האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

EXPECTED PREDICTION ACCURACY AND THE USEFULNESS OF CONTINGENCIES

by

YAAKOV KAREEV, KLAUS FIEDLER and JUDITH AVRAHAMI

Discussion Paper # 455 July 2007

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Expected Prediction Accuracy and the Usefulness of Contingencies*

Yaakov Kareev+, Klaus Fiedler±, & Judith Avrahami+

July 1, 2007

Abstract

Regularities in the environment are used to decide what course of action to take and how to prepare for future events. Here we focus on the utilization of regularities for prediction and argue that the commonly considered measure of regularity - the strength of the contingency between antecedent and outcome events - does not fully capture the goodness of a regularity for predictions. We propose, instead, a new measure - the level of expected prediction accuracy (ExpPA) - which takes into account the fact that, at times, maximal prediction accuracy can be achieved by always predicting the same, most prevalent outcome, and in others, by predicting one outcome for one antecedent and another for the other. Two experiments, testing the ExpPA measure in explaining participants' behavior, found that participants are sensitive to the twin facets of ExpPA and that prediction behavior is best explained by this new measure.

* The research reported in this paper was partly supported by grant 800/04 from the Israel Science Foundation to Yaakov Kareev and by various DFG grants to Klaus Fiedler.

+ School of Education and The Center for the Study of Rationality, The Hebrew University of Jerusalem.

± Department of Psychology, University of Heidelberg.

Introduction

People, like other organisms are constantly on the lookout for regularities, to decide on their actions and prepare for the upcoming state of affairs. It is therefore of great interest to understand how and how well people utilize regularities in their environment. Here we focus on the utilization of regularities and, more specifically, on their utilization for prediction.

The relative frequency of outcome events is the simplest form of regularity that can be used for prediction: When one outcome event is more likely than others, preparing for that event would raise the probability of being well-prepared for the future. A more complex regularity inheres in the relation between antecedent and outcome events: If the likelihood of one outcome event is higher when one antecedent event prevails than when another, we say that the outcome events are contingent on the antecedent events. Often, such a contingency calls for different predictions - and different actions - given the different antecedent states, but this is not always the case.

To illustrate why this is not always the case, consider the following scenario. Springtime and, planning to go shopping in the Old City of Jerusalem, you try to predict the likelihood of finding a parking place by the City Walls. Based on this prediction you will decide whether to take your car or a taxi. The chances for finding a parking place are slim this time of the year; you'd better take a taxi. You also know the day of the week and that the City is more crowded on Fridays, the day many Muslims come there to pray. If you had a choice on what day to go you would, no doubt, take this correlation into account (and, depending on your preference, choose one day or another). Still, if all you do is predict parking conditions, the information related to the day of the week is

irrelevant: You would predict them to be bad irrespective of the day of the week. The situation could, of course, be different and the city crowded only on certain days of the week. If that were the case you would predict different parking conditions for different days and take your car on some days and a taxi on others.

The above scenario demonstrates that even when certain states are contingent on others, greater predictive accuracy may be achieved, at times, by ignoring the contingency altogether and relying on the overall likelihood of one state or another. Stated more generally, when an antecedent event is known one would be best prepared for the future by predicting the most likely outcome given that event; it should be noted, however, that the most likely outcome may be the same for all antecedent events even if the former is contingent on the latter. In other words, the highest likelihood of correct predictions may be attained, at times, from always predicting a single outcome and in others, from predicting different outcomes for different antecedent events (see Fiedler & Kareev, 2006; Kareev, 1995, 2005). Obviously, since contingency-based differential predictions do not always maximize predictive accuracy, the (statistical) strength of the contingency does not fully capture the value of the regularity in the service of prediction.

What is needed to capture this regularity is a measure that would reflect the overall goodness of the regularity in the environment for prediction. Such a measure could then be used to characterize environments and could serve as a yardstick against which prediction behavior, as well as the assessment of contingencies, is compared. We propose that the proportion of accurate predictions expected when the regularity is detected and used be that measure, which we call Expected Prediction Accuracy (ExpPA).

For a more formal definition of ExpPA consider Table 1. This table presents a 2x2 combination of antecedent and outcome events. The cells denoted by *a*, *b*, *c*, and *d* represent the frequency of each of the four combinations assuming, without loss of generality, that $a \geq b$. If event 3 is more likely given both event 1 and event 2, namely, $a > b$ & $c > d$, one would do best to always predict event 3. This is true even if the outcome events are contingent on the antecedent ones such that the likelihood of event 3 given event 1 is different from its likelihood given event 2. Only when one of the outcome events is more likely given one antecedent and another more likely given the other, e.g., if $a > b$ & $d > c$, is the contingency between the events useful for prediction. In this case, differential predictions – predicting event 3 given event 1 and event 4 given event 2 – would result in greater prediction accuracy.

		Outcomes	
		Event 3	Event 4
Antecedents	Event 1	A	B
	Event 2	C	d

Table 1. *The 2x2 Relation between Antecedents and Outcomes*

ExpPA is defined as the maximal level of accuracy that could be attained either by the skew in the marginal distribution of outcome events - if there is one - or by the proportion of cases in the more common diagonal - if there is one. In terms of Table 1 we define the expected proportion of correct predictions as:

$$ExpPA = \max\left(\frac{a+c}{a+b+c+d}, \frac{a+d}{a+b+c+d}\right) \quad (1)$$

with the first term corresponding to the proportion of cases in the prevalent outcome event and the second to the proportion of cases in the more prevalent diagonal. That is, the source of ExpPA can be either the skew in outcomes likelihood, calling for always predicting one outcome value, or the contingency between antecedent and outcome events, calling for predicting one outcome for one antecedent and another for the other. We maintain that this measure captures the goodness or the strength of the regularity in the environment with respect to prediction, and ask how sensitive people are to that measure.

Note the difference between this measure and the measure of the strength of the contingency, Δp, defined as:

$$\Delta p = \frac{a}{a+b} - \frac{c}{c+d} = \frac{ad-bc}{(a+b)(c+d)} \qquad (2)$$

As to the usefulness of detecting and using a contingency in the environment for prediction, equation 1 shows that it can be expressed by the relative frequency of cell *d* versus cell *c*. The comparison between equations 1 and 2 shows that some contingencies, although different from zero may be useless, as far as correct predictions are concerned.

Note that we do not claim that when the proportions of cases in the prevalent outcome is greater than the proportion of cases in the prevalent diagonal the contingency is always useless: When choice between antecedent states is possible and there is a preference for one of the outcome events over others, even a contingency that is useless for prediction is worth using. To illustrate, whether smoking or not, the majority of people are healthy; hence predicting that a certain person is healthy is most likely to be correct irrespective of whether that person smokes or not. Still, when one chooses whether to smoke or not one would take into account the difference, if there is any,

between the likelihood of staying healthy when smoking and that of staying healthy when not. [1]

Since the contingency between variables is important for choice between actions - in view of their consequences - and often for prediction as well, the perception and assessment of contingencies have been studied extensively. Most of the earlier studies have focused on the perception of contingencies through the accuracy of their assessment (for reviews see Allan, 1993; Alloy & Tabachnik, 1984; Beyth-Marom, 1982; Shanks, 1995). These studies found a number of factors affecting the judged strength of a contingency. For example, mode of presentation - whether trial by trial or summarized (e.g., Ward & Jenkins, 1965), type of variables - whether symmetric or asymmetric (e.g., Allan & Jenkins, 1980), way of posing the question - whether the focus is on one, two, or all joint frequencies of the events in question (Crocker, 1982), and the marginal distribution of the variables involved in the contingency (e.g., Dickinson, Shanks, & Evenden, 1984), have all been found to affect the assessment of the strength of a contingency. Thus, whereas research on the assessment of relations between states found that people *are* sensitive to such regularities, it also found that factors that are irrelevant to the statistical strength of the contingency have an effect on its assessment. At the same time, the way regularities are being utilized for prediction has been hitherto almost completely overlooked (for exceptions, with early discussions of the issue, see Fiedler & Kareev, 2006; Kareev, 2005).

[1] It should be mentioned that the proposal of different measures of regularity for prediction and for choice does not apply to the case of continuous variables, where the strength of a contingency is always a good measure of the regularity for the purpose of prediction and choice alike.

In the studies reported here we focus on the role played by ExpPA in determining prediction and ask whether people are sensitive to the strength of ExpPA and to its source. Two experiments were conducted in which participants' reward was a function of their prediction accuracy. In the first, prediction behavior, preferences, and the assessment of the strength of contingency was compared for two data sets that differed only in the skew in the distribution of outcome events. In the second, both the strength of contingency and the skew of outcome events were manipulated - orthogonally - producing four versions of a computer game. Participants' tendency to invest in predictor information was compared for the four versions.

Experiment 1

In this experiment participants observed items from two data sets, each with two values of a predictor and two values of a criterion. For 96 rounds participants observed a predictor value - at random from either data set - and predicted the value of the criterion. They were rewarded for correct predictions. Participants then estimated the strength of contingency in each data set, indicated which of the two they would choose for a subsequent prediction task, and finally made ten predictions for items from the data set of their choice.

The purpose of this experiment was to explore the effects of ExpPA on the use of the regularity in prediction, on the assessment of the strength of the contingency between predictor and criterion value, and on its attraction for choosing between data sets. Therefore, the two data sets differed in their ExpPA and in its source, hence calling for different prediction strategies: In one, maximization of predictive accuracy required

differential prediction whereas for the other, it required always predicting the more common value of the criterion. The two sets had the same Δp but differed in the skew of the marginal distribution of the criterion; hence, participants' sensitivity to the skew, to the contingency or to both, could be revealed.

Method

Materials. The makeup of the two data sets employed in the experiment is depicted in Table 2. The Δp was equal to .42 in both data sets, the marginal distribution of the predictor was .5/.5 in both, but the marginal distribution of the criterion was different: .54/.46 in the data set depicted on the left of Table 2 and .75/.25 in the data set on the right, rendering the skew in the data set on the right higher. The ExpPA of the low-skew data set (the one on the left) - derived from the proportion of cases in the more common diagonal - was 34/48=.70, whereas the ExpPA of the high-skew data set (the one on the right) - derived from the proportion of cases in the more common column - was 36/48=.75.

	Color 1	Color 2	
Box 1	18	6	24
Box 2	8	16	24
	26	22	48

	Color 3	Color 4	
Box 3	23	1	24
Box 4	13	11	24
	36	12	48

Table 2. *The Frequencies of the Two Disc-Colors in the Two Box-Types Separately for the Two Data Sets of Experiment 1*

The materials comprising a data set were 48 tiny boxes, each containing a colored disc. There were four box types and four disc colors such that every set was comprised of its own two distinct types of boxes and two disc colors. Boxes of the same type and discs of the same color were indistinguishable. The assignment of pairwise combinations of box types and disc colors to the two data sets in the pair was counter-balanced between participants. All 96 tiny boxes were stored in a large carton box.

Procedure. Participants were tested individually in a quiet room. Upon entering they were told that they would take part in a study consisting of three phases. For the first phase they were presented with the carton containing the 96 boxes, and told that they would draw, without looking, one box at a time and predict the color of the disc inside it. Before starting, the boxes in the carton were well shuffled in front of them. Once a prediction had been made the box was opened, and if the color of the disc matched the predicted color the participant was awarded 1/2 New Israeli Shekel (about 11 cents at the time). The box (with the disc inside it) was then closed again and placed in another carton. Each participant encountered, during the learning phase, the 96 items comprising the two data sets in a random, intertwined, order. Throughout the experiment the participants had, in front of them, an example of the two pairs of box types and the two disc colors that could be found in them. This ensured that participants distinguished between the two data sets, predicting one of two colors for one set and one of two other colors for the second. On each trial the experimenter recorded the type of box, the participant's prediction, and the actual color of the disc inside the box. Correct predictions were rewarded immediately.

Following the learning phase, the participant performed three tasks:

Direct estimation of the strength of the contingencies: Here the participant was asked to assess, for each data set, the strength of the relation between type of box and color of disc and indicate it by placing a mark on a 150 mm long line that was labeled "no relationship" at its left end, and "perfect relationship" at its right end. Participants were explicitly informed that "no relationship" meant that both types of boxes had the same percentage of each disc color in them, whereas "perfect correlation" meant that each type of box had but one (and different) color of disc in it. The order in which the two data sets were asked about was counterbalanced.

Indirect estimation of the strength of the contingencies: For this task participants were asked to estimate, for each data set, the percentage of discs of one of the colors within the two box types of that data set. The set asked about first, the color asked about, and the order of the types of box, were all counterbalanced across participants.

Choosing a data set: Here the participant was told that the next (third) phase of the study would involve a 10-trials repetition of the first phase (i.e., draw, then predict for a reward), but this time only from one of the two data sets. Participants chose one of the sets knowing, at the time of choice, that the reward for a correct prediction in that phase would be twice as large as that provided before, and that drawing would be conducted from among the same 48 boxes that comprised the data set of their choice.

To counter possible order effects on assessment and choice, each task was performed first by a third of the participants; to keep the counterbalancing in check (see the Design section below for the number of counterbalanced variables) only three of the

six possible orders were used: Choice, Direct, Indirect; Direct, Indirect, Choice; Indirect, Direct, Choice.

Having completed the three tasks of the second phase the participant sampled ten more boxes, this time from the chosen data set. The process was like that performed during learning, namely, random-drawing without replacement, but correct predictions were rewarded by one NIS. The whole experiment was self paced, and typically lasted 25-30 minutes.

Participants. Ninety-six students (80 from the Mt. Scopus campus of the Hebrew University and 16 from Ben Gurion University in Beer Sheba) participated in the experiment for the monetary reward determined by performance, as explained above. The number of participants was dictated by the counterbalancing involved, see below, including the additional stipulation of having an equal number of males and females in each cell of the design.

Design. The factor that was manipulated in this experiment was the skew in the marginal distribution of the criterion. This resulted in two differences between the two data sets:

1. A difference in the level of ExpPA;
2. A difference in the source of ExpPA (the skew in one and the contingency in the other data set) and hence in the usefulness of the contingency.

The factors that were counterbalanced between participants were:

1. Assignment of box type and disc color to data set (2);
2. Order of judgment tasks (3);

3. Characteristics of the Indirect Estimation task ([8]: set asked about first [2] x color asked about [2] x order of box types asked about [2]);

4. Gender (2).

Results and Discussion

We report the results in three sections, the first having to do with the assessment of the strength of the contingency the second with choosing a data set for further predictions and the third with the prediction strategy used.

Assessment. Recall that we used two different ways for eliciting participants' assessment of the strength of the contingencies.

For the direct estimate of the contingency we measured the distance of the mark a participant placed on the line (see above) from the left end labeled 'no-relationship'. The mean for the low-skew data set was 69.4 mm and that for the high-skew data set was 86.1 mm. The data set with the higher skew was thus judged to have stronger contingency than that with the low skew (F[1,95]=11.73, MSE=1138, p=.001).

For the indirect measure of perceived contingency we subtracted the probabilities that participants provided for a criterion value given one predictor value from the probability provided for the same criterion value given the other predictor value. Note that the actual value was positive but estimates could result in a negative value. The mean values were .21 for the low-skew data set (the one on the left of Table 2) and .28 for the high-skew data set indicating, again, that the contingency of the data set with high skew was perceived as stronger than that with low (F[1,95]=5.84, MSE=.041, p=.018).

Choice. This measure indicates which data set a participant preferred for phase 3 - the subsequent draw-and-predict session of another ten trials. A choice was scored -1 if the participant chose the low-skew data set and +1 if the high-skew data set was chosen. Thus, the deviation of the overall mean from zero indicates which side was preferred. The mean value of +.302 reflects the fact that most participants (fully .651 of them) preferred to make further predictions with the high-skew data set - a highly significant difference ($t[95]=4.20$, $p<.001$).

Prediction Behavior. We now turn to an analysis of the actual predictions made by the participants. The question of interest here is whether participants' predictions were contingent on predictor values, or simply reflected the more frequent value of the criterion. To that end a new measure, 'contingency-use', was derived, for every participant in each data set. We tabulated the frequency with which every participant predicted either color given either box type, and identified the diagonal ($a+d$ or $b+c$) and the column ($a+c$ or $b+d$) that had a higher frequency of predictions. Participants were then characterized as contingency-users (and assigned a score of +1) if the frequency of predictions in the common diagonal was higher than the frequency in the common margin. Participants were characterized as margin-users (and assigned a score of -1) if their predictions fell more often in the more common column than in the more common diagonal. Participants received a score of 0 if the more common diagonal and the more common column were equal in frequency.

Pre-Choice Predictions: There was no difference in the utilization of the contingency for the two data sets during the first, learning, phase. The mean score for the low-skew data set was .02 and that for the high-skew data set was -.04 ($F<1$).

Nevertheless, when inspecting only the last ten rounds of the learning phase for each data set a difference in contingency-use does emerge. The mean score for the low-skew data set was .073 and that for the high-skew data set was -.135. This marginally significant difference (F=[1,95]=3.81, MSE=.546, p=.054) does indicate that towards the end of the learning phase participants started to pick up the difference in the source of the ExpPA in the two data sets and used that source for their predictions.

Post-Choice Predictions: The same measure of 'contingency-use' was used to characterize predictions in the last, post-choice, phase of the study. Here every participant had a score only for the data set that had been chosen. The mean score for the low-skew data set was .552 whereas that for the high-skew data set was -.134 (F[1,94]=11.06, MSE=.86, p=.001). These values indicate that participants who chose the low-skew data set must have done so because they judged its contingency to be useful while those who chose the high-skew data set must have done so because they noticed either the utility in the skew or the utility in the contingency with the slight advantage of the margin over the diagonal translating into a slight preference for using the former over the latter.

Prevalence of Maximizing Behavior: In discussing prediction behavior we have characterized participants as "contingency users" or "margin users", based on their more prevalent pattern of predictions. Obviously, the validity of this characterization depends on the extremity of that pattern: The less extreme the pattern, the less justification we would have for using it to label people's behavior. Most of the literature involving choices in probabilistic environments (for reviews see, for example, Estes, 1976; Vulkan, 2000) failed to observe maximizing behavior. Here, in contrast, an analysis of the number of cases (out of the ten in the post-choice prediction task) in which participants' choices

fell within their characteristic pattern revealed the average number to be 9.15: Fully 46 of the 96 participants (.48) had all their 10 predictions either fall in one diagonal or in one column. Another 26 (.27) had 9 such cases, with an additional 18 (.19) having 8 such items. Obviously, and in contrast to what could be expected on the basis of earlier findings, our participants exhibited highly consistent, maximizing behavior.

Discussion

Experiment 1 revealed an overwhelming preference for the high-skew data set and a perception of the contingency in the high-skew data set as stronger. This preference cannot be explained by the strength of the contingency in the data set - which was equal in both. The effect can be explained by the difference in ExpPA between the two data sets.

The way participants utilized the regularities in the data sets indicates that they were sensitive not only to the difference in ExpPA but also to the difference in its source: The tendency to use the contingency for differential predictions was greater with the data set in which the contingency was useful but lower with the data set in which slightly greater accuracy could be achieved by relying on the skew in the margins. In doing so participants optimized their prediction behavior.

Note that the skew in the distribution of criterion values can determine ExpPA in two ways: First it can determine its source, namely, whether the margin or the diagonal. Second, even when the source is the diagonal, the proportion of cases in the diagonal determine ExpPA's level such that two contingencies of equal strength can have a different proportion of cases in the common diagonal hence differ in their ExpPA.

Experiment 1 tested participants' sensitivity to the level of ExpPA in data sets that differed in ExpPA's source. We now turned to ask whether participants would be sensitive to differences in ExpPA even when it is always determined by the more common diagonal. Would their tendency to use information about the predictor value for prediction correspond to the strength of contingency between predictor and outcome or to ExpPA, namely, to the proportion of cases in the prevalent diagonal? To that end, four versions of the task were employed, such that each shared the strength of its contingency with one other and the skew with another. In none of the versions was the proportion of cases in the more common margin higher than the proportion of cases in the more common diagonal.

Experiment 2

In this experiment participants played a game in which a forest scene was gradually revealed while they had to predict what type of animal would be found hiding there. There were two different scenes and two different animals - a hamster and a frog. Participants were rewarded for correct predictions - the higher the faster they responded - and fined for incorrect predictions - again higher the faster they responded. Unlike Experiment 1, in which participants could not avoid noticing the value of the predictor, in Experiment 2 participants could forgo the predictor. To optimize their rewards participants had to assess their chances of correct predictions with and without the predictor value and decide whether or not to wait for that value to be revealed.

Method

Design. Table 3 presents the likelihood of each combination of predictor and criterion values for the four versions of the game:

A	Animal1	Animal2		
Scenery1	.42	.28	.70	
Scenery2	. 08	. 22	.30	Δ_p=.3
	.50	.50		

Expected accuracy:

From contingency .64

From skew .50

B	Animal1	Animal2		
Scenery1	.55	.15	.70	
Scenery2	.15	.15	.30	Δ_p=.3
	.70	.30		

Expected accuracy:

From contingency .70

From skew .70

C	Animal1	Animal2		
Scenery1	. 46	.24	.70	
Scenery2	.04	.26	.30	Δ_p=.6
	.50	.50		

Expected accuracy:

From contingency .72

From skew .50

D	Animal1	Animal2		
Scenery1	.60	.10	.70	
Scenery2	.10	.20	.30	Δ_p=.6
	.70	.30		

Expected accuracy

From contingency .80

From skew .70

Table 3. *The Frequencies of the Two Animals in the Two Sceneries Separately for the Four Versions of the Game of Experiment 2*

The pairs of versions A - B and C - D have the same Δp (.3 and .6, respectively) but a different marginal distribution of criterion values; the pairs of versions A - C and B - D share the marginal distribution but have different Δp. The resulting four versions all differ in ExpPA even though there are only two levels of Δp. ExpPA, here the proportion of cases in the more common diagonal, was .64, .70, .72, and .80 for versions A, B, C, and D, respectively.

Importantly, the versions differ in the usefulness of the contingency: The difference between the probability of making correct predictions when making contingency-based differential predictions (namely, using only the more common diagonal) compared to the expected prediction accuracy when using only the more common margin. In version A the expected accuracy for predictions based on the diagonal is .64 and that for predictions based on the margin is .50, with the contribution of the contingency equal to .14. In version B the expected accuracy for predictions based on the diagonal is .70 and that for predictions based on the margin is also .70, with the contribution of the contingency equal to zero. In version C the expected accuracy for predictions based on the diagonal is .72 and that for predictions based on the margin is .50, with the contribution of the contingency equal to .22. In version D the expected accuracy for predictions based on the diagonal is .80 and that for predictions based on the margin is .70, with the contribution of the contingency equal to .10.

Materials and Procedure. The whole session was computer controlled and administered individually in a quiet room. The instructions, presented as text on the computer monitor, depicted the situation as one involving a bird of prey preparing to swoop down and capture either a hamster or a frog. Participants' task was to prepare, on

each trial, for one type of animal. Points gained (or lost; depending on whether they did or did not prepare for the right animal on that trial) depended on the time elapsed since the onset of the trial, and was indicated by a countdown timer. Participants were also told that on any given trial they would be operating in one of two sceneries, but that the picture of that scenery would be hidden behind a curtain at the onset of the trial, becoming visible only as the curtain came down. Finally, it was explicitly stated that waiting for the scenery to be exposed could, but did not necessarily, improve the accuracy of their predictions, and that they were allowed, but not required to wait for it to be exposed. It was made clear that waiting for the picture to be exposed would result in a lower value on the countdown timer. Immediate feedback was provided after each response. The prediction session lasted for 15 minutes. Time left until the end of the session, was indicated by a bar and a numeric display.

The countdown timer was presented at the top left corner of the monitor, with its initial value set at 20. At the center of the monitor there was a picture of one of two sceneries, covered by a curtain. That picture could serve as a predictor if the participant waited long enough for the curtain to come down sufficiently for the scenery to be uniquely identified. The curtain started to come down 2.5 s after trial onset, making an increasing part of the screen visible, and took about 3.5 seconds to come fully down. After 2.7 s the curtain was already low enough for the revealed scenery to be uniquely identified. The display was in view until the participant indicated his or her prediction (by clicking either on the picture of a hamster or on that of a frog, located at the bottom of the monitor), or until the countdown timer reached 0 (which took about 4 seconds to happen). In either case, the participant was informed of the criterion value, that is whether the

animal was a hamster of a frog, on every trial. A counter displaying the total number of points accumulated until then was updated by adding or subtracting the number of points left on the countdown timer at the time of response. The next trial started automatically after 4 seconds. A session lasted 15 minutes; thus the number of trials in a session varied, depending on the speed with which the participant responded.

At the end of the session the participant was debriefed, and paid in correspondence with the points that had been accumulated.

Participants. A total of 74 students at the University of Heidelberg participated in the experiment for pay. They were randomly assigned to one of the four conditions. One participant had to be removed because he hardly made any predictions.

Results and Discussion

The first measure for evaluating participants' behavior in this task is the percent of correct predictions attained in the four versions. Mean correct predictions were .495, .605, .609, and .686 for versions A, B, C, and D, respectively. As such, they better correspond to the level of ExpPA of these versions - which was .64, .70, .72, and .80, than to the strength of the contingencies which were .3, .3, .6, and .6 (for versions A, B, C, and D, respectively). The correlation - across participants - between correct predictions and ExpPA is .601 whereas that between correct predictions and the strength of the contingency is .430. Although both correlations are pretty strong, the difference between them is significant ($t[70]=2.11$, $p=.038$), indicating, once again, that the strength of a contingency does not capture people's prediction behavior in that environment, as well as ExpPA.

An analysis of variance on participants' accuracy (i.e., the proportion of correct predictions) with the strength of the contingency and the skew as between-participants variables shows that both effects were significant: $F(1,69)=22.00$, $MSE=.008$, $p<.001$, for the strength of the contingency and $F(1,69)=20.57$, $MSE=.008$, $p<.001$, for the skew in the margins. The two factors did not significantly interact ($F<1$).

Recall that a major issue of Experiment 2 is the utilization of the contingency, namely, the extent to which participants deemed the costly predictor worth waiting for. This could be tested since the four versions of the game differed in the usefulness of their contingencies over the likelihood of making correct predictions without it. A second measure of participants' behavior is therefore the degree to which the contingency was used for prediction. To that end, every prediction a participant made was classified as either based on the predictor - the scene - or not. This was determined by the speed of predictions: predictions that were made in less than 5.2 s from trial onset (i.e., in less than 2.7 s from the time the curtain started to come down) - the time required for the scenery to be uniquely identifiable - were classified as predictions made without the value of the predictor whereas predictions that took longer to make were classified as predictions made with the predictor value available. The proportion of trials on which participants waited long enough for the predictor to be revealed was .393, .225, .540, and .395 for versions A, B, C, and D, respectively. These values indicate high sensitivity to the strength of the contingency - more waiting to see the predictor value in C than in A and more waiting in D than in B. They also indicate high sensitivity to the skew in the distribution of the criterion - less waiting in B than in A and less waiting in D than in C. Both main effects were significant: $F(1,69)=5.05$, $MSE=.09$. $p=.028$, for the strength of

the contingency and $F(1,69)=4.92$, $MSE=.09$, $p=.030$, for the skew in the margins. The two factors did not significantly interact ($F<1$). Importantly, the strength of the contingency alone cannot fully account for the pattern of results.

Another way to appreciate the pattern of waiting for the predictor in the four versions is to compare it to the pattern of the contribution, to expected accuracy, of the contingency over the marginal distribution (see above). The values of .393, .225, .540, and .395 indeed correspond to .14, .00, .22, and .10, for versions A, B, C, and D, respectively.

These results show that participants were highly sensitive to regularities in the environment in which they were operating, taking into account both skew and contingency to optimize their performance. In other words, the measure of ExpPA captures their prediction behavior very well.

General Discussion

Regularities in the environment are detected for a purpose: For choosing between actions in view of their outcomes and for predicting outcomes when their antecedents are given. Still, the focus of the majority of studies dealing with the detection of regularities is the way they are judged. Here we studied the detection of regularities from a different angle, that of their utilization for prediction. We argued that, viewed from this angle, a new measure for appreciating regularities is required. The measure proposed - the expected prediction accuracy (ExpPA) - denotes the likelihood of correctly predicting future events when the regularity in the environment is maximally utilized. When the probabilities of all outcomes are equal and unrelated to the antecedents, ExpPA equals

zero. But if the distribution of outcomes is unequal for any of the antecedents then predicting the more likely outcome given that antecedent can raise the likelihood of correct predictions over chance. Importantly, even when the distribution of outcomes is unequal given their antecedents the inequality can have two possible relations to the antecedents: One outcome could be more likely given one antecedent and another more likely given another (and then we would say that there is a useful contingency between antecedents and outcomes) or the same outcome could be more likely for all antecedents. In the first case differential predictions would raise the likelihood of correct predictions whereas in the second, the highest prediction accuracy would be achieved by always predicting the more likely outcome, whether or not the outcomes are contingent on the antecedents. The two experiments reported here demonstrate participants' sensitivity to these aspects of the usefulness of regularity. The experiments thus show that participants' behavior in utilizing the regularity for prediction and in assessing its strength is more closely related to ExpPA than to conventional measures of the strength of a contingency.

In the first experiment participants learned, while predicting for a reward, the characteristics of two data sets. The statistical strength of the contingency was the same in the two sets but their ExpPA was different. Moreover, for one data set maximal prediction accuracy could be achieved by using the contingency to make differential predictions, whereas for the other it could be achieved by unconditional prediction of the more common outcome. Participants' behavior indicated that they were sensitive to the level of ExpPA and to its source: The data set with the higher ExpPA was overwhelmingly preferred to be used in a second, more highly rewarding, task. Furthermore, during the final stages of learning and more so in the subsequent, post-

choice prediction task, participants' prediction behavior closely corresponded to the source of ExpPA: Predictions were predictor-dependent (i.e., differential) for the first data set and undifferentiated (i.e., overall frequency-dependent) for the second. This sensitivity to the source of ExpPA in prediction was particularly interesting given that participants judged the contingency in the second set as stronger. This dissociation between the objective strength of the contingency, its utilization, and evaluation attests to the complex ways in which ExpPA affects behavior and indicates that the assessment and utilization of regularities represent distinct aspects of behavior.

The second experiment tested a straightforward implication of our reasoning, namely, that when obtaining a predictor value is costly, people would be sensitive to its contribution to prediction accuracy and be more willing to forgo that value the lower its contribution. In this experiment, participants again made predictions for a reward but this time the predictor was not immediately available, requiring time for it to be revealed. Analyses of participants' behavior show that their accuracy was related to ExpPA and that the level by which they utilized the predictor values closely reflected the relative usefulness of the predictors: The lesser its utility, the more likely they were to forgo its value and respond before it was exposed. That tendency further supports our contention that people are sensitive not only to the strength of regularities but can also compare the two sources - skew or contingency - and judge whether using a costly predictor for improving prediction accuracy is worth their while.

To recapitulate, the measure of ExpPA differs from the more common measure of contingency, Δp, in two ways. First, when the distribution of outcome events is skewed to the extent that always predicting the more common value would maximize prediction

accuracy, the new measure, ExpPA, assumes the value of the likelihood of the more common value, namely, the proportion of cases in the more common column of the 2x2 table relating the joint frequencies of antecedent events and outcome events. Second, when one outcome event is more likely given one antecedent event and another outcome more likely given another antecedent, ExpPA assumes the likelihood of correctly predicting the outcome when making differential predictions - one outcome for one antecedent and another for the other. In both cases it is the likelihood of making correct predictions when utilizing the regularity in the environment to the maximum.

The first aspect, focusing on the skew in the distribution of different outcomes, may bring to mind another effect of skew that has been studied extensively, the outcome density effect (e.g., Cheng, 1997; Dickinson, et al. 1984; Wasserman, Elek, Chatlosh, & Baker, 1993). In a variety of studies participants judged the power of an antecedent - that was either provided to them or which they could bring about - in causing an outcome. Common to all these studies was the finding that when the overall prevalence of the outcome was high participants judged the causal power of the antecedent to be higher than when its prevalence was low, for objectively identical contingency strength. The effect of outcome density was highly pronounced even for objectively unrelated events. The effect received various explanations, some rule based (e.g., Cheng, 1997; White, 2003) and others association based (e.g., applications of the Rescorla-Wagner, 1972, or Pearce's, 1987, associative learning model) but none of them focused on the utility of the density for predictions. We suggest that, the increased likelihood of correctly predicting the occurrence of the outcome - resulting from the increase in outcome density - may have infiltrated participants' judgment of causation (as it did in participants' assessment of

the strength of the contingency in this paper's Experiment 1) and was mistaken for a stronger causal power. As such, the notion of prediction accuracy could provide an additional explanation of the outcome density effect.

The second aspect by which the measure of ExpPA differs from the more common measures of contingency is that even when the contingency is useful, such that the likelihood of making correct predictions is greater when differential predictions are made for different antecedent events, it corresponds to the proportion of cases in the more common diagonal and not to the difference between conditional probabilities. The four versions of the task used in Experiment 2 demonstrate this difference: There are only two values of Δp but four values of the proportion in the diagonal corresponding to ExpPA. In its focus on the diagonals rather than on Δp, this second aspect is reminiscent of White's measure - PCI - that corresponds to the difference in the proportion of cases in the two diagonals (White, 2003). Indeed, the results of Experiment 2 cannot distinguish between PCI and ExpPA since the proportion of cases in the common diagonal is linearly related to the proportional difference between the diagonals. At the same time, the measure of PCI cannot explain the results of Experiment 1, where the two data sets had the same Δp and the same PCI. The two data sets there did differ in ExpPA and participants' behavior corresponded to that difference.

We contend that ExpPA offers insights for understanding how regularities are detected and used. We further recommend that future studies concerned with the detection of regularities focus on the utilization of regularity and be sensitive to the distinction between using regularities for choice or for prediction.

References

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

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response alternative. *Canadian Journal of Psychology, 34,* 1-11.

Alloy, L. B., & 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.

Beyth-Marom, R. (1982). Perception of correlation reexamined. *Memory & Cognition, 10,* 511-519.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104,* 367-405.

Crocker, J. (1982). Biased questions in judgment of covariation studies. *Personality and Social Psychology Bulletin, 8,* 214-220.

Dickinson, A., Shanks, D., & Evenden, J. (1984). Judgement of act-outcome contingency: The role of selective attribution. *Quarterly Journal of Experimental Psychology, 36A,* 29-50.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37-64.

Fiedler, K., & Kareev, Y. (2006). Does decision quality (always) increase with the size of information samples? Some vicissitudes in applying the law of large numbers. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32,* 883-903.

Kareev, Y. (1995). 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 & Olsson (2005) and Anderson, Doherty, Berg, & Friedrich (2005). *Psychological Review, 112,* 280-285.

Pearce, J. (1987). A model for stimulus generalization in Pavlovian conditioning. *Psychological Review, 94,* 61-73.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations on the effectiveness of reinforcement and non-reinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64-99). New York: Appelton-Century-Crofts.

Shanks, D. (1995). *The psychology of associative learning.* Cambridge, England: Cambridge University Press.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14,* 101–118.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. *Canadian Journal of Psychology, 19,* 231-241.

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.

White, P. A. (2003). Making causal judgments from the proportion of confirming instances: The *p*CI rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 710-727.