterol level was normal, and the other consisting of people whose cholesterol level was high.
>
> Your task is to determine if there exists a relationship between cholesterol level (normal/high) and the gene (normal/mutated).
>
> Here is a file with the data: Green cards represent people with a normal level of cholesterol and yellow cards represent people with a high level of cholesterol. Every card has it written on its back if the person has the normal or the mutant form of the gene.
>
> Sample cards any way you wish, then mark on a scale between 0 and 100 the strength of the relationship between the variables, with 0 indicating no relationship at all, and 100 indicating a perfect relationship. [See below for the contents of all problems]

Both stacks were shuffled while the instructions were read. Participants were completely free with respect to which stack and what part of it to draw the next card from and how many cards to draw in all. Once a card had been drawn, it was turned around to check the value written on its back and was then put aside. Written notes were not allowed. Sampling continued until the participant was ready to provide an estimate of the strength of the correlation in the data set just observed. To indicate his or her estimate, the participant placed a mark on a 100-mm line. The left end of the line had the number 0 and the words "no relationship" written next to it; the right end of the line had the number 100 and the words "perfect relationship" written next to it.

Next, the cards sampled were returned to their original stacks, which were then reshuffled, and the participant was read the instructions for the valuation task, which were as follows:

> You will take part now in an additional task, in which you will be asked to predict the value written on the back of ten cards, five of each color, which will be sampled randomly. Every time you will be right you will get 1 Shekel. To repeat, you will sample at random five cards from each stack, and predict the value on the back of each of these cards.
>
> We shall deal with the scenario in which the green cards represent people whose cholesterol level is normal and the yellow cards represent people whose cholesterol level is high. For each of the ten cards you will predict if the gene for that person is in the normal or mutant form, and will get 1 Shekel if you are right.
>
> Before we start with the prediction task I would like to ask you the following: What amount of money would you demand to give up participation in this task? After you announce that amount you will sample, without looking, one of the ten discs which are inside this bag [pointing to an opaque bag], which bear the numbers 1 through 10. If the value on the disc you draw is larger than or equal to the amount you have asked, you will receive the value written on that disc. If the value on the disc you draw is lower than the amount you have asked, you will perform the prediction task.
>
> If you think about it, you will realize that the amount you ask for should be your true value for participating in the prediction task, because then, if the value you draw is higher than or equal to the value you asked, you will receive the value drawn, and if the value drawn is lower than the value asked, you will perform the task.
>
> For what amount would you give up participation in the prediction task?

At this point, the participant announced the amount asked for and drew a disc. If the value written on the disc was larger than or equal to the amount asked for, the participant was credited with the value of the disc; if the value on the disc was lower than the value asked for, the participant drew ten cards, five from each stack, predicted for each the value on its back, and was credited with 1 shekel for each correct prediction.

The method employed to elicit the value of the task was the random-pricing mechanism (also known as the BDM method) suggested by Becker et al. (1964), which is widely used in economics.

This sequence of sampling, assessment, valuation, and, if necessary, prediction, was repeated three times—once for every problem content. A Latin-square design was used to balance the order of problem contents, with an equal number of participants encountering problems in each of the three orders. Correlation strength varied between participants.

Materials

For each of the three tasks, the data set consisted of 100 opaque plastic cards, arranged in two stacks. One of the two stacks had 70 cards of one color (e.g., green), signifying one value of the predictor (e.g., normal cholesterol level); the other stack had 30 cards of another color (e.g., yellow), signifying the other value of the predictor variable (e.g., high cholesterol level). Other than their color, which was clearly visible, the cards had no distinguishing mark on their top. Each card had a value of the criterion variable written down on its bottom side (e.g., the normal or the mutant form of the gene).

The two stacks presented made up a data set. There were 18 data sets in all; they differed in content (three values; see below), in the strength of the correlation between their variables (three values, see below) and in the variable the values of which were represented by the stacks (two values; the "causal" or the "outcome" variable). Every participant performed the task 3 times, with sets that differed only with respect to their contents.

Design

The effects of three independent variables were studied in the experiment: content area (which was manipulated within participants), strength of the correlation, and field of study (which were between-participants variables).

Content Area The content of the problem assumed one of three values. The "natural science" problem involved the relationship between the status of a gene (normal or abnormal) and cholesterol level (normal or high). The "social science" problem involved the relationship between tribal affiliation (tribe A or tribe B) and burial customs (statues placed or not placed in graves). The "everyday" problem involved the relationship between the shape of the cap of a newly introduced soft drink (round or oval) and the evaluation of its taste (tasty, not tasty).

Strength of the correlation Strength of the correlation assumed one of three values[2]—none (φ = .00), medium (φ = .38), or strong (φ = .76). The frequencies of the items for the three strengths of correlation appear in Tables 2a–c.

Field of study The participants' field of study was either the natural sciences or the social sciences and humanities. In the case of students with a double major, both majors had to be from one of the two fields for the student to be included in the sample. It should be noted that at the Hebrew University of Jerusalem, where the study took place, not only do students take essentially all their courses in the field they major in, but the faculty of natural sciences is located at a campus different from that of the social sciences and humanities.

Table 2 Tables of frequencies of each combination for the three strengths of correlation

	C1	C2	Total
2a – φ = .00			
P1	49	21	70
P2	21	9	30
Total	70	30	100
2b – φ = .38			
P1	57	13	70
P2	13	17	30
Total	70	30	100
2c – φ = .76			
P1	65	5	70
P2	5	25	30
Total	70	30	100

Participants

The participants were 72 students (36 males and 36 females) from the Hebrew University of Jerusalem, who volunteered to participate in the study for remuneration. Half of them were recruited from the Givat Ram (Safra) campus of the University, which houses the natural sciences and the other half were recruited from the Mt. Scopus campus, which houses social sciences and humanities. Each participant's field of study was ascertained before he or she was permitted to participate in the study. There was an equal number of males and females in each cell of the design.

Results

The results will be reported in two sections: one reporting the participants' sampling behavior, the other reporting the assessment of the strength of the correlation and its valuation.

Sampling behavior

Since participants had complete freedom with respect to the number of items sampled, which stack to draw them from, and in what order to do that, we deemed it of interest to analyze each of these measures. The strength of the correlation in the sample at the point sampling was stopped was also treated as a dependent measure.

Sample size In previous studies in which participants were free to sample as many items as they wished before making a judgment or a consequential choice, the median sample size was typically between 15 and 20 items. It was 15 in the

[2] If the cell frequencies in a 2×2 table, starting with the upper left corner (e.g., Table 1a) are *a, b, c,* and *d,* the correlation between the two binary variables is $\varphi = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$.

study by Hertwig, Barron, Weber, and Erev (2004), in which participants had to draw items from two populations, then indicate which they regarded as superior. It was 17 in the study by Weber, Shafir, and Blais (2004), in which participants' task was similar to that employed by Hertwig et al.. In the task closest to the present, the one by Kareev and Fiedler (2006), the mean sample size was 20.68. In the present study, the median sample size was 20, and the mean was 22.10. Sample size was not significantly related to any of the other variables. Apparently, in this task, which called for the assessment and valuation of the correlation, rather than simply detecting whether one exists, the participants' most important consideration was to obtain a reliable estimate, and they felt that a sample of about 11 items from each stack was sufficient for that. The fact that none of the variables was related to total sample size and that the size of the sample drawn form each stack (see below) was close to that previously shown to provide a reasonably accurate estimate of measures of central tendency (Johnson, Budescu, & Wallsten, 2001) indicates that, at least for adults, estimation routines may be quite impervious to situational factors.

Sample composition Sample composition—the proportion of items sampled from either stack making up the data set—is not of much interest when the task calls for an eventual choice of one of them. It is, however, of great interest when the task calls for the assessment of correlation strength. This is so because only proportional sampling would provide the sampler with an unbiased estimate of the strength of the correlation, whereas drawing samples in which each value is about equally represented is likely to result in the observation of a biased, amplified sample correlation (Kareev & Fiedler, 2006). In line with earlier findings (Fiedler, Brinkmann, Betsch, & Wild, 2000; Kareev & Fiedler, 2006), the present data revealed that fully 47.5% of the items in the sample were drawn from the smaller stack.[3] This value is very close to 50% and much higher than 30%—the relative size of the smaller stack. A test of the deviation of the observed proportion from .30 yielded $F(1,66) = 624.39$, $MSE = .011$, $p < .001$, $\eta_p^2 = .904$. In other words, as with the behavior observed in previous research, the participants exhibited a strong tendency to draw samples in which each value was about equally represented. The tendency was pervasive, with items from the smaller stack overrepresented in all the conditions. In addition, there was a significant interaction between problem contents and correlation strength, $F(4, 132) = 2.73$, $MSE = .004$, $p = .032$, $\eta_p^2 = .076$, which was mostly due to the fact that, for the day-to-day problem content, the tendency to draw samples of equal size was least pronounced when the correlation in the population was .00 and most pronounced when it was .38. We find this interaction difficult to explain and of little interest.

Switching between stacks The number of runs is an indicator of how the participants went about accumulating the data. At the one extreme, participants could switch, on every trial, from one stack to the other. Such sampling style, which would result in a number of runs equal to the number of items sampled, would indicate an attempt to have, at every stage, samples of equal or almost equal size from either stack ("parallel search," according to the terminology used by Shaklee & Fischhoff, 1982). At the other extreme, participants could first draw as many items as they wished from one stack, then switch to the other and repeat. Such sampling style, which would result in only two runs, would be indicative of a predetermined sampling strategy.

An analysis of the number of runs revealed that, on average, there were 4.83 runs per sample. With the average number of items per sample being 22.10, the average run length was 4.58. Apparently, on average, participants were not adhering to either of the two pure strategies. Still, a closer inspection of the number of runs revealed that fully 45% of the cases consisted of two runs and another 12% consisted of 3 runs. Thus, it seems that in about half of the cases, participants did adopt a pure sampling strategy. This result is in line with the findings of Shaklee and Fischhoff (1982). Number of runs was not related to any of the other variables.

Termination The participants faced the tasks of assessing the strength of the correlation and valuing its predictive utility, not of deciding whether or not a correlation existed in the population. As a result, the strength of the correlation at the point of termination cannot be taken to represent some threshold. Still, we thought it could be instructive to find out whether any of the independent variables turned out to be related to that value. To this end, we calculated the value of φ in the sample at the point at which the participant terminated the drawing. Not surprisingly, that value was found to be related to the actual strength of the contingency, $F(2, 63) = 125.11$, $MSE = .066$, $p < .001$, $\eta_p^2 = .799$. Again, as was the case with other measures, none of the variables was related to this measure. Apparently, at least for college students, data-sampling behavior reflects the operation of well-established and consistent information search strategies.

Assessment and valuation

Assessment of the correlation Once they stopped sampling, the participants indicated their assessment of the strength of

[3] The median was 48.4%; the mode was 50.0%.

the relationship they had observed. They did that by placing a mark on a line, with the distance (in millimeters) of the mark from the end of the line labeled "no relationship" representing the estimated strength. As was to be expected, the estimates were strongly related to the actual strength: The mean values for the .00, .38, and .76 correlations were 45.42, 69.10, and 88.60 mm, respectively, $F(2, 66) = 73.53$, $MSE = 457.88$, $p < .001$, $\eta_p^2 = .690$. This result is, of course, mostly a manipulation check. It is of interest, though, to note that the assessment of the strength of correlation in the data set with a zero correlation deviated significantly from zero, $F(1, 22) = 166.79$, $MSE = 890.41$, $p < .001$, $\eta_p^2 = .883$. Such biased assessment could be the result of an alignment bias (Kareev, 1995), given that both variables in the data set were unevenly divided. Another possibility is that participants' estimates reflected the fact that expected prediction accuracy—the likelihood of correctly predicting the criterion value—was better than chance even when the correlation was zero. As was observed by Kareev et al. (2009), under such a condition, people tend to judge the correlation as above zero.

Unlike its lack of effect on sampling behavior, here the analysis revealed three significant effects in which problem content was involved. First, there was a main effect of problem content, with the correlations whose content was related to the natural sciences judged as stronger than those related to the social sciences and everyday problems (mean values being 71.46, 66.96, and 64.69, respectively), $F(2, 132) = 4.31$, $MSE = 198.17$, $p = .015$, $\eta_p^2 = .061$. Second, problem content also interacted with correlation strength, $F(4, 132) = 2.84$, $MSE = 198.17$, $p = .027$, $\eta_p^2 = .079$; this interaction was mostly due to differences in the assessment of the strength of the relationship, when actually there was none. When the evidence itself was weak (the zero correlation case), people apparently expected stronger contingencies in the sciences, with the natural sciences correlation judged as stronger than that of the social sciences, but with the everyday life contingency judged as considerably weaker.[4]

Third, and of particular interest to us, was a very strong interaction between problem content and the participants' field of study, $F(2, 132) = 19.05$, $MSE = 198.17$, $p < .001$, $\eta_p^2 = .224$. Students from each faculty judged the correlation of the problem whose content was related to their field of study as strongest, that of the problem from the other science as weakest, and that from the everyday area as falling in between. Students from the natural sciences judged the strength of the gene–cholesterol correlation as 79.31, that between cap shape and taste as 62.67, and that between tribe and grave contents as 60.69. In stark contrast, students from the social sciences judged the strength of the three correlations to be 63.61, 66.72, and 73.22, respectively. These results are particularly striking because problem content was a within-participants factor, with the actual strength of the correlation identical for all three problems. The best explanation of this bias in judgment is offered by invoking prior beliefs or the effects of believability (e.g., Alloy & Tabachnik, 1984; Fugelsang & Thompson, 2000; Griffiths & Tenenbaum, 2009; Jennings et al., 1982; Koslowski et al., 1989; Tenenbaum et al., 2006). Apparently, our participants' prior experiences with correlations in their own field of study affected and amplified their estimate of the strength of the correlation in their own field and attenuated their estimate of that in the less familiar field.

Valuation of the predictive power of the correlation As was discussed in the introduction, the predictive power of a correlation is an aspect that is conceptually different from its strength. To find out the correspondence between the objective and subjective valuations of the predictive power of the correlations we employed, we had our participants indicate the price that they requested in return for forsaking the prediction stage. Since, during this stage, participants were to draw five cards from each stack and be awarded one shekel (about $0.25) when correctly predicting the value written on the cards' backs, the expected value of the task, assuming maximizing behavior, was 7.00, 6.90, and 8.81 NIS for φ values of .00, .38, and .76, respectively (see Tables 2a–c).[5]

The average prices requested differed in relation to the strength of the correlation in the sets. They were 6.46, 6.97, and 8.19 for the φs of .00, .38, and .76, respectively, $F(2, 66) = 14.14$, $MSE = 4.05$, $p < .001$, $\eta_p^2 = .300$. These values corresponded well to the objective ones. In addition, there was a main effect of field of study, with participants schooled in the natural sciences demanding a higher price than their social sciences counterparts (7.49 vs. 6.93 shekels), $F(1, 66) = 4.25$, $MSE = 4.05$, $p = .043$, $\eta_p^2 = .061$). Most interesting, we also observed a highly significant interaction between field of study and content, $F(2, 132) = 9.84$, $MSE = 2.05$, $p < .001$, $\eta_p^2 = .130$, with the natural sciences students demanding most (8.19) for the gene–cholesterol problem and least (6.94) for the tribe–grave problem. The social sciences students, in contrast, demanded least (6.53) for the gene–cholesterol

[4] It is understood that this result is potentially related to the specific contents used for the three correlations.

[5] The first value could be achieved by always predicting the more common value of the variable serving as the criterion; the other values could be achieved by differentially predicting for each value of the predictor its more common value of the criterion. It should be noted that to achieve these values, the participants would have to engage in maximizing behavior.

problem and most (7.39) for the tribe–grave problem. The demand for the everyday problem fell in between the other two for students from both faculties. This interaction points in the same direction as that observed with the direct assessment of the strength of the correlations. Still, whereas the assessment of the correlation was costless, the prices requested have clear monetary implications: Asking for too much or too little to forego the prediction task would have reduced the responder's expected profits. Nevertheless, the results show that the participants' prior beliefs influenced not only their estimates, but also their requests—an "honest" indication of how deeply rooted is the tendency to view a correlation as stronger and more useful in a familiar context than in an unfamiliar one.

Actual profits were only slightly lower than the value that could have been obtained had the participants employed a maximizing strategy: The average overall profit was 7.18 shekels, slightly lower than the 7.57 shekels the participants would have earned had they identified the maximizing behavior called for in each case and engaged in it. These results are very much in line with those observed by Kareev et al. (2009), in which participants also exhibited maximizing behavior when engaged in for-profit predictions. In the present study, profits were, of course, the result of a combination of offers made by the experimenter –when they were above the asking price –and the number of correct predictions made when these offers fell short of that price. Still, the correspondence between actual profits and maximal expected profits is noteworthy. An analysis of profits revealed only one significant effect, that of the strength of the correlation, $F(2, 66) = 13.96$, $MSE = 3.84$, $p < .001$, $\eta_p^2 = .297$, with the mean profits being 6.58, 6.78, and 8.17 for the φs of .00, .38, and .76, respectively.

Discussion

The main object of the present study was to provide a description of the full range of activities—information search, integration, and evaluation—involved in attempting to assess the strength of the relationship between binary variables. The task was set up such that the participants had complete freedom with regard to which and how many items to sample prior to indicating their assessment of the strength of the correlation and their valuation of its predictive power. The variables used were all meaningful, and the distribution of their values unequal. By using different contents, we could study the effects of the correspondence between prior training and problem contents on sampling and assessment behavior. The use of unequal frequencies enabled us to find out whether sampling behavior was more conducive for the accurate assessment of the strength of the correlation or the speedy detection of the correlation, if it existed.

The behavior we observed led us to make a clear distinction between a stage of information search and a stage of information integration and evaluation. Irrespective of their prior training, problem content, or even correlation strength, participants exhibited similar information search behavior: They sampled about 22 items before stopping to pass judgment, evaluate the correlation, and possibly participate in a subsequent prediction task. That sample size is similar to that observed in similar tasks involving self-terminated sampling (Hertwig et al., 2004; Kareev & Fiedler, 2006; Weber et al., 2004) and adds to the growing literature that indicates that people feel samples consisting of 8–11 items provide acceptably accurate estimates of measures of central tendency, with a sample about double that size enabling a decision concerning the difference between two such measures (i.e., an estimate of the strength of a correlation, at least that between two binary variables, which calls for estimating the difference between two proportions). Importantly, although the data sets consisted of two groups of unequal size, the samples consisted of almost identical number of items from each group. This finding replicates and expands on the findings of Kareev and Fiedler and indicates that people adopt sampling techniques that place greater weight on the early detection of a correlation than on the accurate estimate of its value. The finding that no aspect of sampling behavior was related to any of the independent variables may be taken as an indicator that in searching for information, people employ general routines for data acquisition.

Whereas the analysis of sampling behavior can be taken to indicate that information search is impervious to situational factors, the analysis of the assessment of correlation strength and its predictive value shows these facets to be highly sensitive to all the factors whose effect was studied. First, both the assessment of the strength of the correlation and its valuation corresponded closely to the objective correlation and its predictive power. Although such a relationship was to be expected, given the wide range of correlations employed, it still deserves to be pointed out. This overall correspondence notwithstanding, we also observed a strong, persistent interaction between problem content and field of study. Although the contents of the cover stories resembled only loosely the type of research typically encountered by students of the natural and social sciences, changes in contents made students of different backgrounds assess the strength and the value of the correlation as higher when it involved variables more familiar to them. That effect, of "our" correlation being perceived as stronger and more valuable then "theirs" is of particular interest. It demonstrates clearly the strong effect that beliefs, even if vague and general, have on how information is evaluated. It goes to show that data alone may not be sufficient to fully overcome prior notions.

The need for further research to assess the strength of these beliefs and the way in which they affect the assessment of correlations is an obvious conclusion to be drawn from the present findings.

Author Note Shira Soffer, School of Education, The Hebrew University of Jerusalem. Yaakov Kareev, School of Education and The Center for the Study of Rationality, The Hebrew University of Jerusalem. Work was supported by Grants 800/04 and 539/07 from the Israel Science Foundation to Yaakov Kareev.

References

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? *Psychological Bulletin, 114*, 435–448.

Allan, L. G., Hannah, S. D., Crump, M. J. C., & Siegel, S. (2008). The psychophysics of contingency assessment. *Journal of Experimental Psychology: General, 137*, 226–243.

Allan, L. G., Siegel, S., & Tangen, J. M. (2005). A signal detection analysis of contingency data. *Learning & Behavior, 33*, 250–263.

Alloy, L., & Tabachnik, N. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91*, 112–149.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Science, 9*, 226–232.

Chase, W. G., & Simon, H. A. (1979). Perception in chess. In H. A. Simon (Ed.), *Models of thought* (pp. 386–403). New Haven, CT: Yale University Press.

Chater, N., & Oaksford, M. (2006). Mental mechanisms: Speculations on human causal learning and reasoning. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 210–236). New York: Cambridge University Press.

Crocker, J. (1981). Judgment of covariation by social perceivers. *Psychological Bulletin, 90*, 272–292.

Crump, M. J. C., Hannah, S. D., Allan, L. G., & Hord, L. K. (2007). Contingency judgments on the fly. *The Quarterly Journal of Experimental Psychology, 60*, 753–761.

De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theories on human contingency learning. *The Quarterly Journal of Experimental Psychology, 55B*, 289–310.

Fiedler, K., Brinkmann, B., Betsch, T., & Wild, B. (2000). A sampling approach to biases in conditional probability judgments: Beyond base rate neglect and statistical format. *Journal of Experimental Psychology: General, 129*, 399–418.

Fugelsang, J. A., & Thompson, V. A. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology, 54*, 15–32.

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review, 116*, 661–716.

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science, 15*, 534–539.

Jennings, D. L., Amabile, T. M., & Ross, L. (1982). Informal covariation assessment: Data-based versus theory-based judgments. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 211–230). Cambridge: Cambridge University Press.

Johnson, T. R., Budescu, D. V., & Wallsten, T. S. (2001). Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making, 14*, 123–140.

Kareev, Y. (1995). Positive bias in the perception of correlation. *Psychological Review, 102*, 490–502.

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

Kareev, Y., & Fiedler, K. (2006). Nonproportional sampling and the amplification of correlations. *Psychological Science, 17*, 715–720.

Kareev, Y., Fiedler, K., & Avrahami, J. (2009). Base rates, contingencies, and prediction behavior. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*, 371–380.

Klahr, D., & Dunbar, K. (1988). Dual-space search during scientific reasoning. *Cognitive Science, 12*, 1–48.

Klahr, D., Fay, A. L., & Dunbar, K. (1993). Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology, 25*, 111–146.

Koslowski, B., Okagaki, L., Lorenz, C., & Umbach, D. (1989). When covariation is not enough: The role of causal mechanism, sampling method, and sample size in causal reasoning. *Child Development, 60*, 1316–1327.

Kuhn, D., & Dean, D. (2005). Is developing scientific thinking all about learning to control variables? *Psychological Science, 16*, 866–870.

Kuhn, D., & Pease, M. (2008). What needs to develop in the development of inquiry skills? *Cognition and Instruction, 26*, 512–559.

Kuhn, D., Schauble, L., & Garciamila, M. (1992). Cross-domain development of scientific reasoning. *Cognition and Instruction, 9*, 285–327.

Lau, R. R., Smith, R. A., & Fiske, S. T. (1991). Political beliefs, policy interpretations and political persuasion. *Journal of Politics, 53*, 646–675.

McKenzie, C. R. M., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology, 54*, 33–61.

Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review, 101*, 608–631.

Oaksford, M., & Chater, N. (2003). Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review, 10*, 289–318.

Perales, J. C., Catena, A., Shanks, D. R., & González, J. A. (2005). Dissociation between judgments and outcome-expectancy measures in covariation learning: A signal detection theory approach. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 31*, 1105–1120.

Pryor, J. B., & Merluzzi, T. V. (1985). The role of expertise in processing social interaction scripts. *Journal of Experimental Social Psychology, 21*, 362–379.

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*, 31–57.

Schauble, L. (1996). The development of scientific reasoning in knowledge-rich contexts. *Developmental Psychology, 32*, 102–119.

Schauble, L., Klopfer, L. E., & Raghavan, K. (1991). Students' transition from an engineering model to a science model of experimentation. *Journal of Research in Science Teaching, 28*, 859–882.

Shaklee, H., & Fischhoff, B. (1982). Strategies of information search in causal analysis. *Memory & Cognition, 10*, 520–530.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of correlations between events. *Memory & Cognition, 8*, 459–467.

Shanks, D. R. (2004). Judging covariation and causation. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 220–239). Malden, MA: Blackwell.

Simon, H. A., & Gobet, F. (2000). Expertise effects in memory recall: Comment on Vicente & Wang (1998). *Psychological Review, 107*, 593–600.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*, 309–318.

Vadillo, M. A., & Matute, H. (2007). Predictions and causal estimations are not supported by the same associative structure. *The Quarterly Journal of Experimental Psychology, 60*, 433–447.

Vadillo, M. A., Miller, R. R., & Matute, H. (2005). Causal and predictive-value judgments, but not predictions, are based on cue–outcome contingency. *Learning & Behavior, 33*, 172–183.

Vicente, K. J. (1992). Memory recall in a process control system: A measure of expertise and display effectiveness. *Memory & Cognition, 20*, 356–373.

Waldmann, M. R. (2001). Predictive versus diagnostic causal learning: Evidence from an over-shadowing paradigm. *Psychonomic Bulletin & Review, 8*, 600–608.

Wasserman, E. A., Elek, S. M., Chatlosh, D. L., & Baker, A. G. (1993). Rating causal relations: Role of probability in judgments of response–outcome contingency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 174–188.

Weber, E. U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111*, 430–445.