Psychology in Spain, 2003, Vol. 7. No 1, 3-9

RISK-TAKERS: DO THEY KNOW HOW MUCH OF A RISK THEY ARE TAKING?

Rosa Bersabé Morán*, Rosario Martínez Arias** and Ricardo Tejeiro Salguero*
*University of Málaga. **Complutense University of Madrid

In a real-world setting, two groups of undergraduate subjects showing different levels of risk were compared in several measures of probability judgment accuracy. After taking a test with 20 true-false items, they were asked to estimate the subjective probability that each item was true. Subjects were split into two risk groups. Both achieved the very same grades, but the risky group gave a larger number of answers. The data obtained suggest that risk has an effect on calibration and noisiness of the probability judgments. The more reckless subjects were more poorly calibrated, i.e., the subjective probabilities they estimated deviated further from the actual proportion of true items. Therefore, it seems that those who risk more know less about how much they risk. The discussion focuses on the implications of these findings for gambling behaviour. Calibration and covariance graphs are provided.

En una situación real, dos grupos de sujetos que manifestaban diferente nivel de riesgo se compararon en diferentes medidas de precisión en los juicios de probabilidad. Después de contestar un examen de 20 ítems tipo verdadero-falso, se pidió a los alumnos que estimaran la probabilidad de que cada una de las preguntas fuera verdadera. Se formaron dos grupos según el nivel de riesgo de los sujetos. Ambos consiguieron las mismas notas en el examen, pero el grupo de arriesgados contestó un mayor número de preguntas. Los datos obtenidos sugieren que el riesgo afecta a la calibración y al ruido de los juicios probabilísticos. Los sujetos más arriesgados calibraron peor, es decir, las probabilidades que estimaron se desviaban más de las proporciones reales de ítems verdaderos. Por tanto, parece que los que se arriesgan más, saben menos cuánto se arriesgan. La discusión se centra en las implicaciones de estos resultados sobre la conducta en los juegos de azar. Se acompañan gráficos de calibración y covariación.

Probability judgements are relevant in decision-making. Let us imagine, for example, an individual who was deciding whether or not to take out an insurance policy for his house. In order to choose one or the other alternative, it would be suitable to have information on various matters. On the one hand, he should weigh up the *consequences* (advantages and disadvantages) of subscribing to the policy or not. On the other hand, it is important to know the *probabilities* of theft, flood and other events that might cause material losses in his house. If he were absolutely sure that the probability of these events were nil (perhaps in a fictional world) it would be advisable not to take out the policy. But the majority of decisions are taken with neither exact knowledge nor total ignorance of the probabilities of future events. In general, we are faced with situations of relative uncertainty in which we can assess to some extent the likelihood of certain events, either because we have some idea of their probability distribution or because we have some beliefs about them (Martínez Arias, 1991).

The most commonly used procedure for showing the likelihood of an event is by requesting the subject to judge the probability (in percentages) of that event occurring. In the present study we analyze various measures of the accuracy of probability judgements. Of these, calibration has been easily that which has received most attention in the literature in this field, and refers to the extent to which a subject's probability judgements about an event coincide with the proportion of times it actually occurs.

Several studies have attempted to determine whether there are individual differences with regard to the calibration of probabilities (the following reviews are relevant: Keren, 1991; Lichtenstein, Fischhoff & Phillips, 1982; O'Connor, 1989). Variables that these studies have analyzed in relation to the subject are: sex; knowledge of the matter; experience in different fields; cultural differences; depressive mood; and personality variables (authoritarianism, conservatism, dogmatism and intolerance to ambiguity).

The original Spanish version of this paper has been previously published in *Apuntes de Psicología*, 2002, Vol. 20. No 1, 3-16

.....

Address for correspondence: Rosa Bersabé Morán. Dpto. de Psicobiología y Metodología de las CC. del Comportamiento. Facultad de Psicología. Universidad de Málaga. 29071 Málaga. Spain. E_mail: bersabe@uma.es

In the present work we compare students with the same level of knowledge of an academic subject (the same grade), but different level of risk on replying to questions in an exam. Why do we think risk may be related to the calibration of probabilities? Some research shows that when subjects bet in games of chance, in which they risk losing what they have bet, a multitude of cognitive biases appear (Bersabé, 1996; Bersabé & Martínez Arias, 1999; Bersabé & Martínez Arias, 2000; Gaboury & Ladouceur, 1989; Griffiths, 1994; Ladouceur & Gaboury, 1988; Wagenaar, 1988; Walker, 1992). One of these is the illusion of control that emerges when players believe they can control luck with their skill (e.g., by choosing particular numbers in the lottery, throwing the dice in a special way, etc.). Langer (1975) observed that, in these situations, the expectation of personal success was inappropriately higher than what was guaranteed by objective probability. Thus, when we face a situation of uncertainty in which we have something to lose, we are influenced by cognitive biases, and overconfidence ensues. Therefore, as an initial hypothesis, we shall suppose that the most reckless, that is, those who expose themselves to greater losses or failures, will be more overconfident, so that they will also calibrate more poorly. In any case, we think it appropriate to analyze not only calibration, but all the accuracy components in probability judgements, so that we are in a better position to understand how judgements are formed, and how they can be modified.

METHOD

Participants

The sample was made up of 218 students on the Psychometrics course at the Psychology Faculty of the Complutense University in Madrid (Spain). Age range was 19 to 25 years (Mean=19.53; SD=1.07). The lowest-risk group consisted of 22 men and 87 women, while the highest-risk group comprised 25 men and 84 women.

Materials

A theoretical examination in Psychometrics made up of 20 items, each with two response alternatives (True-False). An instruction sheet and questionnaire with the same 20 items on which subjects had to judge the probabilities of an item being true and false (redundant information). Of the 20 items presented, 11 were true and 9 false. The theoretical examination, together with a practical one, determined the grade students obtained in that subject.

Procedure

First of all, in a natural situation, subjects took a multiple-choice theoretical examination with 20 items, each of which had just two alternatives (T-F). They were asked to circle the option they considered correct, and could leave questions out if they wished. Once the examination was over, they were asked to read the instruction sheet, on which it was explained how they were to fill out the questionnaire. They were presented with the same 20 questions as in the examination, and asked to judge the probabilities of that item being true and false. This was done by means of percentages. For example, a response of "true=0% and false=100%" would indicate that the subject was totally sure the item was false. On the other hand, a response of T=50% and F=50% would mean that the two alternatives were equally probable. It was made clear on the instruction sheet, moreover, that the test was voluntary, that it would not affect the subject's mark in the exam, and that the purpose was to study the quality of multiple-choice tests. It was hoped in this way to eliminate possible "social desirability" effects, which may lead students to want to appear more confident in their probability judgements than they actually were.

In the majority of studies on calibration, the method employed is different from the one used here. In general, subjects are first presented with questions and told to reply to all of them: a forced-choice task. Secondly, they are asked to indicate (also by means of percentages) the extent to which they think the response given to each item is correct. Thus, these responses are between 50% and 100%, because it is not expected to be mistaken more than chance would dictate. In our study, the first task, responding to the items, is a non-forced-choice task (items can be left unanswered), and the judgement required of subjects refers to whether or not the item is true (not whether their response is correct). It is important to stress these distinctions in the procedure, given their repercussion with regard to the interpretation of the measures of accuracy in the judgments. The percentages estimated by subjects are converted into eleven categories of subjective probabilities: [0-0.05) [0.05-0.15) [0.15-0.25) [0.25-0.35) [0.35-0.45) [0.45-0.55] (0.55-0.65] (0.65-0.75] (0.75-0.85] (0.85-0.95] (0.95-1].

For the purposes of our research it was necessary to form two groups of subjects according to the level of risk they displayed. As Yates (1990) points out, the gen-

eral concept of risk can be described as facing the possibility of losing or failing in some way in order to obtain greater benefit. Thus, it was considered that a student would more reckless than another if, with the same knowledge (the same grade), he or she answered more questions. Students acting in this way supposedly do so to try and answer more questions at the risk of also making more mistakes. Following this idea, two risk groups were formed as follows: students were classified according to the grade they obtained in the 20 items (correct answers minus errors). Sixteen different grades were found. Students who obtained the same grade were divided into two groups according to number of responses. When there was an odd number of students with the same grade, the central one was removed. Thus, the students were distributed uniformly in two groups with identical grades (Mean = 6.69; SD=3.29), but different numbers of responses (Mean=13.61; SD=1.66 in the low-risk group, and Mean=17.17; SD=1.41 in the high-risk group). The mean number of responses was indeed different in the two risk groups ($t_{(216)}$= -17.04; p<.001).

In order to ensure that the risk groups were properly formed, Figure 1 was drawn, showing the percentages of responses for each subjective probability. As it can be seen, the most reckless students do not need to be so sure that the item is true or false to respond to it. In other words, their "response threshold" is lower.

RESULTS

Several probability judgement accuracy measures were computed individually for each subject. Let us consider each of these measures, described more fully in Yates (1990).

Overall Accuracy

Mean Probability Score ($\overline{PS}$). The probability judgement accuracy measure most commonly used is attributed to Brier (1950), and is known as "Brier Score", Quadratic Score, or mean Probability Score. It is defined as:

$$\overline{PS} = \frac{\sum (f_i - d_i)^2}{N}$$

In the present study,

f_i subjective probability that the item is true

$d_i=1$ if item is true

$d_i=0$ if item is false

Therefore, $\overline{PS}$ is a quadratic function that measures the difference between each one of the subjects' judgements and what actually occurs. $\overline{PS}$ scores vary within the [0,1] interval. The more accurate the judgements, the lower the $\overline{PS}$. In the present study, the least reckless subjects showed a mean $\overline{PS}$ somewhat lower than that of the conservative subjects, though the difference was not statistically significant (see Table 1).

If a student were absolutely unable to predict whether an item was true or false, he or she could give for all items a subjective probability equal to 50%. This strategy is called *uniform judge*. Following Yates (1990), it can be shown that to this subject would correspond a $\overline{PS}$=.25. As Table 1 indicates, the percentage of subjects over the uniform judgement accuracy level was also slightly higher in the low-risk group.

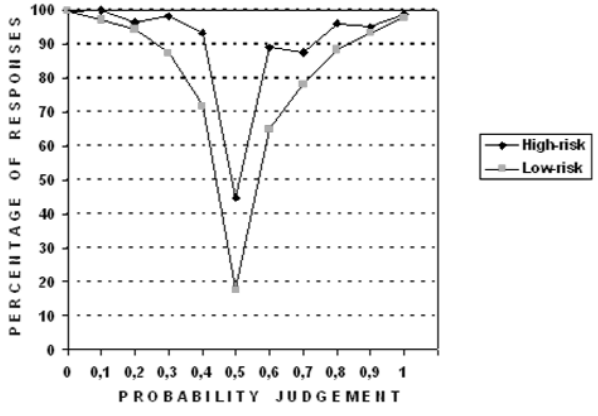
Accuracy in probability judgements is not an undifferentiated concept; rather, it can be broken into different components: calibration, discrimination and noisiness (Murphy, 1973; Yates, 1982).

Calibration

Three indices of calibration can be obtained:

Calibration-in-the-small is measured by the calibration-in-the-small index (CIS). We have already seen how, in the Probability Score, each probability judgement was compared with what actually happened in each one of the items (0: false; 1: true). The CIS, however, takes all the items in which the student has estimated the same subjective probability and verifies how many of these items are actually true, in order to compare subjective probability with actual proportion. For example, let us suppose a student has estimated a probability of .7 in 10 items. With perfect calibration, 7 of

Figure 1
Percentage of responses according to subjective probability estimated by high- and low-risk groups



these 10 items would actually be true, and 3 would be false. The CIS is computed as follows:

$$CIS = \frac{\sum N_j (f_j - \bar{d}_j)^2}{N}$$

where,

- j indicates the different categories of subjective probabilities (0, .1, .2, .3, .4, .5, .6, .7, .8, .9, and 1),
- N_j is the number of judgements registered in category j of subjective probability;
- $\bar{d}_j$ is the actual proportion of true items, considering those where category j of subjective probability was estimated.

As emerges from the formula, the lower the CIS score, the better the calibration. The data reveal that the mean of the CIS index is poorer (higher) in the high-risk group (Table 1). It seems that the subjective probabilities of the risk-takers are farther removed from the reality. Thus, those who take most risk are those that are least aware of how much risk they are taking.

Most of the works on calibration provide calibration graphs. In order to draw these, it is necessary to collect the judgements of all the subjects in each group, that is, as though the students had taken it in turns to make their judgements. Therefore, each group (or "macrosubject") obtained a single score in each judgement accuracy measure. These are the scores that are shown in the graphs.

In the calibration graphs, the CIS is represented as the deviations of the function with respect to the 1:1 diagonal. If a subject calibrates perfectly in all the categories of his or her judgements, the calibration curve will coincide with the 1:1 diagonal. Figure 2 shows the calibration curves of each risk group. In them it is difficult to appreciate the differences found in the analysis of the individual data. On combining the judgements in each group, the differences in the CIS seem to disappear.

A second type of calibration is the *calibration-in-the-large* (CIL). If the calibration were perfect, then the mean of *all* judgements on an event ($\bar{f}$) should coincide with the proportion of times it actually occurs ($\bar{d}$). Calibration-in-the-large is operativized by means of the *Bias* statistic:

$$Bias = \bar{f} - \bar{d}$$

or with its square, the *calibration-in-the-large* index (CIL):

$$CIL = Bias^2$$

The higher the absolute value of the Bias, the poorer the CIL. With a single subject, the Bias and the CIL provide redundant information. However, with a group of subjects, they can throw light o a number of issues. For example, if half the subjects in a group obtained a Bias of 0.15 and the other half -0.15, the mean Bias would be nil. On the other hand, the subjects' judgements taken individually would provide a poor CIL.

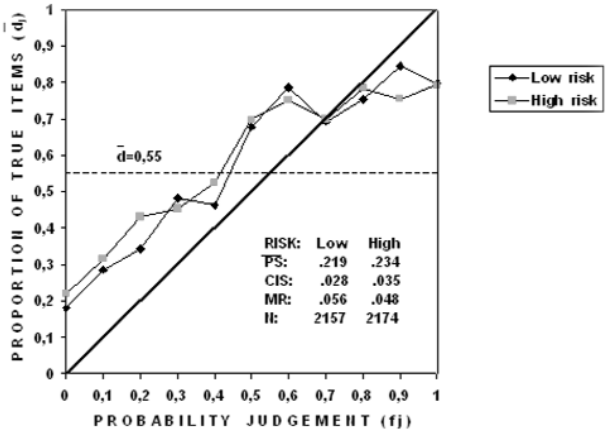
In those studies in which subjects are asked to estimate the probability (.5-1) that the response given is correct, Bias is a good indicator of over/underconfidence (Lichtenstein & Fischhoff, 1977). A positive Bias is found when there is overconfidence, that is, when subjects judge themselves to have a given more right answers than they actually have. On the other hand, underconfidence results in a negative Bias. However, in this

Table 1
Probability judgement accuracy measures in each risk group

Accuracy measure [a]	Low risk (N=109) Mean	(SD)	High risk (N=109) Mean	(SD)	t (216)	Two-tailed significance
Overall						
$\overline{PS}$ ↓	.221	(.062)	.234	(.064)	-1.527	.128
$\overline{PS}$ <.25 ↑[b]	61.9%		54.8%		.811	.368
Calibration						
CIS ↓	.084	(.049)	.098	(.047)	-2.041	**.042**
Bias 0	-.103	(.066)	-.110	(.068)	.776	.439
CIL ↓	.015	(.015)	.017	(.016)	-.856	.393
Discrimination						
MR ↑	.112	(.048)	.110	(.048)	.371	.711
Slope ↑	.284	(.139)	.284	(.161)	.049	.961
Noisiness						
Scat (f) ↓	.074	(.033)	.085	(.036)	-2.171	**.031**

[a] ↓: smaller values better; ↑: larger values better; 0: the best value is zero.
[b] Comparison between percentages via Chi-square test with continuity correction.

Figure 2
Calibration curves in the high- and low-risk groups



study the task consists in judging the probability of the item (answered or not) being true (not of the answer given being correct). Thus, a negative Bias, as found in the two risk groups (see Table 1), does not indicate underconfidence, but rather greater confidence in the falsity than in the truth of the 20 items. This negative Bias may be due to the fact that there are more true items than false ones, so that the base rate ($\bar{d}$) is higher than .5. It may also be due to a response bias whereby there was greater confidence in the false questions than in the true ones. In any case, what does seem clear is that subjects' risk is not significantly related to Bias or to CIL.

Bias is shown in the covariance graphs (Figures 3a and 3b) by means of the intersection between the vertical line of the base rate ($\bar{d}$) and the horizontal line of the mean subjective probability ($\bar{f}$). If this intersection falls on the 1:1 diagonal, then the Bias is nil, since the mean subjective probability coincides with base rate. Bias is positive when the intersection falls above the diagonal. Bias is negative when it falls below, as in the two graphs presented in Figure 3 (a and b). This confirms what had already been obtained numerically.

Discrimination

Calibration refers to the ability to indicate appropriately the different probabilities of an event occurring. In contrast, discrimination refers to the tendency to say something somewhat different when an event occurs than when it does not. The extent to which a set of judgements approaches the ideal of discrimination is reflected in the *Murphy Resolution* statistic (MR):

$$MR = \frac{\sum N_j(\bar{d}_j - \bar{d})^2}{N}$$

The means of the MR statistic were found to be quite similar in the two risk groups (Table 1). This is logical if we consider that the groups were formed by evening up the grade (correct answers minus errors), which is directly related to the ability to discriminate between true and false items.

The MR statistic is also reflected in the calibration graphs (Figure 2). In these, the base rate ($\bar{d}$) is drawn in a horizontal line. MR will be extremely poor when this horizontal line overlaps with the calibration curve. This case would occur when the subject gave the different subjective probabilities randomly, without being able to discriminate when the item is true and false. The actual proportion of true items will then be expected to equal the base rate over the long run for any subjective probability. On the other hand, any subject capable of discriminating the items perfectly would judge probabilities higher than 0.5 when the item is true, and under 0.5 when it is false. Thus, the calibration curve for all the subjective probabilities over 0.5 would have a height of

Figure 3.a.
Covariance graph for the low-risk group

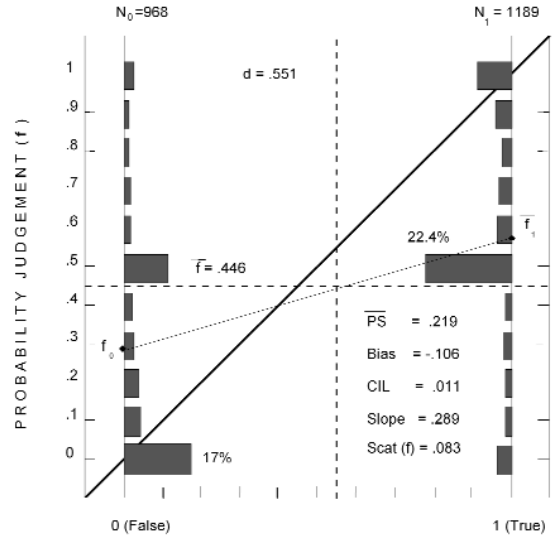
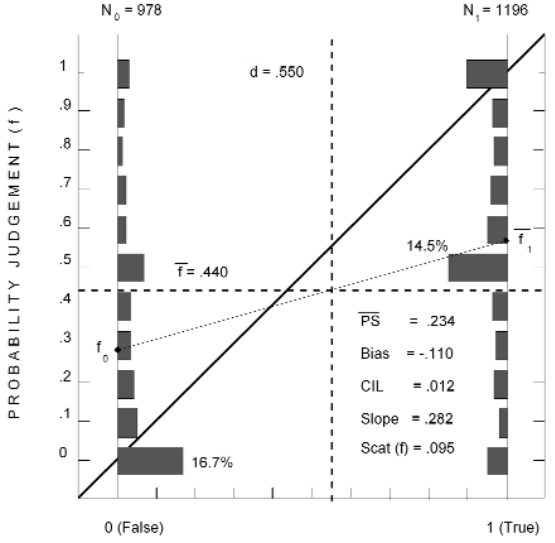


Figure 3.b.
Covariance graph for the high-risk group



1 (all those items would be true), and for those under 0.5, of 0 (all those items would be false). Therefore, MR is represented as the distances between each point (height) of the calibration curve and the height of the base rate. As can be seen in Figure 2, the calibration curves of the two risk groups approach the base rate in a similar way, that is, they show similar discrimination.

Slope is another measure of the ability to discriminate when an event takes place and when it does not. When judgements are accurate in this respect, the mean of subjective probabilities when the item is true ($\bar{f}_1$) will tend to be greater than when it is false ($\bar{f}_0$). Therefore, the Slope is computed as:

$$Slope = \bar{f}_1 - \bar{f}_0$$

Naturally, this measure of discrimination between true and false items also depends on knowledge of the material or subject. For this reason, the mean of the Slopes were practically identical in the two risk groups (see Table 1), since the grade was evened up.

In the covariance graphs, Slope is reflected in the gradient of the straight line that links $\bar{f}_0$ with $\bar{f}_1$. Maximum discrimination corresponds to the 1:1 diagonal (Slope = 1-0 = 1). As it can be seen in the covariation graphs (Figures 3a and 3b), both slopes have almost the same gradient.

Noisiness

The last aspect of judgement accuracy analyzed was the noisiness, which involves a particular form of variation. The amount of random variation, or "scatter", in a collection of judgements is indexed by the variances of the judgements conditional on whether the item is true ($Var(f_1)$) or false ($Var(f_0)$). The *Scatter* measure provides joint information on this variability, which is just a weighted mean of these conditional variances. Thus, the *Scatter* statistic (Scat (f)) is:

$$Scat(f) = \frac{N_1 Var(f_1) + N_0 Var(f_0)}{N_1 + N_0}$$

where, N_1 is the number of true items, and
N_0 the number of false items.

The reckless students obtained a mean Scat(f) value significantly higher than the conservative ones (see Table 1). Noisiness is shown in covariance graphs. These graphs are actually opposing bar charts: that on the left for the subjective probabilities when the item is false; and that on the right for the subjective probabilities when the item is true. The greater the variance in both the left and the right bar graphs, the greater the Scat(f). In the covariance graphs, it can be seen that in the high-risk group (Figure 3a) the variance of the judgements is somewhat higher than in the low-risk group (Figure 3b).

DISCUSSION

The students were divided into two groups with identical grade but different level of risk. The most reckless answered more questions with the same level of subjective confidence, exposing themselves to more mistakes in order to try and make more correct responses.

Both the grade and the measures of discrimination between true and false items reflect in some way the subject's knowledge of the material. Thus, as expected the discrimination measures were similar, since the grades were the same. In any case, what interested us especially was to answer a question along the lines of the title of this article: are those who take risks more aware of how much of a risk they are taking? According to our results, just the opposite occurs. Those most inclined to risk calibrated more poorly: that is, the subjective probabilities they estimated deviated more from the real proportion of true items. Thus, it seems that those who take most risks know least about the risks they are taking.

The difference in calibration between the two risk groups was not clearly appreciated in the calibration graphs. These graphs should represent the probability judgements of all the subjects in each group, so that each group obtains a single score in each accuracy measure. As Björkman (1992) points out, individual analysis shows that the group scores do not reflect the full story. Calibration and resolution are attributes of individuals, with considerable variation from subject to subject. For this reason, individual analysis of the data is more appropriate, even if it is via the means and standard deviations of each group.

Our results have implications not only for the theoretical study of judgements, but also for the understanding of the attitude of risk-taking that characterizes gamblers. Thus, the clearly reckless behaviour of gamblers who systematically risk losing their money may be related to

an inadequate calibration of their probability judgements. If this were confirmed, it would be interesting to incorporate methods for correcting Bias (Fischhoff, 1982) into cognitive therapies for the treatment of pathological gambling. Indeed, some of these methods have already been applied with relative success as part of wider cognitive restructuring therapy (Ladouceur, Sylvain, Letarte, Giroux & Jacques, 1998; Sylvain, Ladouceur & Boisvert, 1997).

Finally, we should like to mention an aspect that deserves special attention in the study of risk: the naturalness of the situation in which the tests are applied. In our study, the first task (taking a real exam) allowed us to observe subjects' level of risk – not only because it was a non-forced-choice task, but also, and more important, because the possible consequences of answering the questions or not, passing or failing, were relevant for the student. Many of the studies on decision-making in situations of risk use imaginary settings in which subjects have to make a choice between two or more fictional alternatives, described with information on the possible consequences and their probabilities. Subjects' level of risk in the same task may vary depending on whether the situation is natural or laboratory-based, so that it is crucial to ensure ecological validity in this type of study.

REFERENCES

Bersabé, R. (1996). *Sesgos cognitivos en los juegos de azar: La ilusión de control.* Published doctoral dissertation. Universidad Complutense de Madrid.

Bersabé, R., & Martínez Arias, R. (1999). La superstición en los juegos de azar. *Clínica y Salud, 10,* 151-167.

Bersabé, R., & Martínez Arias, R. (2000). Superstition in gambling. *Psychology in Spain, 4,* 28-34.

Björkman, M. (1992). Knowledge, Calibration, and Resolution: A linear Model. O*rganizational Behavior and Human Decision Processes, 51,* 1-21.

Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1-3.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristic and biases* (pp. 422-444). New York: Cambridge University Press.

Gaboury, A., & Ladouceur, R. (1989). Erroneous perceptions and gambling. *Journal of Social Behaviour and Personality, 4,* 411-420.

Griffiths, M. D. (1994). The role of cognitive bias and skill in fruit machine gambling. *British Journal of Psychology, 85,* 351-369.

Keren, G. (1991). Calibration and probability judgments: Conceptual and methodological issues. *Acta Psychologica, 77,* 217-273.

Ladouceur, R., & Gaboury, A. (1988). Effects of limited and unlimited stakes on gambling behavior. *Journal of Gambling Behavior, 4,* 119-126.

Ladouceur, R., Sylvain, C., Letarte, H., Giroux, I, & Jacques, C. (1998). Cognitive treatment of pathological gamblers. *Behaviour Research and Therapy, 36,* 1111-1119.

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

Langer, E. J. (1975). *The psychology of control*. Beverly Hills, CA: Sage.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 306-334). New York: Cambridge University Press.

Martínez Arias, R. (1991). El proceso de toma de decisiones. In R. Martínez Arias & M. Yela (Eds.), *Pensamiento e Inteligencia* (pp. 411-494). Madrid: Alhambra Universidad.

Murphy, A. H. (1983). A new vector partition of the probability score. *Journal of Applied Meteorology, 12,* 595-600.

O'Connor, M. J. (1989). Models of human behavior and confidence in judgment: a review. *International Journal of Forecasting, 5,* 159-169.

Sylvain, C., Ladouceur, R., & Boisvert, J. M. (1997). Cognitive and behavioral treatment of pathological gambling: A controlled study. *Journal of Consulting and Clinical Psychology, 65,* 727-732.

Wagenaar, W. A. (1988). *Paradoxes of Gambling Behaviour.* Hove: Lawrence Erlbaum.

Walker, M. B. (1992). *The Psychology of Gambling.* Oxford: Pergamon.

Yates, J. F. (1982). External correspondence: Decompositions of the mean probability score. *Organizational Behavior and Human Performance, 30,* 132-156.

Yates, J. F. (1990). *Judgment and Decision Making.* Englewood Cliffs, New Jersey: Prentice Hall.