ECONOMETRICS AND DECISION MAKING: EFFECTS OF COMMUNICATION OF RESULTS

Emre Soyer

TESI DOCTORAL UPF / 2012

DIRECTOR DE LA TESI

Prof. Robin M. Hogarth

DEPARTMENT OF BUSINESS AND ECONOMICS

UNIVERSITAT POMPEU FABRA

to Anne and Baba

Acknowledgements

“This is a neat idea,” he said with a kind smile on his face. “I would encourage you to pursue it. You’ll have my support. But you have to know that this would be a hard and dangerous road to take.”

Robin Hogarth was right!

At the end, it is thanks to him that I managed to overcome the obstacles he had warned me about. I am grateful for all his time, energy, wisdom, walks, laughs, ideas. Thanks to him, I am a better person.

There are many other friends and colleagues who, upon my unpredictable requests, often disrupted their routines to help me with my PhD mission. Among them are Rosemarie Nagel, Gert Cornelissen, Johannes Mueller-Trede, Gaël Le Mens, Jose Apesteguia, Thijs van Rens, Nick Longford, Marta Araque, Berkay Özcan and Eser Sakarya. I am grateful for their teachings and continuous support.

Finally, without Hale, Mehmet and Ipek, I would be truly lost and unhappy - I love you.

Abstract

This thesis incorporates three studies that analyze how information is presented in various contexts, how these different modes of presentation affect decision makers' perceptions and how to improve communication of information to eliminate distortions. Chapter 1 features a scenario where experts make inferences given different presentations of a regression analysis, a widely used statistical method. Chapter 2 introduces an experience-based presentation mode and tests its effectiveness on decision makers with varying statistical abilities, across multiple probabilistic tasks. Chapter 3 demonstrates the effects of presentation mode and the number of available options on the amounts and distributions of donations to NGOs and their campaigns. Overall, the findings suggest that presentation mode is an important determinant of judgments and decisions, and they can be restructured to improve the accuracy of inferences.

Resumen

Esta tesis incluye tres estudios que analizan cómo la información se presenta en varios contextos, cómo estos diferentes modos de presentación influyen las percepciones de los tomadores de decisiones y cómo mejorar la comunicación de la información para eliminar distorsiones. Capítulo 1 analiza una situación donde expertos hacen inferencias utilizando diferentes presentaciones de un análisis de regresión, un método de estadística ampliamente utilizado. Capítulo 2 introduce un modo de presentación basado en experiencia y pone a prueba su eficacia a través de múltiples problemas probabilísticas. Capítulo 3 demuestra los efectos del modo de presentación y el número de opciones disponibles sobre las cantidades y la distribución de las donaciones a las ONG y sus campañas. En general, los resultados sugieren que el modo de presentación es un determinante importante de las percepciones y decisiones, y pueden ser reestructuradas para mejorar la precisión de las inferencias.

Preface

The assumption of rationality asserts that presentation mode should not affect interpretation and analysis. However, there is considerable psychological evidence suggesting that different presentations of the same problem might lead to different inferences. This notion is mainly fueled by the works of Daniel Kahneman and Amos Tversky, who show that even subtle changes in questions designed to induce preferences are subject to contextual influences (Kahneman & Tversky, 1979; Tversky & Kahneman, 1981, 1983, 1986).

Description is considered to be the primary method of presenting information and its effects are widely scrutinized in the judgment and decision making literature. Among influential studies that investigate how individuals might be framed by the contents and structures of descriptions are Brunswik (1952), Simon (1978), Hogarth (1982), Sedlmeier (1999), Hoffrage, Lindsey, Hertwig and Gigerenzer (2000). More recently, Thaler and Sunstein (2008) further popularized the notion by advocating the optimization of descriptions to improve decision outcomes; a phenomenon referred to as “choice architecture.”

Overall, the aforementioned literature and the current thesis identify presentation mode as an important determinant for judgments and decisions. The fact that descriptions are easy to modify and restructure only adds to the relevance of the topic and leads to questions on how they could be constructed to effectively improve the accuracy of inferences. Hence, studies featured in this thesis will not only investigate presentation effects, but they will also introduce and prescribe methodologies to improve communication of information. Some other essential aspects of the issue, such as the freedom of choice, the accessibility of relevant

information and the number of available options, will also be explored in the discussion sections of appropriate chapters.

Chapter 1 aims to demonstrate the implications of presentation mode, and in particular of description, in a specific situation and to define its boundary conditions. Noting the predominant role of regression analysis in empirical economics, it surveys the ability of knowledgeable decision makers to make inferences based on the outputs of this statistical tool. The findings demonstrate that currently employed presentations of statistical outputs of regression analyses induce an illusion of predictability of outcomes, i.e. an erroneous belief that the analyzed outcomes are more predictable than what the estimation indicates. The survey also reveals that the inferences of participants are most accurate when only graphs are provided. The implications of this study suggest, inter alia, the need to reconsider how to present estimation outputs to better acknowledge what Ziliak and McCloskey (2008) call the "economic significance" of empirical results.

The chapter is based on Soyer and Hogarth (2012a), which is debated in a series of discussion papers among scholars in the fields of decision making and prediction (Armstrong, 2012; Ord, 2012; Taleb & Goldstein, 2012; Ziliak; 2012, Soyer & Hogarth, 2012b).

Providing only graphs, however, is not a credible solution to the problem in hand. While such an approach helps to identify the source of the problem, it eliminates parts of the presentation that are essential for the interpretation of other important aspects of the analysis, such as the average effects and the statistical validity of findings. Therefore, the first chapter ends with a proposal about the possible provision of add-on and

easy-to-use simulation tools that would enable consumers of empirical analyses to make accurate inferences.

Chapter 2 elaborates on the simulation methodology proposed in Chapter 1. It argues that such simulations would provide decision makers with the appropriate experience that would constitute a valid basis for their judgments. In that sense, it features and tests the reliability of a presentation mode that is based on experience, and not on description. Hence, not only it provides a viable alternative to the descriptions featured in Chapter 1, but it also introduces a presentation mode that complies with the recent research on risky decision making, which has argued that in many situations, people do not have access to synthetic descriptions of probabilistic information (Hertwig, Barron, Weber, & Erev, 2004; Weber, Shafir, & Blais, 2004).

The chapter is based on Hogarth and Soyer (2011) and hypothesizes that experiencing data for statistical problems in the form of sequentially simulated outcomes can lead to more accurate inferences than typical, analytical descriptions. It features two experiments to test the idea. The first one involves seven well-known probabilistic inference tasks and demonstrates that individuals relate easily to the simulated experience technology. The second experiment features a hypothetical investment decision comparing responses of a group given an analytical presentation with that of two groups exposed to simulated experience. Results indeed show significant positive effects of experience over analysis in the accuracy of statistical inferences, regardless of decision makers' statistical knowledge.

The effectiveness of the simulated experience methodology in aiding judgments prompts to question its applicability in settings where

probabilistic structures are complex and hard to describe. A first attempt to adapt this presentation mode to a real decision scenario involved simulating natural disaster scenarios based on the models proposed by Kunreuther and Michel-Kerjan (2010) in an effort to promote multi-period insurance schemes over single-period ones.

Another recent study that features the simulated experience methodology to aid judgments is Hogarth, Mukherjee and Soyer (2012). The probabilistic problem posed in this case is a contest-entry situation, where decision makers show considerable difficulty in assessing their chances of success. The paper features an experiment where participants are provided with experience in the form of sequentially simulated outcomes, which is shown to help them improve their assessments of success probabilities and consequently their contest-entry decisions.

Chapter 3 deals with how the presentation and the availability of different number of alternatives affect judgments and decisions, specifically in the context of charitable giving. Through a field study conducted on the general population in Spain, it analyses the effects of number of NGOs and charity campaigns on the amounts and distributions of donations. The findings suggest that when individuals are presented with more options; a) they contribute more, b) they give more to recipients that they are more knowledgeable about, and c) the distributions of their contributions change with the number of available options and this change is different in the case of NGOs and campaigns.

The chapter is based on Soyer and Hogarth (2011) and explores also the possible reasons why donors would behave differently when they are provided with varying numbers of options. It argues that in the context of charitable giving, more options induce a perception of a larger need for

aid, which in turn leads to more contributions. Moreover, the experimental conditions that feature campaigns reveal that the structures of current online interfaces employed by NGOs when asking for donations lead to a reduction in individuals' willingness to contribute, e.g. the use of drop-down menus when offering multiple options reduces potential donations.

The findings of this final chapter have been well received by the NGO community. Several organizations from Spain and Turkey, including Intermon Oxfam and TEMA, showed interest in the implications of the analysis and are currently considering incorporating the suggested strategies in their resource generation processes.

Table of contents

Soyer, E., & Hogarth R. M. (2012a). The illusion of predictability: How regression statistics mislead experts**.** *International Journal of Forecasting* (in press).

1. THE ILLUSION OF PREDICTABILITY: HOW REGRESSION STATISTICS MISLEAD EXPERTS

(Based on Soyer & Hogarth, 2012a)

1.1. Introduction

Much academic research in empirical economics involves determining whether or not one or several variables have causal effects on another. Typically, the statistical tool used to make such affirmations is regression analysis where the terms "independent" and "dependent" are used to distinguish cause(s) from outcomes. The results from most analyses consist of statements as to whether particular independent variables are or are not "significant" in affecting outcomes (the dependent variable) and discussions of the importance of such variables focus on the "average" effects on outcomes due to possible changes in inputs.

However, if the analysis is used for prediction, emphasizing only statistically significant average effects is an incomplete characterization of the relation between an independent and dependent variable. It is also essential to acknowledge the level of uncertainty in outcomes of the dependent variable conditional on values of the independent variable. For example, consider a decision maker who is pondering which actions to take and how much to do so in order to reach a certain goal. This requires forming conjectures about individual outcomes that would result from specific inputs. Moreover, the answers to these questions depend not only on estimating average effects but the distribution of possible effects around the average as well.

In this chapter, we argue that the emphasis on determining average causal effects in the economics literature limits the ability to make correct probabilistic forecasts. In particular, the way results are presented in

regression analyses obfuscates the uncertainty inherent in the dependent variable. As a consequence, consumers of economic literature can be subject to what we call the "illusion of predictability."

Whereas it can be argued that how information is presented should not affect rational interpretation and analysis, there is abundant psychological evidence demonstrating presentation effects. Many studies have shown, for example, how subtle changes in questions designed to elicit preferences are subject to contextual influences (see, e.g., Kahneman & Tversky, 1979). Moreover, these have been reported in both controlled laboratory conditions and field studies involving appropriately motivated experts (Camerer, 2000; Thaler & Sunstein, 2008). Human information processing capacity is limited and the manner in which attention is allocated has important implications for both revealed preferences and inferences (Simon, 1978).

Recently, Gigerenzer and his colleagues (Gigerenzer et al., 2007) reviewed research on how probabilities and statistical information are presented and consequently perceived by individuals or specific groups that use them frequently in their decisions. They show that mistakes in probabilistic reasoning and miscommunication of statistical information are common. Their work focuses mainly on the fields of medicine and law, where in particular situations, doctors, lawyers and judges fail to communicate crucial statistical information appropriately thereby leading to biased judgments that impact negatively on others. One example is the failure of gynecologists to infer correctly the probability of cancer given the way mammography results are communicated.

We examine how economists communicate statistical information. Specifically, we note that much work in empirical economics involves the estimation of average causal effects through the technique of regression

analysis. However, when we asked a large sample of economists to use the standard reported outputs of the simplest form of regression analysis to make probabilistic forecasts for decision making purposes, nearly 70% of them experienced difficulty. The reason, we believe, is that current reporting practices focus attention on the uncertainty surrounding model parameter estimates and fail to highlight the uncertainty concerning outcomes of the dependent variable conditional on the model identified. When attention was directed appropriately – by graphical as opposed to tabular means – over 90% of our respondents made accurate inferences.

In the next section (1.2), we provide some background on the practice and evolution of reporting empirical results in journals in economics. In section 1.3 we provide information concerning the survey we conducted with economists that involved answering four decision-oriented questions based on a standard format for reporting results of regression analysis. We employed six different conditions designed to assess differential effects due to model fit (R^2) and different forms of graphical presentation (with and without accompanying statistics). In section 1.4, we present our results: In brief, our study shows that the typical presentation format of econometric models and results – one mainly based on regression coefficients and their standard errors – leads economists to ignore the level of predictive uncertainty implied by the model and captured by the standard deviation of the estimated residuals. As a consequence, there is a considerable illusion of predictability. Adding graphs to the standard presentation of coefficients and standard errors does little to improve inferences. However, presenting results in graphical fashion alone improved accuracy. The implications of our findings, including suggestions on how to improve statistical reporting, are discussed in section 1.5.

1.2. Current practice

There are many sources of empirical analyses in economics. To obtain a representative sample of current practice, we selected all the articles published in the 3rd issues (of each year) of four leading journals between 1998 and 2007 (441 articles). The journals were *American Economic Review* (AER), *Quarterly Journal of Economics* (QJE), *Review of Economic Studies* (RES) and *Journal of Political Economy* (JPE). Among these articles, we excluded those with time series analyses and only included those with cross-sectional analyses where authors identify one or more independent variables as a statistically significant cause for relevant economic and social outcomes. Our aim is to determine how the consumers of this literature translate findings about average causal effects into perceptions of predictability.

Many articles published in these journals are empirical. Over 70% of the empirical analyses use variations of regression analysis of which 75% have linear specifications. Regression analysis is clearly the most prominent tool used by economists to test hypotheses and identify relations among economic and social variables.

In economics journals empirical studies follow a common procedure to display and evaluate results. Typically, authors provide a table that displays descriptive statistics of the sample used in the analysis. Before or after this display, they describe the specification of the model on which the analysis is based. Then the regression results are provided in detailed tables. In most cases, these results include the coefficient estimates and their standard errors along with other frequently reported statistics, such as the number of observations and R^2.

Table 1.1 summarizes these details for the sample of studies referred to above. It shows that, apart from the regression coefficients and their standard errors (or *t*-statistics), there is not much agreement on what else should be reported. The data suggest, therefore, that economists probably understand well the inferences that can be made about regression coefficients or the average impact of manipulating an independent variable; however, their ability to make inferences about other probabilistic implications is possibly less well developed (e.g., predicting individual outcomes conditional on specific inputs).

Table 1.1. Distribution of types of statistics provided by studies in sample of economics journals.

Studies that	Journals: AER	QJE	JPE	RES	Total	% of Total
...use linear regression analysis	42	41	15	13	111	x
...provide both the sample standard deviation of the dependent variable(s) and the R^2 statistic	16	27	11	12	66	59
...provide R^2 statistics	30	32	15	12	89	80
...provide the sample standard deviation of the dependent variable(s)	21	32	11	13	77	69
...provide the estimated constant, along with its standard error	19	14	4	1	38	34
...provide a scatter plot	19	16	5	2	42	38
...provide the standard error of the regression (*SER*)	5	3	1	1	10	9

It is not clear when, how, and why the above manner of presenting regression results in publications emerged. No procedure is made explicit in the submission guidelines for the highly ranked journals. Moreover, popular econometric textbooks, such as Greene (2003), Judge et al. (1985) and Gujarati and Porter (2009) do not explain specifically how to present results or how to use them for decision making. Hendry and Nielsen

(2007) address issues regarding prediction in more detail than other similar textbooks. Another exception is Wooldridge (2008), who dedicates several sections to issues of presentation. His outline suggests that a good summary consists of a table with selected coefficient estimates and their standard errors, R^2 statistic, constant, and the number of observations. Indeed, this is consistent with today's practice. More than 60% of the articles in Table 1.1 follow a similar procedure.

Zellner (1984) conducted a survey of statistical practice based on articles published in 1978 in the AER, JPE, *International Economic Review*, *Journal of Econometrics* and *Econometrica*. He documented confusion as to the meaning of tests of significance and proposed Bayesian methods to overcome theoretical and practical problems. Similarly, McCloskey and Ziliak (1996) provided an illuminating study of statistical practice based on articles published in AER in the 1980s. They demonstrated widespread confusion in the interpretation of statistical results due to confounding the concepts of statistical and economic or substantive significance. Too many results depended on whether *t* or other statistics exceeded arbitrarily defined limits. In follow-up studies, Ziliak and McCloskey (2004; 2008) report that, if anything, this situation worsened in the 1990s. (See also Zellner, 2004.)

Empirical finance has developed an illuminating manner of determining the significance of findings. In this field, once statistical analysis has identified a variable as "important" in affecting, say, stock returns, it is standard to assess "how important" by evaluating the performance of simulated stock portfolios that use the variable (see, e.g., Jensen, 1968; Carhart, 1997).

In psychology, augmenting significance tests with effect size became a common practice in the 1980's. For example, in its submission guidelines, *Psychological Science*, the flagship journal of the Association for Psychological Science, explicitly states, "effect sizes should accompany major results. When relevant, bar and line graphs should include distributional information usually confidence intervals or standard errors of the mean."

In forecasting, Armstrong (2007) initiated a discussion on not only the necessity to use effect size measures when identifying relations among variables, but also on how significance tests should be avoided when doing so. He argues that significance tests are often misinterpreted and, even when presented and interpreted correctly, they fail to contribute to the decision making process. Schwab and Starbuck (2009) make an analogous argument for management science.

In interpreting the results of linear regression analysis from a decision making and predictive perspective, two statistics can convey meaning to readers about the level of uncertainty in results. These are R^2 and the Standard Error of the Regression (*SER*).[1] As a bounded and standardized quantity, R^2 describes the fit of a model. *SER*, on the other hand, provides information on the degree of predictability in the metric of the dependent variable.

Table 1.1 shows that *SER* is practically absent from the presentation of results. Less than 10% of the studies with linear specifications provide it. $R^2$is the prevalent statistic reported to provide an idea of model fit. This is

[1] Some sources refer to *SER* as the Standard Error of Estimate or *SEE* (see RATS), some others as root Mean Squared Error or root-MSE (see STATA). Wooldridge (2008) uses Standard Error of the Regression (*SER*) defining it as "an estimator of the standard deviation of the error term."

the case for 80% of the published articles with a linear specification. Table 1.1 also shows that more than 40% of the publications in our sample that utilize a linear regression analysis (excluding studies that base their main results on IV regression) provide no information on either R^2 or the standard deviation of the dependent variable. Hence, a decision maker consulting these studies cannot infer much about the unexplained variance within the dependent variable and the cloud of data points on which the regression line is fit. Alternatively, a scatter plot would be essential to perceive the degree of uncertainty. However, less than 40% of publications in our sample provide a graph with actual observations.

Given the prevalence of empirical analyses and their potential use for decision making and prediction, debates about how to present results are important. However, it is also important that debates be informed by evidence of how knowledgeable individuals use current tools for making probabilistic inferences, and how different presentation formats affect judgment. Our goal is to provide such evidence.

1.3. Survey

a) Goal and design

How do knowledgeable individuals (economists) interpret specific decision making implications of the standard output of a regression analysis? To find out, we applied the following criteria to select the survey questions. First, we provided information about a well-specified model that strictly met the underlying assumptions of linear regression analysis. Second, the model was straightforward in that it had only one independent variable. Third, all the information necessary to solve the problems posed was available from the output provided. Fourth, although sufficient information was available, respondents had to apply knowledge

about statistical inference to make the calculations necessary to answer the questions.

This last criterion is the most demanding because whereas economists may be used to interpreting the statistical significance of regression coefficients, they typically do not assess the uncertainties involved in prediction when an independent variable is changed or manipulated (apart from making "on average" statements that give no hint as to the distribution around the average).

Our study required that respondents answer four decision making questions after being provided with information about a correctly specified regression analysis. There were six different conditions that varied in the overall fit of the regression model (Conditions 1, 3, and 5 with R^2 = .50, the others with R^2 = .25), as well as the amount and type of information provided. Figures 1.1 and 1.2 report the information provided to the respondents for Conditions 1 and 2, which is similar in form and content to the outputs of many reports in the economic literature (and consistent with Wooldridge, 2008). Conditions 3 and 4 used the same tables but additionally provided the bivariate scatter-plots of the dependent and independent variables as well as the standard deviation of the estimated residuals – see Figures 1.3 and 1.4. In Conditions 5 and 6, the statistical outputs of the regression analyses were not provided but the bivariate graphs of the dependent and independent variables were, as in Figures 1.3 and 1.4.[2] In other words, for these two conditions we were intrigued by what would happen if respondents were limited to only consulting graphs.

[2] We thank Rosemarie Nagel for suggesting that we include Conditions 5 and 6.

Consider the econometric model

$$Y_i = C + \beta X_i + e_i$$

Where:

- Y : Economic payoff, given the choice of X.
- X : A continuous choice variable which is costly to undertake
- C : Constant
- β : The effect of X on Y
- e : Random perturbation; $e_i \mid X_i \sim N[0, \sigma^2]$ with $E(e_i)=0$, $Cov(e_i, e_j)=0$ and $Cov(e_i, X_i)=0$.

In this setting, the goal is to estimate β and C, based on a random sample of X and Y with 1000 observations. The sample statistics are as follows:

Variable	Mean	Std. Dev.
X	50.72	28.12
Y	51.11	40.78

The OLS fit for of the model to this sample gives the following results:

	Dependent Variable: Y
X	1.001 (0.033)**
Constant	0.32 (1.92)
R^2	0.50
N	1 000

Standard errors in parentheses

** Significant at 95% confidence level

N is the number of observations

Results indicate that constant C is not statistically different from zero and that X has a statistically significant positive effect on Y. β is estimated to be 1.001.

Suppose that this model is indeed a very good approximation of the real world relation between X and Y, and that the linear estimation is suitable. Furthermore, among alternative specifications, this model is the one that gives the highest R-squared.

The above result is a useful tool for decision-making purposes: It links the economic payoffs Y to the choice variable X. One can now use this relation to predict one's payoffs or to select their X and to obtain desired levels of Y. More importantly, the above model links Y and X correctly. This is crucial because increasing X is costly and knowing this true relationship helps individuals make more accurate decisions.

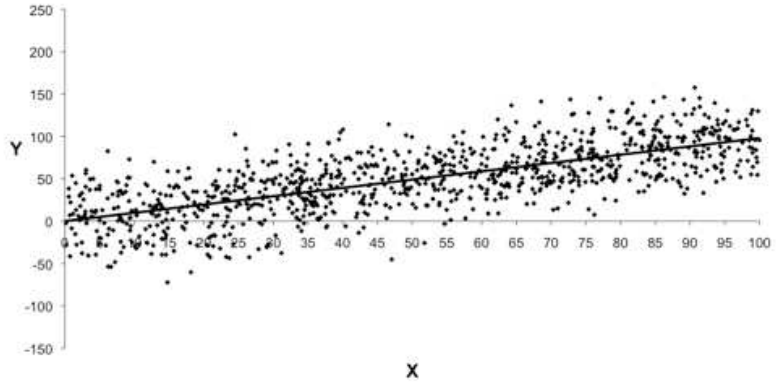
Figure 1.1. Presentation of Condition 1. This mimics the methodology of 60% of the publications that were surveyed and the suggestions of Wooldridge (2008).

Variable	Mean	Std. Dev.
X	49.51	28.74
Y	51.22	59.25

	Dependent Variable: *Y*
X	1.02
	(0.056)**
Constant	0.61
	(3.74)
R^2	0.25
N	1 000

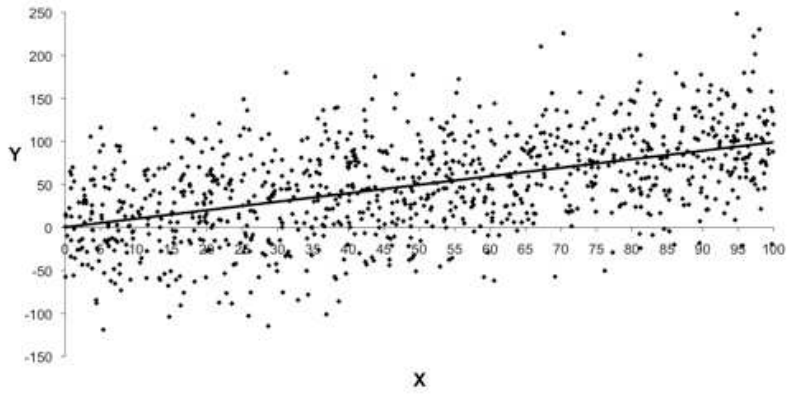
Standard errors in parentheses
** Significant at 95% confidence level
N is the number of observations

Figure 1.2. Tables in Condition 2. The rest of the presentation is the same as Figure 1.1.



The standard error of the regression ($\hat{\sigma}_\varepsilon$) is 29.

Figure 1.3. Bivariate scatter plot of Condition 1 and information on *SER*. Both were provided to participants along with estimation results in Condition 3. Only the graph was provided in Condition 5.



The standard error of the regression ($\hat{\sigma}_\varepsilon$) is 51.

Figure 1.4. Bivariate scatter plot of Condition 2 and information on *SER*. Both were provided to participants along with estimation results in Condition 4. Only the graph was provided in Condition 6.

Similar to our survey on current practice in section 1.2, we again limit attention to cross-sectional analyses in our experimental conditions. We are primarily concerned in determining how findings on average causal effects are used for predictions and decision making. Our variations in different conditions would not be valid for time series studies where the R^2 statistic does not provide information on model fit. It is important to add that in published papers results are also discussed verbally. These discussions, which are mostly confined to certain coefficient estimates and their statistical significance, might distract decision makers from the uncertainties about outcomes. None of our conditions involve such discussions.

b) Questions

For Conditions 1, 3, and 5, we asked the following questions:

1. What would be the minimum value of X that an individual would need to make sure that s/he obtains a positive outcome ($Y > 0$) with 95% probability?
2. What minimum, positive value of X would make sure, with 95% probability, that the individual obtains more Y than a person who has $X = 0$?
3. Given that the 95% confidence interval for β is (0.936, 1.067), if an individual has $X = 1$, what would be the probability that s/he gets $Y > 0.936$?
4. If an individual has $X = 1$, what would be the probability that s/he gets $Y > 1.001$ (i.e. the point estimate)?

The questions for Conditions 2, 4, and 6 were the same except that the confidence interval for β is (0.911, 1.130), and we ask about the probabilities of obtaining $Y > 0.911$ and $Y > 1.02$, given $X = 1$, in questions 3 and 4 respectively. All four questions are reasonable in that they seek answers to questions that would be of interest to decision makers. However, they are not the types of questions that reports in economics journals usually lead readers to pose. They therefore test a respondent's ability to reason correctly in a statistical manner given the information provided. In Appendix 1.A, we provide the rationale behind the questions and the correct answers.

c) Respondents and method

We sent web-based surveys to faculty members in economics departments at leading universities worldwide. From the top 150 departments, ranked by econometric publications between 1989 and 2005 (Baltagi, 2007, Table

3), we randomly selected 113.[3] Within each department, we randomly selected up to 36 faculty members. We ordered them alphabetically by their names and assigned Condition 1 to the first person, Condition 2 to the second person, … , Condition 6 to the sixth person, then again Condition 1 to the seventh person and so on.

We conducted the survey online by personally sending a link for the survey along with a short explanation to the professional email address of each prospective participant. In this way, we managed to keep the survey strictly anonymous. We do know the large pool of institutions to which the participants belong but have no means of identifying the individual sources of the answers. The participants answered the survey voluntarily. They had no time constraints and were allowed to use calculators or computers if they wished. We told all prospective participants that, at the completion of the research, the study along with the feedback on questions and answers would be posted on the web and that they would be notified.[4] We did not offer respondents any economic incentives for participation.

As can be seen from Table 1.2, we dispatched a total of 3,013 requests to participate. About one-fourth of potential respondents (26%) opened the survey and, we presume, looked at the set-ups and questions. About a third (or 9% of all potential respondents) actually completed the survey. The proportion of potential respondents who opened the surveys and responded was highest for Conditions 5 and 6 (40%) as opposed to the 30% and 32% in Conditions 1 and 2, and 3 and 4, respectively. The

[3] We stopped sampling universities once we had at least 30 individual responses for all questions asked. A few universities were not included in our sample because their webpages did not facilitate accessing potential respondents. This was more frequent for non-US universities. For reasons of confidentiality, we do not identify any of these universities.

[4] This was, in fact, done right after a first draft of the paper was written.

average time taken to complete the survey was also lowest for Conditions 5 and 6 (see foot of Table 1.2). We consider these outcomes again when we discuss the results below.

Table 1.2. Characteristics of respondents

Condition	1	2	3	4	5	6	Total	%'s
Requests to participate	568	531	548	510	438	418	3,013	
Requests opened	143	152	140	131	113	98	777	26
Surveys completed	45	45	49	38	36	44	257	9
Position								
Professor	17	14	19	18	17	22	107	42
Associate Professor	8	7	12	10	6	2	45	18
Assistant Professor	12	18	16	9	9	12	76	30
Lecturer	6	4	1	1	3	3	18	7
Other	2	2	1	0	1	5	11	4
Total	45	45	49	38	36	44	257	
Use of regression analysis								
Never	7	5	11	11	6	15	55	23
Some	11	16	17	10	17	13	84	36
Often	16	14	7	7	7	8	59	25
Always	5	5	8	6	6	7	37	16
Total	39	40	43	34	36	43	235	
Average minutes spent	11.6	10.3	7.4	7.5	5.7	6.5	8.1	
<Std. dev.>	<12.0>	<7.8>	<7.1>	<5.3>	<3.9>	<6.0>	<7.7>	

Table 1.2 documents characteristics of our respondents. In terms of position, a majority (59%) are at the rank of Associate Professor or higher. They also work in a wide variety of fields within the economics profession. Thirteen percent of respondents classified themselves as econometricians and more than two-thirds (77%) used regression analysis in their work (41% “often” or “always”).

1.4. Results

a) Condition 1

Respondents' answers to Condition 1 are summarized in Figure 1.5. Three answers incorporating only "I don't know", or "?" were removed from the data. For the first two questions, responses within plus or minus five of the correct amount were considered correct. For questions 3 and 4 we considered correct responses that were between plus or minus five percent of the answer. We also regarded as correct the responses of four participants who did not provide numerical estimates, but mentioned that the answer was mainly related to the error term and its variance (across all conditions there were 21 such responses). The questions and the correct answers are displayed in the titles of the histograms in Figure 1.5.

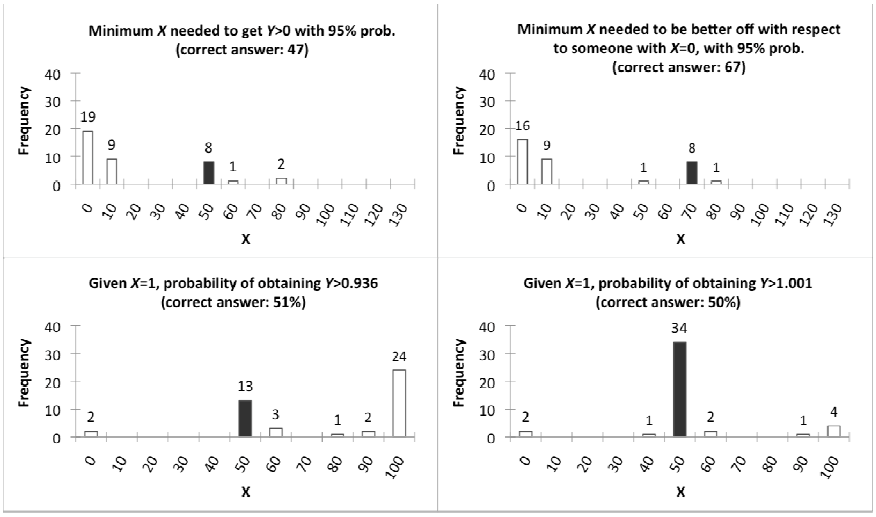


Figure 1.5. Histograms for the responses to Condition 1. The figure on top-left shows answers to Question 1, the one on top-right to Question 2, the one on bottom-left to Question 3 and the one on bottom-right to Question 4. Each histogram displays also the question and the approximate correct answer. The dark column identifies those responses that we considered as correct. Above each column are the numbers of participants who gave that particular answer. Numbers of responses were 39, 35, 45 and 44 for questions 1, 2, 3 and 4 respectively.

Most answers to the first three questions are incorrect. They suggest that the presentation directs the respondents into evaluating the results only through the coefficient estimates and obscures the uncertainty implicit in the dependent variable. Specifically, Figure 1.5 shows that:

1. 72% of the participants believe that for an individual to obtain a positive outcome with 95% probability, a small X ($X < 10$) would be enough, given the regression results. A majority state that any small positive amount of X would be sufficient to obtain a positive outcome with 95% probability. However, in order to obtain a positive outcome with 95% probability, a decision maker should choose approximately X=47.

2. 71% of the answers to the second question suggest that for an individual to be better off with 95% probability than another person with $X = 0$, a small amount of X ($X < 10$) would be sufficient. In fact, given that the person with $X = 0$ will also be subject to a random shock, the X needed to ensure this condition is approximately 67.

3. 60% of the participants suggest that given $X = 1$, the probability of obtaining an outcome that is above the lower bound of the estimated coefficient's 95% confidence interval is very high (greater than 80%). Instead, the correct probability is approximately 51%, as in this case the uncertainty around the coefficient estimates is small compared to the uncertainty due to the error term.

4. 84% of participants gave an approximately correct answer of 50% to question 4.

Participants' answers to the first two questions suggest that the uncertainty affecting *Y* is not directly visible in the presentation of the results. The answers to question 3, on the other hand, shed light on what the majority of our sample sees as the main source of fluctuation in the dependent variable. The results suggest that it is the uncertainty concerning the estimated coefficients that is seen to be important and not the magnitude of the *SER*. In the jargon of popular econometrics texts, whereas respondents were sensitive to one of the two sources of prediction error, the sampling error, they ignored the error term of the regression equation. The apparent invisibility of the random component in the presentation lures respondents into disregarding the error term and to confuse an outcome with its estimated expected value.

In their answers to questions 3 and 4, the majority of participants claim that if someone chooses $X = 1$, the probability of obtaining $Y > 1.001$ has a 50% chance, but obtaining $Y > 0.936$ is almost certain. Incidentally, the high rate of correct answers to question 4 suggests that failure to respond accurately to questions 1-3 was not because participants failed to pay attention to the task (i.e., they were not responding "randomly").

Our findings echo those of Lawrence and Makridakis (1989) who showed in an experiment that decision makers tend to construct confidence intervals of forecasts through estimated coefficients and fail to take into account correctly the randomness inherent in the process they are evaluating. Our results are also consistent with Goldstein and Taleb (2007) who have shown how failing to interpret a statistic appropriately can lead to incorrect assessments of risk.

In sum, the results of Condition 1 show that the common way of displaying results in the empirical economics literature leads to an illusion of predictability in that part of the uncertainty is invisible to the

respondents. In Condition 2, we tested this interpretation by seeing whether the answers to Condition 1 are robust to different levels of uncertainty.

b) Conditions 2 through 4

If the presentation of the results causes the error term to be ignored, then regardless of its variance, the answers of the decision makers should not change in different set-ups, provided that its expectation is zero. To test this, we change only the variance of the error term in Condition 2 – see Figure 1.2. Conditions 3 and 4 replicate Conditions 1 and 2 except that we add scatter plots and *SER* statistics – see Figures 1.3 and 1.4.

The histograms of the responses to the four questions of Conditions 2, 3, and 4 are remarkably similar to that of Condition 1 (see Appendix 1.B). These similarities are displayed in Table 1.3.

The similarity in responses between Conditions 1 and 2 shows that – under the influence of the current methodology – economists are led to overestimate the effects of explanatory factors on economic outcomes. The misperceptions in the respondents' answers suggest that the way regression results are presented in publications can blind even knowledgeable individuals from differentiating among different clouds of data points and uncertainties. At an early stage of our investigation, we also conducted the same survey (using Conditions 1 and 2) with a group of 50 graduate students in economics at Universitat Pompeu Fabra who had recently taken an advanced econometrics course as well as with 30 academic social scientists (recruited through the European Association for Decision Making). The results (not reported here) were similar to those of our sample of economists. They suggest that the origins of the misperceptions can be traced to the methodology as opposed to professional backgrounds.

Table 1.3. Comparison of results for Conditions 1 through 6

Condition	1	2	3	4	5	6
R^2	*0.50*	*0.25*	*0.50*	*0.25*	*0.50*	*0.25*
Scatter plot	*no*	*no*	*yes*	*yes*	*yes*	*yes*
Estimation results	*yes*	*yes*	*yes*	*yes*	*no*	*no*
Percentage of participants whose answer to:						
Question (1) was X < 10 *(Incorrect)*	**72**	**67**	**61**	**41**	**3**	**7**
Question (2) was X < 10 *(Incorrect)*	**71**	**70**	**67**	**47**	**3**	**15**
Question (3) was above 80% *(Incorrect)*	**60**	**64**	**63**	**32**	**9**	**7**
Question (4) was approx. 50% *(Correct)*	**84**	**88**	**76**	**84**	**91**	**93**
Approximate correct answers are						
Question 1	47	82	47	82	47	82
Question 2	67	116	67	116	67	116
Question 3 (%)	51	51	51	51	51	51
Question 4 (%)	50	50	50	50	50	50
Number of participants						
Question 1	39	36	44	32	31	41
Question 2	35	30	39	32	30	39
Question 3	45	42	49	37	32	43
Question 4	44	41	49	37	32	43

Notes:

Question 1) What would be the minimum value of X that an individual would need to make sure that s/he obtains a positive outcome ($Y > 0$) with 95% probability?

Question 2) What minimum, positive value of X would make sure, with 95% probability, that the individual obtains more Y than a person who has $X = 0$?

Question 3) Given that the 95% confidence interval for β is (a, b), if an individual has $X = 1$, what would be the probability that s/he gets $Y > a$?

Question 4) If an individual has $X = 1$, what would be the probability that s/he gets $Y > \hat{\beta}$?

Where $a = 0.936$, $b = 1.067$ and $\hat{\beta} = 1.001$ in Conditions 1, 3 and 5; and $a = 0.911$, $b = 1.13$ and $\hat{\beta} = 1.02$ inConditions 2, 4 and 6.

Table 1.3 indicates that when the representation is augmented with a graph of actual observations and with statistical information on the magnitude of the error term (*SER*), the perceptions of the relevant uncertainty and consequently the predictions improve. However, around half of the participants still fail to take the error term into account when making predictions and give similar answers to those in Conditions 1 and 2 (see Appendix 1.B for histograms of responses to Conditions 3 and 4). This suggests that respondents still mainly rely on the table showing the estimated coefficients and their standard errors as the main tool for assessing uncertainty. Since the information provided in Conditions 3 and 4 is rarely provided in published papers, this does not provide much hope for improvement. Possibly more drastic changes are necessary. Conditions 5 and 6 were designed to test this suggestion.

c) Conditions 5 and 6

Our results so far suggest that, when making predictions using regression analysis, economists pay excessive attention to coefficient estimates and their standard errors and fail to consider the uncertainty inherent in the relation between the dependent and independent variables. What happens, therefore, when they cannot see estimates of coefficients and related statistics but only have a bivariate scatter plot? This is the essence of Conditions 5 and 6 – see the graphs in Figures 1.3 and 1.4.

Figure 1.6 displays the histograms for responses to the four questions in Condition 5. The responses to Condition 6 were similar and the histograms are displayed in Appendix 1.B. These show that participants are now much more accurate in their assessments of uncertainty compared to the previous Conditions (see also Table 1.3). In fact, when the coefficient estimates are not available, they are forced to attend solely to the graph, which depicts adequately the uncertainty within the dependent

variable. This further suggests that scant attention was paid to the graphs when coefficient estimates were present. Despite the unrealistic manner of presenting the results, Conditions 5 and 6 show that a simple graph can be better suited to assessing the predictability of an outcome than a table with coefficient estimates or a presentation that includes both a graph and a table.

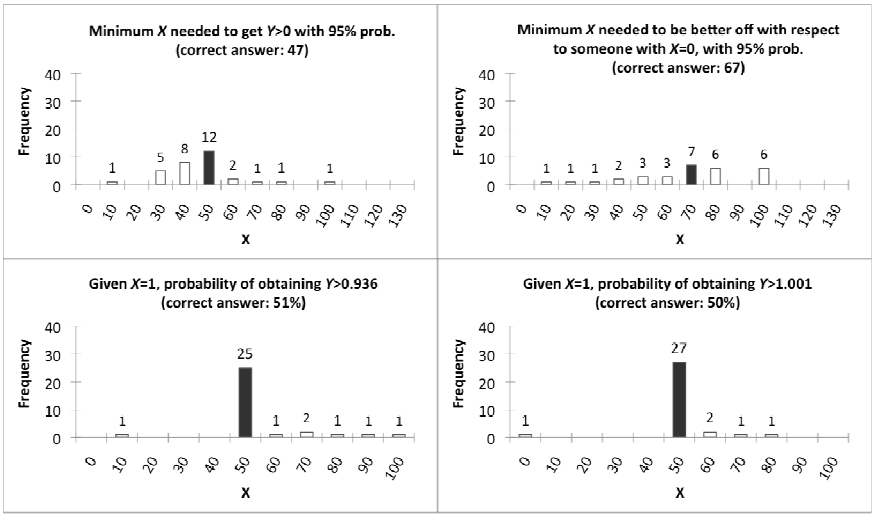


Figure 1.6. Histograms for the responses to Condition 5. The figure on top-left shows answers to Question 1, the one on top-right to Question 2, the one on bottom-left to Question 3 and the one on bottom-right to Question 4. Each histogram displays also the question and the approximate correct answer. The dark column identifies those responses that we considered as correct. Above each column are the numbers of participants who gave that particular answer. Numbers of responses were 31, 30, 32 and 32 for questions 1, 2, 3 and 4 respectively.

In Conditions 5 and 6, most of the participants, including some who made the most accurate predictions, protested in their comments about the insufficiency of information provided for the task. They claimed that, without the coefficient estimates, it was impossible to determine the answers and that all they did was to "guess" the outcomes approximately. Yet their guesses were more accurate than the predictions in the previous

conditions that resulted from careful investigation of the coefficient estimates and time-consuming computations. Indeed, as indicated in Table 1.2, respondents in Conditions 5 and 6 spent significantly less time on the task than those in Conditions 1 and 2 ($t(40) = 2.71$ and $t(40) = 2.38$, $p = 0.01$ and 0.02, respectively).

d) Effects of training and experience

Table 1.2 shows that our sample of 257 economists varied widely in terms of professorial rank and use of regression analysis in their work. We failed to find any relation between the numbers of correct answers and professorial rank or frequency of using regression analysis. A higher percentage of statisticians, financial economists and econometricians performed well relative to the average respondent (with, respectively, 64%, 56%, and 51% providing correct answers compared to the overall average of 35%). When answers were accurate, the average time spent was also slightly higher (10.2 versus 9.3 minutes). Appendix 1.C shows in detail the characteristics and proportions of respondents, who gave accurate answers in Conditions 1 through 4.

1.5. Discussion

We conducted a survey on probabilistic predictions made by economists on the basis of regression outputs similar to those published in leading economics journals. Given only the regression statistics typically reported in such journals, many respondents made inappropriate inferences. In particular, they seemed to locate the uncertainty of prediction in estimates of the regression coefficients and not in the standard error of the regression (*SER*). Indeed, responses hardly differed between cases where the fit of the estimated model varied between .25 and .50.

We also provided some respondents with scatter plots of the regression together with explicit information on the *SER*. However, this had only a small ameliorative effect and suggests that respondents relied principally on the regression statistics (e.g., coefficients and their standard errors) to make their judgments. Finally, we forced other respondents to rely on graphical representation by only providing a scatter plot and no regression statistics. Members of this group complained they had insufficient information but – most importantly – were more accurate in their responses and took less time to answer than the other groups.

Several issues can be raised about our study concerning the nature of the questions asked, the specific respondents recruited, and motivations to answer our questions. We now address these issues.

First, we deliberately asked questions that are usually not posed in journal articles because we sought to illuminate economists' appreciation of the predictability of economic relations as opposed to the assessment of "significance" of certain variables (McCloskey & Ziliak, 1996; Ziliak & McCloskey, 2004; 2008). This is important. For example, even though economics articles typically do not address explicit decision making questions, models can be used to estimate, say, the probability of reaching given levels of output for specific levels of input as well as the economic significance of the findings. It is also important to understand that a policy that achieves a significantly positive effect "on average" might still be undesirable because it leaves a large fraction of the population worse off. Hence, the questions are essential but "tricky" only in the sense that they are not what economists typically ask.

Second, as noted earlier, 26% of potential respondents took the time to open (and look at?) our survey questions and 9% answered. Does this

mean that our respondents were biased and, if so, in what direction? We clearly cannot answer this question but can state that our sample contained a substantial number of respondents (257) who represent different characteristics of academic economists. Moreover, they were relevant respondents in that they were recruited worldwide from leading departments of economics as judged by publications in econometrics (Baltagi, 2007).

Third, by maintaining anonymity in responses, we were unable to offer incentives to our respondents. However, would incentives make a difference? Clearly, without conducting a specific study we cannot say. However, the consensus from results in experimental economics is that incentives increase effort and reduce variance in responses but do not necessarily increase average accuracy (Camerer & Hogarth, 1999). We also note that when professionals are asked questions relating to their competence, there is little incentive to provide casual answers. Interestingly, our survey simulates well the circumstances under which many economists read journal articles: There are no explicit monetary incentives; readers do not wish to make additional computations or do work to fill in gaps left by the authors; and time is precious. The presentation of results is, thus, crucial.

Since our investigation speaks to how statistical results are presented in academic journals, it is important to ask what specific audience authors have in mind. The goal in the leading economics journals is scientific: to identify which variables impact some economic output and to assess the strength of the relation. Indeed, the discussion of results often involves terms such as a “strong” effect where the rhetoric reflects the size of *t*-statistics and the like. Moreover, the strength of a relation is often described only from the perspective of an average effect, e.g., that a unit

increase in an independent variable implies, on average, a δ increase in the dependent variable.

As preliminary statements of the relevance of specific economic variables, this practice is acceptable. Indeed, although authors undoubtedly want to emphasize the scientific importance of their findings, we see no evidence of deliberate attempts to mislead readers into believing that results imply more control over the dependent variable than is, in fact, the case. In addition, the papers have been reviewed by peers who are typically not shy about expressing reservations. However, from a decision making perspective, the typical form of presentation can lead to an illusion of predictability over the outcomes, given the underlying regression model. Specifically, there can be considerable variability around expectations of effects that needs to be calibrated in the interpretation of results. Thus, readers who don't "go beyond the information" given and take the trouble to calculate, say, the implications of some decision-oriented questions may gain an inaccurate view of the results obtained.

At one level, it can be argued that the principle of *caveat emptor* should apply. That is, consumers of economic research should know better how to use the information provided and it is their responsibility to assess uncertainty appropriately. It is not the fault of the authors or the journals. We make two arguments against the *caveat emptor* principle as applied here.

First, as demonstrated by our survey, even knowledgeable economists experience difficulty in going beyond the information provided in typical outputs of regression analysis. If one wants to make the argument that people "ought" to do something, then it should be also clearly demonstrated that they "can."

Second, given the vast quantities of economic reports available, it is unlikely that most readers will take the necessary steps to go beyond the information provided. As a consequence, by reading journals in economics they will necessarily acquire a false impression of what knowledge gained from economic research allows one to say. In short, they will believe that economic outputs are far more predictable than is in fact the case.

We make all of the above statements assuming that econometric models describe empirical phenomena appropriately. In reality, such models might suffer from a variety of problems associated with the omission of key variables, measurement error, multicollinearity, or estimating future values of predictors. It can only be shown that model assumptions are at best approximately satisfied (they are not "rejected" by the data). Moreover, whereas the model-data fit is maximized within the particular sample observed, there is no guarantee that the estimated relations will be maintained in other samples. Indeed, the R^2 estimated on a fitting sample inevitably "shrinks" when predicting to a new sample and it is problematic to estimate *a priori* the amount of shrinkage. There is also evidence that statistical significance is often wrongly associated with replicability (Tversky & Kahneman, 1971; see also Hubbard & Armstrong, 1994). Possibly, if authors discussed these issues further, perceptions on predictability of outcomes would improve. However, these considerations are beyond the scope of the present study.

Furthermore, because our aim was to isolate the impact of presentation mode on predictions, we made many simplifying assumptions. For instance, errors that are heteroskedastic and non-normally distributed or fewer observations at the more extreme values of the dependent variable would also increase prediction error. Even though many estimation

procedures do not require assumptions, such as normally distributed random disturbances, to obtain consistent estimates, the explanations they provide through coefficient estimates and average values would be less accurate if the law of large numbers did not hold. Hence, in more realistic scenarios, where our assumptions are not valid, decisions that are weighted towards expected values and coefficient estimates would be even less accurate than our results indicate.

How then can current practice be improved? Our results show that providing graphs alone led to the most accurate inferences. However, since this excludes the actual statistical analysis evaluating the relation between different variables, we do not deem it a practical solution. But we do believe it is appropriate to present graphs together with summary statistics as we did in Conditions 3 and 4, although this methodology does not eliminate the problem.

We seriously doubt that any substantial modification of current practice will be accepted. We therefore suggest *augmenting* reports by requiring authors to provide internet links to simulation tools. These could explore different implications of the analysis as well as let readers pose different probabilistic questions. In short, we propose providing tools that allow readers to experience the uncertainty in the outcomes of the regression.[5]

In fact, we embarked on testing the effectiveness of simulations in facilitating probabilistic inferences (Hogarth & Soyer, 2011). In two experiments conducted with participants varying in statistical

[5] For example, by following the link http://www.econ.upf.edu/~soyer/Emre_Soyer/Econometrics_Project.html the reader can investigate many questions concerning the two regression set-ups that we examined in this paper as well as experience simulated outcomes.

sophistication, respondents were provided with an interface where they sampled sequentially the outcomes predicted by an underlying model. In the first, we tested responses to seven well-known probabilistic puzzles. The second involved simulating the predictions of an estimated regression model given one's choices, in order to make investment decisions. The results of both experiments are unequivocal. Experience obtained through simulation led to far more accurate inferences than attempts at analysis. Also, participants preferred using the experiential methodology over analysis. Moreover, when aided by simulation, participants who are naïve with respect to probabilistic reasoning performed as well as those with university training in statistical inference. The results support our suggestion that authors of empirical papers supplement the outputs of their analyses with simulation models that allow decision makers to "go beyond the information given" and "experience" outcomes of the model given their inputs.

Whereas our suggestion imposes an additional burden on authors, it reduces effort and misinterpretation on the part of readers, and makes any empirical article a more accessible scientific product. Moreover, it has the potential to correct statistical misinterpretations that were not identified by our study. As such we believe our suggestion goes a long way to toward increasing understanding of economic phenomena. At the same time, our suggestion calls for additional research into understanding when and why different presentation formats lead to misinterpretation.

In addition to suggesting changes in how statistical results should be reported in journals to produce better inferences, our results also have implications for the teaching of statistical techniques. First, textbooks should provide more coverage of how to report statistical results as well as instruction in how to make probabilistic predictions. Even a cursory

examination of leading textbooks shows that the topic of reporting currently receives little attention and decision making is only considered through the construction of confidence intervals around predicted outcomes.

Together with estimating average effects, evaluating the predictive ability of economic models should become an important component of the teaching of econometrics. Indeed, if this is linked to the development and use of simulation methods, it could become a most attractive (and illuminating) part of any econometrics syllabus.

Finally, we note that scientific knowledge advances to the extent that we are able to forecast and control different phenomena. However, if we cannot make appropriate probabilistic statements about our predictions, our ability to assess our knowledge accurately is seriously compromised.

Appendix 1.A. Rationale for answers to the four questions

a) Preliminary comments

We test whether or not decision makers knowledgeable about regression analysis correctly evaluate the unpredictability of an outcome, given the standard presentation of linear regression results in an empirical study. To isolate the effects of a possible misperception, we created a basic specification. In this hypothetical situation, a continuous variable *X* causes an outcome *Y* and the effect of one more *X* is estimated to be almost exactly equal to 1. The majority of the fluctuation in *Y* is due to a random disturbance uncorrelated with *X*, which is normally and independently distributed with constant variance. Hence, the decision maker knows that all the assumptions of the classical linear regression model hold (see, e.g., Greene, 2003).

b) Answers to Questions 1 and 2

In the first two questions, participants are asked to advise a hypothetical individual who desires to have a certain level of control over the outcomes. This corresponds to the desire to obtain a certain amount of *Y* through some action *X*. The first question reflects the desire to obtain a positive outcome, whereas the second reflects the desire to be better off with respect to an alternative of no-action. If one considers only averages, the estimation results suggest that an individual should expect the relation between *X* and *Y* to be one to one. However, when could an individual claim that a certain outcome has occurred because of their actions, and not due to chance? How much does chance have to say in the realization of an outcome? The answers to these questions depend on the standard deviation of the estimated residuals (*SER*).

In a linear regression analysis, SER^2 corresponds to the variance of the dependent variable that is unexplained by the independent variables and is captured by the statistic ($1\text{-}R^2$). In Conditions 1 and 3 this is given as 50%. One can compute the *SER* using the ($1\text{-}R^2$) statistic and the variance of *Y*:

$$SER = se(\hat{e}) = \sqrt{(Var(Y)(1-R^2)} = \sqrt{(40.78^2)(0.5)} \approx 29 \quad \text{(A1)}$$

The answer to the first question can be approximately calculated by constructing a one-sided 95% confidence interval using (A1). We are looking for *X* where,

$$\text{Prob}(\, Z > -\tfrac{\hat{C}+\hat{\beta}X)}{se(\hat{e})}) = \text{Prob}(\, Z > -\tfrac{0.32+1.001X}{29}) = 0.95, \text{ where } Z \sim N(0,1) \quad \text{(A2)}$$

Thus, to obtain a positive payoff with 95% probability, an individual has to choose:

$$X = \tfrac{(1.645*29-0.32)}{1.001} \approx 47 \quad \text{(A3)}$$

The answer to the second question requires one additional calculation. Specifically, we need to know the standard deviation of the difference between two random variables, that is

$$(Y_i \mid X_i = x_i) - (Y_i \mid X_i = 0), \text{ where } x_i > 0 \quad \text{(A4)}$$

We know that ($Y_i \mid X_i$) is an identically, independently and normally distributed random error with an estimated standard deviation of again 29. Given that a different and independent shock occurs for different individuals and actions, the standard deviation of (A4) becomes:

$$\sqrt{Var[(Y_i \mid X_i = x_i) - (Y_i \mid X_i = 0)]} =$$
$$= \sqrt{Var(Y_i \mid X_i = x_i) + Var(Y_i \mid X_i = 0)} = \sqrt{29^2 + 29^2} \approx 41 \quad \text{(A5)}$$

Thus, the answer to question 2 is:

$$X = \frac{(1.645*41-0.32)}{1.001} \approx 67 \quad \text{(A6)}$$

For Condition 2 (and thus also 4 and 6), similar reasoning is involved. For these conditions, the equivalent of equation (A1) is

$$SER = se(\hat{e}) = \sqrt{(Var(Y)(1-R^2)} = \sqrt{(59.25^2)(0.75)} \approx 51 \quad \text{(A7)}$$

such that the answer to question 1 is:

$$X = \frac{(1.645*51-0.62)}{1.02} \approx 82 \quad \text{(A8)}$$

As for question 2, we need to find out about (A4) in this condition:

$$\sqrt{Var(Y_i \mid X_i = x_i) + Var(Y_i \mid X_i = 0)} = \sqrt{51^2 + 51^2} \approx 72 \quad \text{(A9)}$$

So that the answer to question 2 in Condition 2 becomes:

$$X = \frac{(1.645*72-0.62)}{1.02} \approx 116 \quad \text{(A10)}$$

c) Answers to Questions 3 and 4

Here, we inquire about how decision makers weight the different sources of uncertainty within the dependent variable. These questions provide insight as to whether or not the typical presentation of the results directs the participants into considering that the fluctuation around the estimated coefficient is a larger source of uncertainty in the realization of *Y* than it really is.

Question 3 asks about the probability of obtaining an outcome above the lower-bound of the 95% confidence interval of the estimated coefficient, given a value of *X*=1.

In Conditions 1, 3 and 5, the lower-bound is 0.936. We can find an approximate answer to this question using the estimated model and the *SER* from equation (A1),

$$\Pr(Y_i > 0.936 \mid X_i = 1) = \Pr(\widehat{C} + \widehat{\beta}X_i + \hat{e} > 0.936 \mid X_i = 1) =$$
$$= \Pr(\hat{e} > 0.936 - \widehat{C} - \widehat{\beta}X_i \mid X_i = 1) = \Pr\left(\frac{\hat{e}}{se(\hat{e})} > \frac{0.936 - \widehat{C} - \widehat{\beta}X_i}{se(\hat{e})} \mid X_i = 1\right) =$$
$$= 1 - \Phi\left(\frac{0.936 - 0.32 - 1.001}{29}\right) = 1 - \Phi(-0.013) \approx 0.51 \qquad \text{(A11)}$$

where Φ is the cumulative standard normal distribution.

Question 4 asks about the probability of obtaining an outcome above the point estimate, given a value of *X*=1. In Conditions 1, 3 and 5, the point estimate is 1.001. We can use similar calculations to in order to obtain an answer.

$$\Pr(Y_i > 1.001 \mid X_i = 1) = \Pr(\widehat{C} + \widehat{\beta}X_i + \hat{e} > 1.001 \mid X_i = 1) =$$
$$= \Pr(\hat{e} > 1.001 - \widehat{C} - \widehat{\beta}X_i \mid X_i = 1) = \Pr\left(\frac{\hat{e}}{se(\hat{e})} > \frac{1.001 - \widehat{C} - \widehat{\beta}X_i}{se(\hat{e})} \mid X_i = 1\right) =$$
$$= 1 - \Phi\left(\frac{1.001 - 0.32 - 1.001}{29}\right) = 1 - \Phi(-0.01) \approx 0.5 \qquad \text{(A12)}$$

For questions 3 and 4 of Condition 2 (and thus 4 and 6), we follow similar reasoning using the appropriate estimates. Thus, for question 3,

$$\Pr(Y_i > 0.911 \mid X_i = 1) = \Pr(\widehat{C} + \widehat{\beta}X_i + \hat{e} > 0.911 \mid X_i = 1) =$$
$$= \Pr(\hat{e} > 0.911 - \widehat{C} - \widehat{\beta}X_i \mid X_i = 1) = \Pr\left(\frac{\hat{e}}{se(\hat{e})} > \frac{0.911 - \widehat{C} - \widehat{\beta}X_i}{se(\hat{e})} \mid X_i = 1\right) =$$
$$= 1 - \Phi\left(\frac{0.911 - 0.61 - 1.02}{51}\right) = 1 - \Phi(-0.015) \approx 0.51 \qquad \text{(A13)}$$

And for question 4,

$$\Pr(Y_i > 1.02 \mid X_i = 1) = \Pr(\widehat{C} + \widehat{\beta}X_i + \hat{e} > 1.02 \mid X_i = 1) =$$

$$= \Pr(\hat{e} > 1.02 - \widehat{C} - \widehat{\beta}X_i \mid X_i = 1) = \Pr\left(\frac{\hat{e}}{se(\hat{e})} > \frac{1.02-\widehat{C}-\widehat{\beta}X_i}{se(\hat{e})} \mid X_i = 1\right) =$$

$$= 1 - \Phi\left(\frac{1.02-0.61-1.02}{51}\right) = 1 - \Phi(-0.01) \approx 0.5 \qquad \text{(A14)}$$

Appendix 1.B. Histograms for the answers to Conditions 2, 3, 4 and 6

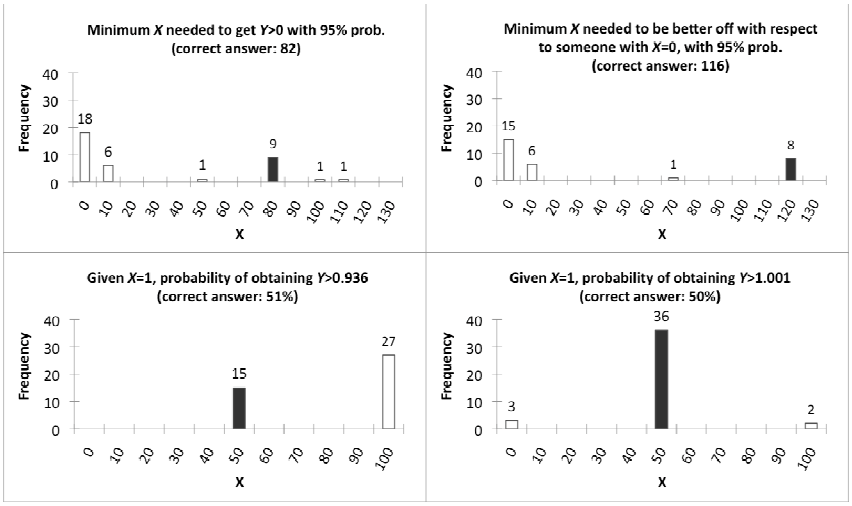


Figure 1.B1. Histograms for the responses to Condition 2. The figure on top-left shows answers to Question 1, the one on top-right to Question 2, the one on bottom-left to Question 3 and the one on bottom-right to Question 4. Each histogram displays also the question and the approximate correct answer. The dark column identifies those responses that we considered as correct. Above each column are the numbers of participants who gave that particular answer. Numbers of responses were 36, 30, 42 and 41 for questions 1, 2, 3 and 4 respectively.

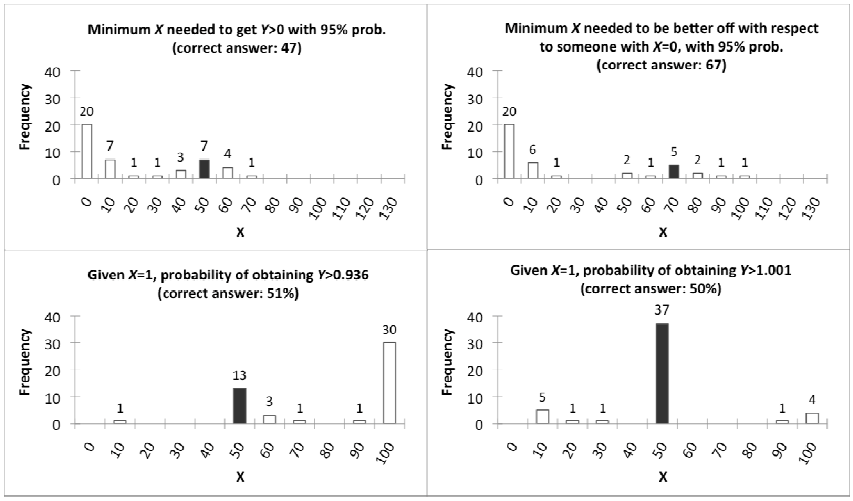


Figure 1.B2. Histograms for the responses to Condition 3. The figure on top-left shows answers to Question 1, the one on top-right to Question 2, the one on bottom-left to Question 3 and the one on bottom-right to Question 4. Each histogram displays also the question and the approximate correct answer. The dark column identifies those responses that we considered as correct. Above each column are the numbers of participants who gave that particular answer. Numbers of responses were 44, 39, 49 and 49 for questions 1, 2, 3 and 4 respectively.

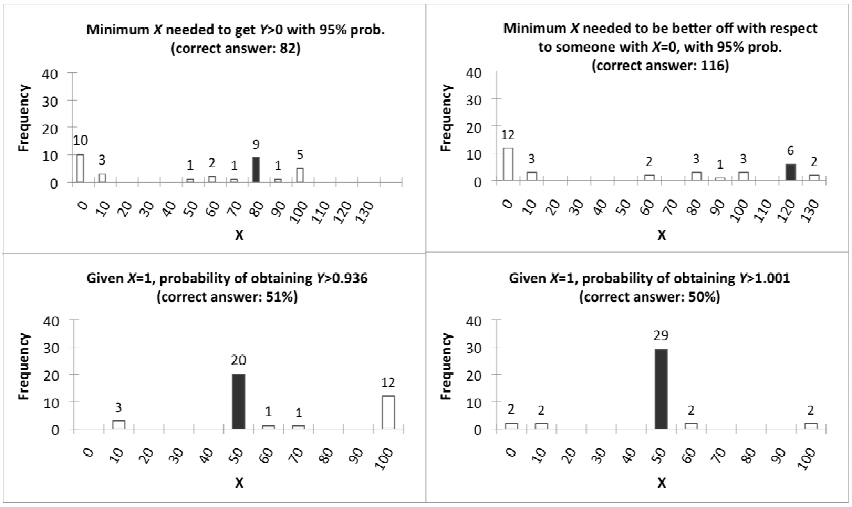


Figure 1.B3. Histograms for the responses to Condition 4. The figure on top-left shows answers to Question 1, the one on top-right to Question 2, the one on bottom-left to Question 3 and the one on bottom-right to Question 4. Each histogram displays also the question and the approximate correct answer. The dark column identifies those responses that we considered as correct. Above each column are the numbers of participants who gave that particular answer. Numbers of responses were 32, 32, 37 and 37 for questions 1, 2, 3 and 4 respectively.

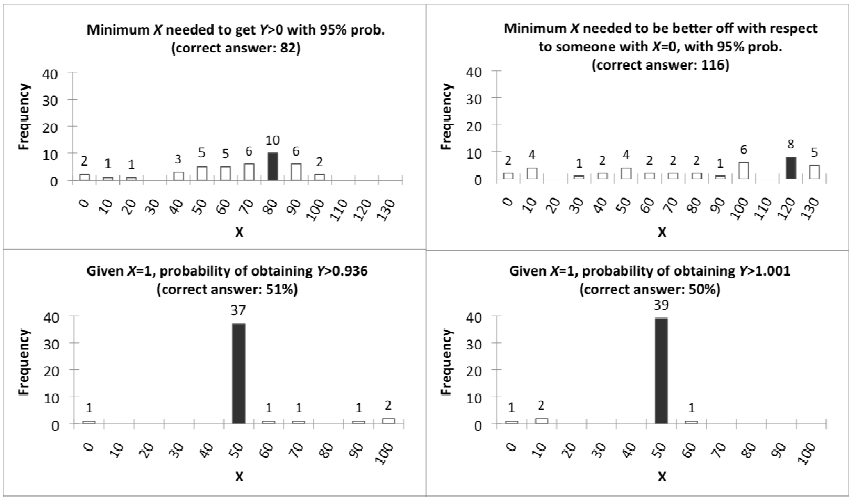


Figure 1.B4. Histograms for the responses to Condition 6. The figure on top-left shows answers to Question 1, the one on top-right to Question 2, the one on bottom-left to Question 3 and the one on bottom-right to Question 4. Each histogram displays also the question and the approximate correct answer. The dark column identifies those responses that we considered as correct. Above each column are the numbers of participants who gave that particular answer. Numbers of responses were 41, 39, 43 and 43 for questions 1, 2, 3 and 4 respectively.

Appendix 1.C. Relations between training, experience and responses in Conditions 1 to 4

(Number of respondents with correct answers in parentheses)

Condition	1	2	3	4	Total over four conditions	Percentage of respondents with correct answers
Position						
Professor	17 (4)	14 (5)	19 (6)	18 (11)	68 (26)	38
Associate Professor	8 (2)	7 (3)	12 (4)	10 (8)	37 (17)	46
Assistant Professor	12 (5)	18 (4)	16 (6)	9 (2)	55 (17)	31
Senior Lecturer	0 (0)	2 (1)	1 (0)	0 (0)	3 (1)	33
Lecturer	6 (1)	4 (0)	1 (0)	0 (0)	12 (1)	8
Post-Doctoral Researcher	2 (0)	0 (0)	0 (0)	0 (0)	2 (0)	0
Total	45 (12)	45 (13)	49 (13)	38 (21)	177 (62)	35
Research fields						
Econometrics	14 (6)	11 (6)	10 (5)	14 (8)	49 (25)	51
Labor economics	12 (5)	11 (2)	14 (3)	10 (7)	47 (17)	36
Monetary economics	5 (1)	2 (0)	5 (2)	2 (0)	14 (3)	21
Financial economics	4 (1)	5 (3)	4 (3)	3 (2)	16 (9)	56
Behavioral economics	3 (1)	7 (2)	2 (1)	3 (0)	15 (4)	27
Developmental economics	8 (1)	2 (1)	9 (3)	5 (1)	24 (6)	25
Health economics	4 (0)	3 (0)	5 (1)	1 (1)	13 (2)	15
Political economy	3 (1)	5 (1)	7 (3)	4 (2)	19 (7)	37
Public economics	9 (1)	6 (1)	10 (4)	8 (6)	33 (12)	36
Environmental economics	1 (0)	2 (1)	3 (0)	2 (1)	8 (2)	25
Industrial organization	2 (1)	6 (1)	6 (1)	2 (1)	16 (3)	19
Game theory	4 (1)	1 (1)	4 (1)	5 (2)	14 (5)	36
International economics	6 (2)	6 (0)	7 (1)	2 (1)	21 (4)	19
Macroeconomics	9 (2)	9 (2)	13 (2)	6 (5)	37 (11)	30
Microeconomics	11 (2)	4 (2)	11 (5)	7 (4)	33 (13)	39
Economic history	2 (0)	2 (0)	6 (3)	2 (1)	12 (4)	33
Statistics	3 (1)	4 (4)	1 (1)	1 (1)	11 (7)	64
Other	0 (0)	0 (0)	1 (1)	0 (0)	1 (1)	100

Appendix 1.C. continued…

Condition	1	2	3	4	Total over four conditions	Percentage of respondents with correct answers
Use of regression analysis						
Never	7 (1)	5 (0)	11 (7)	11 (5)	34 (13)	38
Some	11 (4)	16 (6)	17 (0)	10 (5)	54 (15)	28
Often	16 (4)	14 (5)	7 (2)	7 (6)	44 (17)	39
Always	5 (3)	5 (1)	8 (4)	6 (2)	24 (10)	42
Total	39 (12)	40 (12)	43 (13)	34 (18)	156 (55)	35
Average minutes spent	12 (10.9)	10.6 (12.6)	7.4 (11.2)	7.5 (7.4)	8.1 (10.2)	8.1
Std. dev.	12 (9.4)	7.8 (9)	7.1 (12.3)	5.3 (5.2)	7.7 (9)	7.7

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience vs. non-transparent description. *Journal of Experimental Psychology: General*, 140, 434-463.

2. SEQUENTIALLY SIMULATED OUTCOMES: KIND EXPERIENCEVS. NON-TRANSPARENT DESCRIPTION

(Based on Hogarth & Soyer, 2011)

2.1. Introduction

Recently, research on risky decision making has drawn attention to the fact that, in many naturally occurring situations, people do not have access to synthetic descriptions of probabilistic information that are characteristic of the experimental literature (Hertwig, Barron, Weber, & Erev, 2004; Weber, Shafir, & Blais, 2004). For example, imagine a motorist who is considering whether to exceed the speed limit on a particular highway. Lacking an externally provided estimate of the probability of detection (i.e., description), she would necessarily base her decision on what had happened in the same or similar situations in the past (i.e., experience). That is, probabilistic information about possible outcomes of decisions is often acquired through a process of sequential sampling.

Most research comparing decisions based on description as opposed to experience has naturally centered on when and why decisions differ between these two modes. The main finding concerns low-probability events. Specifically, whereas Kahneman and Tversky's (1979) influential description-based prospect theory predicts decisions consistent with the overweighting of small probabilities, decisions based on experience are consistent with underweighting (Hertwig et al., 2004; but see also Fox & Hadar, 2006).

In an important extension of this paradigm, Lejarraga (2010) asked when people might actually prefer and/or be better served to make choices after experience as opposed to description. He investigated two types of

situation. The first involved temporal gaps between the times of acquiring information and deciding. Here judgments based on experience were found to be more accurate than those based on description, a result Lejarraga attributed to differential degradation in memory of the two types of information. Second, he noted that descriptions of probabilistic information can vary in complexity. Thus, if faced with descriptions that are difficult to interpret, people might prefer to sample outcomes as opposed to drawing inferences from description (i.e., to prefer experience over description). In one experiment, he manipulated the complexity of description by varying the number of events used to define the relevant probability, that is, as a single event (simple), a function of two events (more complex), or a function of three events (even more complex). Results showed that as complexity increases, so does preference for experience over description.

The goal of the present chapter is to investigate when judgments of probability based on experience are more accurate than those based on description and to suggest theoretical and practical implications. To do so, we first define what we mean by "description" and "experience" and specify relevant psychological dimensions on which they can be characterized.

In Hertwig et al. (2004), description was made operational by providing experimental participants with the specific probabilities associated with the outcomes of the choices they faced (e.g., a sure gain of $4 versus a 0.80 chance of winning $5). As noted above, in his complexity manipulation Lejarraga (2010) presented probability information in a format involving two or more uncertain events but always such that the probabilities of specific outcomes could be calculated. In other words, description can be defined as providing all the information necessary to

specify the probabilities of relevant events even though the actual values might not be stated (i.e., all the information is present for rational calculation be it simple multiplication or addition or more complex operations such as required by Bayes' theorem).

From a strictly rational perspective, this definition implies that all descriptions are equivalent. However, this is not the case psychologically (Einhorn & Hogarth, 1981) and raises the issue of how to characterize description. In this work, we say that descriptions vary on the dimension of *transparency* (Tversky & Kahneman, 1986). Thus, for example, the description of a problem in terms of a single probability affecting the outcomes would be transparent to most people (e.g., the gamble in the preceding paragraph). However, if this probability had to be inferred from, say, the conjunction of several events (see, e.g., Lejarraga, 2010), the problem would be less transparent. Transparency, therefore, depends on both objective characteristics of the problem description and those of the decision maker. Thus, whereas a complex version of a problem might not be transparent to somebody with a low level of statistical sophistication, it could be transparent to an expert in probability theory.

The term experience also covers many variations (Shanks, 1991). The key notion is that, across time, a person observes sequences of outcomes that can be used to infer characteristics of the underlying data generating process. Examples can therefore vary from tightly controlled associative learning tasks in a laboratory to observations of actions and outcomes in naturally occurring settings. Moreover, the outcomes observed can be generated with or without the person's intervention and, importantly, may or may not provide accurate information as to the relevant characteristics of the underlying process.

In discussing the conditions that affect learning from experience, Hogarth (2001) distinguished between what he calls *kind* and *wicked* learning environments. In kind environments, people receive accurate and complete feedback that correctly represents the situation they face and thereby enables appropriate learning. Thus, observing outcomes in a kind environment typically leads people to reach unbiased estimates of characteristics of the process. In contrast, feedback in wicked environments is incomplete or missing, or systematically biased, and does not enable the learner to acquire an accurate representation.

Given these characterizations of description and experience, it is possible to depict task environments as varying in a "transparency x kindness" space as shown in Figure 2.1.Thus, the choices in the study by Hertwig et al. (2004) would be located in the lower left-hand corner since they were transparent on description and kind on experience. Although kind, the complex stimuli of Lejarraga's (2010) were not so transparent and would therefore be placed more to the right in the figure. Presumably, people would be better off trusting description in the higher left-hand section of the figure (transparent and wicked), but it is not clear what to predict for situations that are both wicked and non-transparent.

In this work, we deliberately explore situations that are kind but lack transparency (i.e., the lower right-hand area of the figure). There are two reasons. First, in naturally occurring situations, many important problems lack transparency. Second, we wish to explore the extent to which kind experience – in the form of sequentially simulated outcomes – can overcome lack of transparency. This issue lies at the heart of this chapter that is organized as follows.

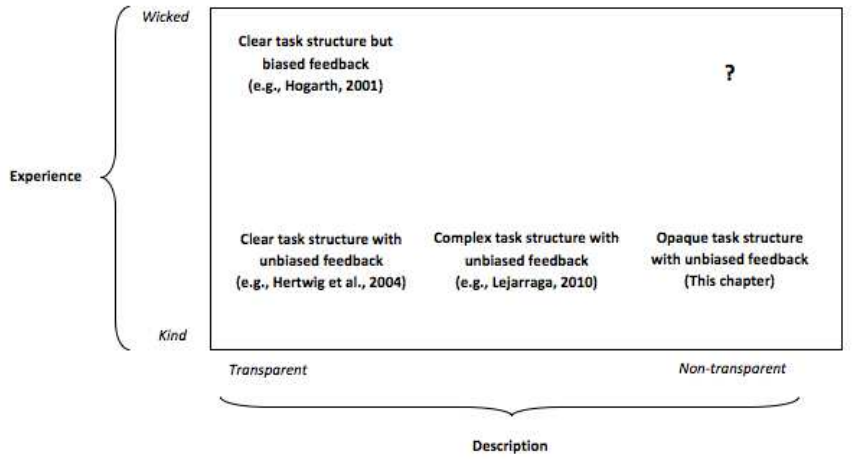


Figure 2.1. Characterizations of decision tasks by description and experience. Task structures can vary on description (the horizontal axis) from transparent to non-transparent (i.e., from clear to opaque). Experience (on the vertical axis) can vary from kind to wicked (i.e., with feedback that varies from being unbiased to biased).

We first review literature that discusses when the presentation of sequentially sampled information impacts the accuracy of judgments of probability. We next present Experiment 1 in which, using seven well-known probability problems, we contrast estimates made after description and experience. The descriptions we provide follow closely those used in the literature and would usually be categorized as non-transparent. In the experience condition, participants face a kind environment that is made operational by a simulation model that allows them to sample – and thus experience – outcomes of the relevant probabilistic process. After making estimates based on both description and experience, participants are required to provide a final response. In short, results show that estimates based on simulated experience are more accurate than those based on description and that, for their final responses, participants express a preference for experience over description. Moreover, these results hold across participants with different levels of statistical sophistication.

Whereas statistical experts might consider descriptions of some problems in Experiment 1 to be transparent, this is not the case of an investment scenario used in Experiment 2 where we extend testing the value of experience in a complex (i.e., non-transparent) situation. Once again, we find that judgments based on simulated experience are more accurate than those based on description and that there is no effect of statistical sophistication on the accuracy of experience-based responses. We conclude by discussing practical and theoretical issues raised by our work.

2.2. Frequency data and probabilistic reasoning

a) Transparency of probabilistic information

In an extensive review of issues of risk perception and communication in the medical domain, Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, and Woloshin (2007)note that the presentation of statistical information has large, and often predictable effects on the inferences people draw. For example, people are impacted far more by descriptions of risk reduction – due, say, to some intervention or treatment – when this is expressed in relative as opposed to absolute terms, e.g., as 50% instead of from 2 in 1,000 to 1 in 1,000. Similarly, in interpreting test results (e.g., mammograms), physicians' probabilistic judgments are more accurate when data are presented in *naturalfrequency* format as opposed to typical probabilistic descriptions (Gigerenzer & Hoffrage, 1995; Cosmides & Tooby, 1966; Hoffrage & Gigerenzer, 1998; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000; Brase, 2008). Indeed, frequency representations have also been observed to improve inferences in the Linda problem (Tversky & Kahneman, 1983; Fiedler, 1988; Hertwig & Gigerenzer, 1999), sample size tasks (Sedlmeier, 1998), and the Monty Hall problem (Krauss & Wang, 2003). In summarizing these and other studies, Gigerenzer et al. (2007) state

> …statistical literacy is largely a function of the outside world and …can be fostered by education and, *even more simply, by representing numbers in ways that are transparent to the human mind* (p. 54, italics added).

The argument for natural frequencies is based on the importance of experience. Specifically, Gigerenzer and Hoffrage (1995, p. 686) define natural frequencies as being data "actually experienced in a series of events" noting that "from animals to neural networks, systems seem to learn about contingencies through sequential encoding and updating of event frequencies…". In addition, they define *natural sampling* as involving the "sequential acquisition of information by updating event frequencies *without* artificially fixing the marginal frequencies" (p. 686). Paradoxically, participants in their experiments never actually experienced data *sequentially*, that is, "as a series of events." Instead, they observed totals. That is, Gigerenzer and his colleagues presented data in the form of *summarized* natural frequencies (see also Edgell, Harbison, Neace, Nahinsky, & Lajoie, 2004, p. 214).

The importance of this comment lies in the fact that experience is typically not just in the form of summed frequencies. Instead, frequencies are characterized by being experienced sequentially – that is, one-at-a-time. The foraging animal, for example, does not consult a table of data in a natural frequency format when deciding which of two potential sites has produced more food. Instead, over time it has accumulated experience – either directly or by observation – about the two sources. Moreover, numerous studies conducted with animals have shown appropriate sensitivity to environmental probabilities and, for the most part, "rational" behavior (see, e.g., Real, 1991; 1996; Weber et al., 2004).

b) Kind and wicked environments

Encoding frequency information is central to human learning and is largely an automatic process (Holyoak & Spellman, 1993). The literature has been summarized by, amongst others, Hasher and Zacks (1979; 1984) and Zacks and Hasher (2002). As their studies show, humans have a remarkable capacity for the accurate encoding of frequency information. Moreover, this cognitive activity demands little by way of attention, does not require intention, is invariant to learning, age, and many individual differences, and also involves recognizing the frequencies of subcategories of experienced events. That it is a basic cognitive mechanism that was probably developed through evolutionary pressures is reinforced by the findings that several non-human species show similar capacities.

Several studies have explored how exposure to frequency information (i.e., experience) affects the accuracy of probabilistic judgments. These show that exposure does lead to accurate judgments but that accuracy is limited to the actual stimulus-outcome relations observed. For example, Christensen-Szalanski and Beach (1982) investigated whether sequentially observing 100 instances of either base-rate or base-rate and diagnostic information would impact subsequent assessments of Bayesian posterior probabilities. They found no effect for base-rate information alone, but a favorable impact for base-rate and diagnostic information. In a further study, Betsch, Biel, Eddelbüttel, and Mock (1998) showed that, when people explicitly sampled frequency information, they were more appropriately sensitive to base rates in a Bayesian updating task than if provided with conventional probabilistic task descriptions (see also, Koehler, 1996). In a related investigation, Sedlmeier (1999, Chs. 10, 11) reported accurate probabilistic inferences when participants observed data

dynamically using a “flexible urn” in the shape of a computer simulation model. Similarly, Fiedler and Unkelbach (2011) have also considered effects of experiencing data in the spatial domain.

On the other hand, it is important to emphasize that people’s untutored skills seem to be limited to the data and relations that they actually observe. They do not necessarily imply the ability to make accurately other probabilistic judgments that could be inferred from the same observations. For example, using a medical decision making task, Edgell et al. (2004) established that although people could learn the forward conditional probabilities of the data they had observed, they were deficient when it came to assessing the corresponding inverse conditional probabilities and tended to substitute the former for the latter. However, when trained on joint probabilities they were able to overcome this tendency. Similar results have been reported by Cobos, Almaraz, and García-Madruga (2003) using an associative learning paradigm to investigate biases in probabilistic judgments. They showed, inter alia, how learning a conditional probability in one direction induced conjunction fallacies (Tversky & Kahneman, 1983) when the conditional probability required was in the other direction (i.e., the inverse probability). Likewise, Nilsson (2008) found that people were able to avoid conjunction fallacies when experience was in terms of joint probabilities but not when it involved separate experiences of the marginal components. He attributed the inferential errors to incorrect combination rules.

The work of Fiedler and his colleagues has also emphasized that people base their judgments on the data they actually observe (Fiedler, 2000; Fiedler, Brinkmann, Betsch, & Wild, 2000; Fiedler & Juslin, 2006). These authors blame inferential errors on difficulties people face in

understanding the sampling processes they encounter as opposed to the encoding of observations. They explicitly state

> ….the central empirical message of the reported studies … is that the inductive operation of quantifying the occurrence rate of a focal event in a sample is largely unimpaired and rather accurate. Judgment biases do not arise during this quantification process within available samples, but only because judges lack the necessary metacognitive skills to detect and correct for the biases that are already inherent in the available samples. (Fiedler et al., 2000, p. 412).

It is not true, of course, that people lack all metacognitive skills, just that these are limited (see, e.g., Elwin,Juslin, Olsson, &Enkvist, 2007).

A further limitation is the failure to gain insights that allow generalizing beyond the actual characteristics of the data experienced. For example, after repeated experience with the Monty Hall game (made operational by card or computer games) people do learn to take the optimal decision (Granberg & Brown, 1995; Friedman, 1998; Granberg & Dorr, 1998). However, there is no evidence that experience with an analogous game leads to understanding the probabilities affecting outcomes (Franco-Watkins, Derks, & Dougherty, 2003).

In summary, people can encode sequentially generated frequency data accurately (e.g., actions and outcomes) but are limited in their ability to make probabilistic inferences that go beyond the data observed (e.g., to infer inverse probabilities). Thus, there should be no expectation that, by themselves, kind environments can teach people to make all inferences that are logically implied by the data experienced.

2.3. Simulated experience

Since there are many ways in which experience can be made operational, we define here the method used in this paper. In fact, we created specific simulation tools for each problem in the two experiments. However, from the participants' perspective, all provided the same functions. These were to simulate and observe outcomes of each process (i.e., problem) modeled, one trial at a time and for as many trials as they wished, as well as the possibility to review subsets of outcomes of past trials. To explain the simulation methodology to participants in Experiment 1, we used the example of a coin toss and Figure 2.2 shows the computer interface specifically designed for this purpose.

Participants are told that a click on the SIMULATE button corresponds to tossing a coin once (i.e., one trial) and that the associated outcome – "1" for "heads" and "0" for "tails" – appears in the last row of a column. By clicking on the button several times, a participant can observe a series of simulated coin tosses (i.e., trials) the outcomes of which are recorded in sequence in the column. Moreover, the tool lets users select subsamples of their past experience (of simulated outcomes) in order to obtain statistical summaries (count, sum and average). For example, Figure 2.2 shows a situation where the user has clicked nine times and generated nine outcomes. Here the user has also manipulated the mouse to select a subset of five outcomes (the dashed area) of which summary statistics are provided in the table.

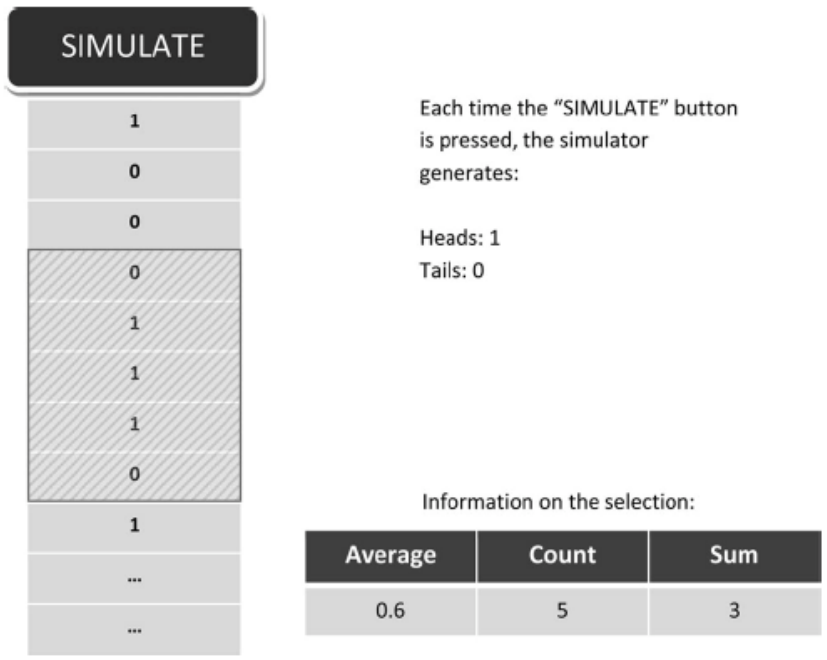


Figure 2.2. Simulation tool for coin toss. Each time the SIMULATE button is clicked, a coin toss is simulated producing a "1" (heads) or "0" (tails). The figure depicts the outcomes of nine clicks of which five have been selected (dashed area). Summary statistics of this subsample are shown in the table provided. Participants were free to sample as many outcomes as they wished and to obtain statistical summaries of subsamples they selected with the mouse.

The specific simulation interfaces designed for the problems in Experiment 1 are analogous to the coin toss simulator in their functions; they reduce the outcomes to binary values of "1" and "0", provide a new outcome with each click, make past outcomes available for visual inspection and can provide statistical summaries of subsets of previously experienced outcomes that users select. The only difference with the tool provided in Experiment 2 is that, due to the different nature of the questions, it does not show the outcomes as a string of "1"s and "0"s, but generates integer values instead.

It is important to emphasize that our simulation methodology gives participants total control over their experience (i.e., numbers of trials, evaluated subsets). It also provides a cognitive aid that participants can use to summarize subsets of their experience. Our methodology thus has some differences from that used in, for example, Hertwig et al. (2004) and this is an issue to which we will return. The simulators for each problem in the two experiments are discussed in detail in Appendix 2.B.

2.4. Experiment 1

The goals of Experiment 1 were to assess people's ability to make probabilistic judgments after gaining simulated experience on a range of problems and to observe their preferences between such experience and objective, non-transparent descriptions of the same problems.

a) Design

We varied two between-subject factors in an incomplete 2 x 2 design in which all participants gave three answers to each of seven questions. One between-subject factor was level of statistical sophistication. We compared responses of advanced undergraduate students who had taken classes in statistical reasoning and probability theory with those of a group of older, university-educated adults with less formal statistical knowledge. The second between-subject factor was whether participants first answered the questions before experiencing sequentially simulated outcomes, and then again after having done so, as opposed to the reverse, that is, first after having experienced sequentially simulated outcomes, and then without having done so. This second factor, however, was incomplete in that it was only varied for the advanced undergraduate students.

The experimental design is illustrated in Table 2.1. As shown there, one group of advanced undergraduates first answered all seven questions "analytically," i.e., without having experienced sequentially simulated outcomes. In contrast, the second group of undergraduates first answered after experiencing sequentially simulated outcomes. After each answer provided in the second task (with and without simulated outcomes, as appropriate), both groups were required to state a final answer that, if correct, would earn them €1.00 (for each correct answer). We refer to these two groups as "Sophisticated A-E" and "Sophisticated E-A," respectively.

The group of older university-educated adults was recruited through personal contacts of one of the authors. They only answered the questions in one order: before and then after having experienced sequentially simulated outcomes and also gave a final answer. Given their volunteer status, it was not deemed appropriate to remunerate them financially for either their accuracy or participation. We refer to this group as "Naïve."

Table 2.1. Design for Experiment 1

Group	**1st Task**	**2nd Task**	**3rd Task***	**Remuneration**
Sophisticated A-E	Answer analytically	*Coin toss example* Answer with experience	Final answer	1 Euro / correct final answer
Sophisticated E-A	*Coin toss example* Answer with experience	Answer analytically	Final answer	1 Euro / correct final answer
Naïve	Answer analytically	*Coin toss example* Answer with experience	Final answer	None

***** Final answers were given to each problem right after the 2nd task for that problem was completed.

Table 2.2. The seven probabilistic inference problems

1. Bayesian updating (Gigerenzer et al. 2007 version) Assume you conduct breast cancer screening using mammography in a certain region. You know the following information about the women in this region: The probability that a woman has breast cancer is 1% (prevalence) If a woman has breast cancer, the probability that she tests positive is 90% (sensitivity) If a woman does not have breast cancer, the probability that she nevertheless tests positive is 9% (false-positive rate) A woman – chosen at random – gets breast screening and the test results show that she has cancer. What is the probability that she has cancer? a) The probability that she has breast cancer is about 81%. b) Out of 10 women with a positive mammogram, about 9 have breast cancer. c) Out of 10 women with a positive mammogram, about 1 has breast cancer. d) The probability that she has breast cancer is about 1%.
2. Birthday problem In a group that has 25 people in it, what is the probability that 2 or more people have the same birthday?
3. Conjunction problem A project has 7 parts. The success of the project depends on the success of these parts. In order to be successful, all its parts need to be successful. Assume that each part is independent from the others and each has a 75% success rate. What is the probability that the project will be successful?

Table 2.2. Continued

4. Linda problem (Tversky & Kahneman, 1983)

Jessica is 31 years old, single, candid, and very promising. She graduated in philosophy. As a student, she was anxious about discrimination issues and social justice, and also took part in anti-nuclear demonstrations.

Assign a rank to the following statements from most probable to least probable:

a) Jessica works in a bookstore and takes Yoga classes.
b) Jessica is active in the feminist movement.
c) Jessica is a psychiatric social worker.
d) Jessica is a member of the League of Women Voters.
e) Jessica is a bank teller.
f) Jessica is an insurance salesperson.
g) Jessica is a bank teller and is active in the feminist movement.

5. Hospital problem (Tversky & Kahneman, 1974)

A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day. In the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are girls. However, the exact percentage varies from day to day. Sometimes it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were girls.

Which hospital do you think recorded more such days?

a) The larger hospital?
b) The smaller hospital?
c) About the same for both hospitals?

6. Regression toward the mean

A class of students enters in a TOEFL exam (it is a standardized test of English language). One of the students gets a better result than 90% of the class.

The same class, including the person who had done better than 90% of his class, enters another TOEFL exam. Past data suggest that the correlation between the scores of the different exams is about 0.8.

Which statement is correct?

a) It is more likely that the student in question now gets a better ranking.
b) It is more likely that the student in question now gets a worse ranking.
c) The chances that the student gets a better ranking or a worse one are approximately equal.

Table 2.2. Continued

7. Monty Hall problem There are three doors A, B and C. We randomly selected one of them and put a Ferrari behind it. Behind the remaining two doors there is nothing. You will select a door and we will open it. You will win the game if there is Ferrari behind it. Now select a door. (The participant makes a selection, say A). Before we open the door you selected, we open B and show you that there is nothing behind it. Now two doors remain: A and C. Behind one of them is a Ferrari. Given this situation, please state if you would like to a) Stay with your original selection b) Change to the other door

Table 2.2 provides descriptions of the seven problems used in the experiment. They were chosen because they represent a range of well-known problems that people typically answer incorrectly. Answers are provided in Appendix 2.A.

We note, parenthetically, that the initial stimuli presented to participants were the same whether they were answering the problems with or without simulated experience. Strictly speaking, therefore, Experiment 1 is a test of answers given after "description" versus "description and experience" as opposed to simply "description" versus "experience." Given the nature of our stimuli (we had to describe these for both conditions), this was unavoidable. However, this is not an issue in Experiment 2.

b) Procedure

To handle possible technical issues concerning the simulation technology, the experiment was run on an individual basis. Thus, participants made individual appointments to meet with the experimenter and were alone with him when answering questions. To facilitate presentation, we describe the procedures separately for the Sophisticated A-E and Naïve groups, on the one hand, and for the Sophisticated E-A group, on the other.

- Sophisticated A-E and Naïve

The experimenter told the participants that they would have to answer seven problems that could be solved through probabilistic reasoning. The experimenter further stressed the importance of getting the answers right. Sophisticated participants were informed that their remuneration would depend on the accuracy of their final answers.

The participants provided their age and were asked to indicate their level of comfort in probabilistic and statistical reasoning on a 5-point scale: (1) Does not know or remember anything; (2) Knows or remembers little; (3) Remembers some of the things and did well in related courses; (4) Remembers all or most of the things; (5) Expert and can teach others.

Then the participants were given written descriptions of the problems (see Table 2.2) in a language in which they were fluent (Spanish, English, or Turkish) and in a randomized order. The descriptions of each problem were also read to them out loud. Participants were given a pen, paper(s) and a calculator during this analytic task of the experiment.

Once the participants had provided answers to all the questions, they were told that now they would face the same questions again in the same order as before. However, this time, they would have a tool for each problem which would allow them to "live through" the same problem again and again, observing the outcomes sequentially. At this point the participants were referred to a coin toss, as a simple example of a probabilistic problem with two outcomes and were introduced to simulation through this example (see Figure 2.2). Participants learned how to click on the SIMULATE button, how each click results in an outcome depending on the structure of the problem, and how to use the mouse to obtain statistical summaries of subsets of previously sampled outcomes. Any questions about simulation were then clarified by the experimenter. This coin toss exercise took approximately two minutes for each participant.

Once the participants had familiarized themselves with experiencing outcomes using the coin toss example, the experimenter told them that they would now experience in a similar way the outcomes associated with the same seven problems for which they had previously provided answers. They were told that, in a metaphorical sense, they will be "tossing" groups of 25 people, TOEFL exams, and so on. The manner of sampling outcomes was left to each participant's discretion, that is, number of trials, time taken per trial, and even different numbers of samples of trials for given problems. Similar to the coin toss interface, at any time while sampling, participants were free to stop, select a subsample of past experiences using the mouse and observe the count, sum and average of the selected subset of the data in a table provided on the same screen. Occasionally, participants asked questions about the simulation mechanism and the experimenter always answered in a standard manner: "The program simulates correctly the current situation/problem and provides you with an outcome each time you click."

The instructions for each of the seven problems were similar but varied in their details and, together with descriptions of the individual simulation models, are provided in Appendix 2.B. However, by way of illustration, we provide here the instructions for the Bayesian updating task:

> This program has the same functions as the coin toss program. However, it is designed for this particular problem. With each click on the SIMULATE button you will "meet" a different woman with a positive test result. For each of them it will show (1) if she really has cancer and (0) otherwise. Hence, with each click you will see the outcome associated with one randomly selected woman. You can click as many times you wish and select and obtain statistical summaries of a subset of outcomes that you have previously sampled by selecting it with the mouse.

The participants were asked to provide two more responses for each problem; a response based on experiencing sequentially simulated outcomes, and a final response to the problem that could take all information they wanted into account. The written descriptions of the problems were again made available for inspection when the participants needed to consult them. The experimenter told Sophisticated participants that each accurate final answer would earn them 1 Euro.

When participants had provided all three answers (analytic, experience, and final) for each problem, they were paid according to their performance (if appropriate), thanked, and dismissed.

- Sophisticated E-A

The same procedure was applied to the participants in this group except that they completed the task with simulated experience first. Thus, they were told that they would have to solve seven probabilistic problems and notified about the importance of giving accurate responses on which their remuneration would depend. Then, after providing details of their age and

level of statistical sophistication (on the same 5-point scale as above), they were directly introduced to the coin toss example and asked to experience each of the seven problems using their specific simulators and to provide answers given their experiences. Before experiencing sequentially simulated outcomes for each problem, a written description of the question was provided in a language in which the participants were fluent (mainly Spanish), which was also read out loud.

Once the experience task was over, the experimenter provided the participants with a pen, paper(s) and a calculator and asked them to solve each question again, this time analytically. The participants were required to provide an analytical answer and then a final answer to each problem that would take account of all the information they wanted to consider. The written descriptions of the problems were made again available for inspection when the participants needed to consult them. The experimenter told the participants that each accurate final answer would earn them 1 Euro.

When participants had provided all three answers (experience, analytic and final) for each problem, they were paid according to their performance, thanked, and dismissed.

c) Participants

Sixty-two undergraduate students were recruited from two classes at Universitat Pompeu Fabra and assigned at random to the Sophisticated A-E and Sophisticated E-A groups (31 to each). The students were in the $3^{rd}/4^{th}$ year of undergraduate studies in business and/or economics and had all taken courses in probability and statistics. When asked to indicate their level of comfort in probabilistic and statistical reasoning on the 5-point

scale described above the mean self-reported rating for both groups was 3 (*SD* = .4), "remembers some of the things and did well in related courses." The average age of the Sophisticated groups was 22 and 52% were female. The mean remuneration participants received – for the correctness of their final answers – was €4.52(*SD* = 1.5).

The mean age of the 20 university-educated adults in the Naïve group was 39 (range from 24 to 59) and 50% were female. In terms of statistical sophistication, the mean self-reported rating (using the same scale as the undergraduates above) was 2 (or "knows or remembers little," *SD* = .6).

d) Results

- Numbers of trials and time taken

The mean sizes of samples (i.e., numbers of simulated outcomes) per respondent across all the problems were almost the same for the two Sophisticated groups, 66 (*SD* = 30) and 65 (*SD* = 40) for A-E and E-A, respectively, but lower in the Naïve group, mean of 48 (*SD* = 27). The Sophisticated-Naïve difference is significant ($t(41) = 2.28$, $p = .03$).

The groups did not differ much in how long they took to answer the problems. Members of Sophisticated A-E spent on average 19.6 minutes (*SD* = 4.3) on the first task of solving the problems analytically and 23.1 minutes (*SD* = 3.9) on the second task. Members of Sophisticated E-A spent on average 25.5 minutes (*SD* = 4.7) on the first task of experiencing the outcomes through simulation and 15.4 minutes (*SD* = 3.9) on their second task. The time spent on the third task (providing a final answer) was below one minute for all participants. Only the total time spent was recorded for the Naïve group; the average was 48.5 (*SD* = 5.6). Across all groups, experimental sessions lasted, on average, 42.9 minutes per participant (*SD*= 8.0).

- Accuracy of responses

Table 2.3 provides an overview of the percentage of correct responses to the seven problems broken down by experimental conditions and groups. To simplify presentation, we refer to answers made without having experienced simulated outcomes by the term "Analytic." "Experience" refers to answers made after experiencing simulated outcomes.

Some general trends can be observed from Table 2.3. First, across all problems and groups, the percentage of correct answers after experience exceeds that of the analytic responses, and typically by a large margin. Second, with one exception, the percentage of correct final answers lies between their experience and analytic counterparts (the exception is the conjunction problem). This suggests that whereas participants were capable of interpreting their experience, they still wanted to give some weight to their analytic responses. Third, there are order effects. For some problems, more participants in Sophisticated E-A (who answered after using experience first) gave accurate analytic responses than those in Sophisticated A-E. Finally, whereas the analytic responses of the Naïve group are generally less accurate than their Sophisticated counterparts, their post-experience and final answers are quite comparable. Statistical tests supporting all the above statements are provided in Appendix 2.C.

Table 2.3. Percentages of correct answers to inferential problems by experimental conditions

	Sophisticated		Naive	Mean
	A-E	E-A		
1. Bayesian updating				
Analytic	17	42	20	27
Experience	97	97	100	98
Final	79	58	70	69
2. Birthday problem				
Analytic	3	13	0	6
Experience	55	61	65	60
Final	35	61	30	44
3. Conjunction problem				
Analytic	55	52	25	47
Experience	74	77	75	75
Final	77	77	75	77
4. Linda problem				
Analytic	10	32	10	18
Experience	97	97	90	95
Final	65	71	60	66
5. Hospital problem				
Analytic	39	61	25	44
Experience	97	97	100	98
Final	81	68	65	72
6. Regression toward mean				
Analytic	32	45	25	35
Experience	68	90	70	77
Final	55	65	35	54
7. Monty Hall				
Analytic	31	48	15	34
Experience	93	97	95	95
Final	69	58	55	61
n =	31 (29)	31	20	

Since the pattern of responses was similar across all problems, we first discuss only the Bayesian updating task in detail and then draw attention to distinguishing features of responses to other problems separately below.

Figure 2.3 reports the result of the Bayesian updating task. We display nine graphs. The three graphs in the top row report the data for the analytic responses for the three groups (from left to right, Sophisticated A-E, Sophisticated E-A, and Naïve, respectively). The middle and bottom rows show the analogous data for the experience and final responses.

The specific version of the Bayesian updating problem that we used was taken from Gigerenzer et al. (2007) and provides four options to choose from, as shown in Table 2.2. Gigerenzer et al. (2007) employed this version in a continuing education program in which 160 gynecologists were instructed how to use natural frequencies for solving Bayesian updating problems. The results of that session were quite successful. Whereas only 21% of the 160 gynecologists provided the correct answer before training, the percentage rose to 87% after training.

The comparison with our results can be seen by looking down the left-most column of graphs in Figure 2.3. Only 5 out of 29 (17%) answer correctly initially (similar to Gigerenzer's 21%). However, after experience 28 out of 29 (97%) answer correctly although this figure drops to 23 out of 29 (79%) for the final answer. In short, our results are comparable to those achieved with the natural frequency method.

Figures analogous to Figure 2.3 that provide full information on responses made in all conditions by all groups to the six remaining problems can be found in Appendix 2.D (for details on the simulators for each problem, see Appendix 2.B). Results show:

In the Birthday problem, analytic responses were skewed for all three groups toward incorrect, low values. Experience made a dramatic difference. Whereas the actual percentage correct was less than in other problems, the answers of a clear majority were close to correct (between 50% and 60%). This pattern was largely maintained in the final response by the Sophisticated groups but the Naïve group exhibited wide dispersion.

The analytic responses to the conjunction problem were somewhat dispersed. But experience made a difference that was largely maintained by all groups in their final responses.

For the Linda problem we considered only whether participants recognized that the event "bank teller and active in the feminist movement" could not be more likely than "bank teller." (Our text referred to Jessica as opposed to Linda to avoid the possibility that the Sophisticated participants might have heard of the "Linda problem"). The analytic and experience–based responses were generally opposites for all groups (incorrect and correct, respectively). The majority responses for the final answer, however, were correct.

Experience led to almost 100% correct responses for the hospital problem and the majority of final answers were also correct. For this problem there was an order effect. In the Sophisticated E-A group, there was a majority of correct analytic responses. In this case, prior experience was probably particularly relevant because no calculations were needed to answer the analytic question.

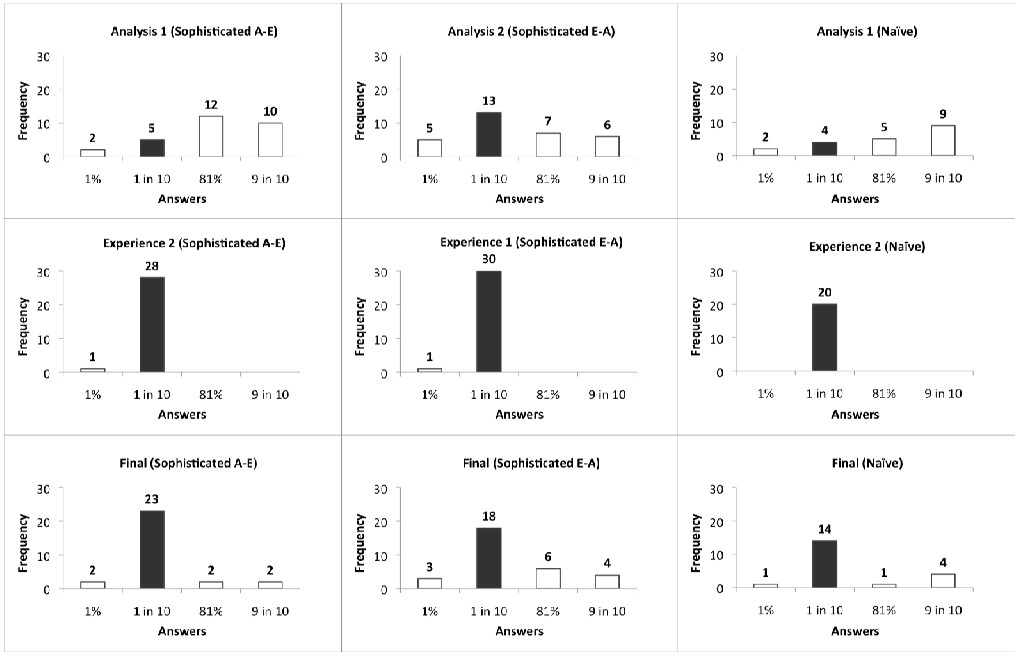


Figure 2.3. Histograms of responses given to the Bayesian updating problem. From left to right, columns represent Sophisticated A-E, Sophisticated E-A and Naïve, with sample sizes 29, 31 and 20, respectively. From top to bottom, rows represent analytical responses, responses after simulated experience, and final responses. The numbers in the columns represent the numbers of answers. The dark column identifies the correct answer.

In the regression toward the mean problem, the modal responses of all groups to the analytic question were centered on the incorrect answer of "equal," thereby suggesting that the respondents did not understand the principle behind the question. The effect of experience was to shift answers to being more correct. However, at the final stage the Naïve group is not convinced.

Experience had a big impact for the Monty Hall problem. Almost everybody chose the correct answer of "change" after experience. However, a minority regressed to the incorrect answer at the final stage.

e) Discussion

Previous research on the stimuli in Experiment 1 has shown that, when presented in the standard probabilistic format, responses to these problems are typically incorrect. And yet, when the presentation allowed experiencing sequentially simulated outcomes, responses for all questions were remarkably accurate. To this we add three points.

First, training people to participate in the simulations by using the coin toss example was easy and took little time, on average some two minutes per person. Participants related well to the task of experiencing the outcomes of simulations.

Second, whereas our participants varied on levels of statistical sophistication, the accuracy of all participants' responses benefited from experience.

Third, our participants made third (and final) answers that implied preference for answers achieved with or without the aid of simulated

experience or some combination of the two. Whereas some participants did revert toward answers made without experience, most gave more weight to those achieved through experience.

Parenthetically, we note that solving a problem analytically and through experience are two quite distinct approaches. Nonetheless, participants did not question the design of the experiment that required them to provide answers to the same problems multiple times.

Participants were not given any indications as to how many trials to sample. Sophisticated participants sampled more than the Naïve and some problems involved larger samples than others. For example, the Bayesian and birthday problems both involved the largest numbers of trials (means of approximately 80 to 90 for the Sophisticated groups) whereas the Linda problem stimulated far fewer trials (around 30 for all groups). However, in this problem, participants had to simulate multiple outcomes on each trial thereby experiencing vectors of "1's" or "0's" for each category and not just single "1's" or "0's." Thus, the task was more cognitively taxing (see Appendix 2.B for more information).

An interesting benchmark for the size of samples is the behavior of Hertwig et al.'s (2004) participants who learned the features of two alternative choice options by active sampling of experience (in a manner similar to ours, i.e., by clicking a key on a personal computer). In this study, the median sample size was 15, far less than the medians we observed of 52, 51, and 30 (for Sophisticated A-E, E-A, and Naïve, respectively). A possible reason could be that, unlike Hertwig et al., we provided participants with the means to summarize selected past experiences. Indeed, in a recent experiment we tested the effects of such a memory aid by directly contrasting probability estimates made with and

without the ability to summarize selected past outcomes (Hogarth, Mukherjee, & Soyer, 2012). Results showed little difference in mean accuracy of estimates but the lack of a memory aid was associated with greater dispersion of responses and the sampling of less trials. Using the same paradigm as Hertwig et al., Lejarraga (2010) found that more analytically oriented participants sampled more than the less analytical, a result that parallels our Sophisticated-Naïve difference.

The seven inferential problems we used as stimuli were selected for two reasons. The first was that we wanted to test our ideas on well-known problems. The second was that if our "method" were to work well across a range of problems (as opposed to within variations of the same), it would provide a stronger test of its efficacy. Indeed, the methodology was successful across a range of problems and has important implications. Specifically, in cases where it is difficult to provide transparent descriptions of problems, simulated experience can be used to foster accurate perceptions of probabilities. As specific examples, consider the birthday and Monty Hall problems.

For the Analytic task, most participants made calculations and more calculations did not lead to more correct answers. Nor were there differences for the language in which the instructions were provided. As noted before, participants in the Sophisticated E-A condition often transformed their calculations to obtain the result they had experienced in the simulation, using this as a cue to the answer. This suggests that simulated experience can play an important role in providing insights to improve the quality of analytical thinking.

In sum, the results of Experiment 1 show that people relate easily to experience in the form of sequentially simulated outcomes; they use it

well to make accurate inferences across a range of problems; and they prefer it to non-transparent description. However, despite these results, one might still argue that in many cases an alternative method should be preferred such as, say, presenting natural frequencies for Bayesian updating tasks. There would be several advantages. First, there is no need to construct a simulation model. Second, the transparency of the method provides insight into the structure of the problem. And third, the method can be applied to solve other similar problems. We therefore sought to examine the efficacy of simulated experience in a situation where it is not obvious that an alternative presentation format, such as description by natural frequencies, could be employed. This was the main goal of Experiment 2.

A further goal of Experiment 2 was to assess more accurately the difference between the accuracy of inferences based on description alone and experience alone since, as noted above, Experiment 1 only estimated the additional effect of experience over description.

2.5. Experiment 2

Experiment 2 involves a decision making situation where a transparent description of the problem cannot be easily constructed.

a) Design

The design of Experiment 2 involved between-subject comparisons of responses of two conditions that were required to answer questions based either on the analytical description of a problem involving regression analysis or after gaining experience with a simulation tool. We label the conditions as Analytic and Experience, respectively, except that there were two subgroups in the Experience condition. One involved

statistically sophisticated, graduate students in economics whom we label Sophisticated, and who were similar to respondents in the Analytic condition. The other was comprised of university-educated adults without advanced statistical knowledge whom we refer to as Naïve (specifically, these participants did not know what "regression analysis" is). We therefore make comparisons between three subgroups: Analytic (the only group in the Analytic condition), Sophisticated, and Naïve.

Table 2.4. Experiment 2 Analytic group set-up

Thank you for participating in this experiment. It is anonymous, please do not write you name.

Here you will be asked to make an investment decision. You are given 40 credits. You can allocate these 40 credits in 3 ways:

I_1 : You can invest some in "Investment 1"
I_2 : You can invest some in "Investment 2"
N : You can choose not to invest some of it.

You can choose how much to put in each of these 3 options, provided that your choices add up to 40.The relationship between the investments and their effect on the outcome is given by the following linear equation:

$$\Delta Y_i = \alpha + \beta_1 \mathrm{I}_{1,i} + \beta_2 \mathrm{I}_{2,i} + e_i$$

Where "ΔY" is the **change** in resulting credits, "I_1" is the amount invested in investment 1, "I_2" is the amount invested in investment 2, "β_1" and "β_2" are the effects of investments on the change in credits and "*e*" is the random perturbation.

The return to each investment is estimated through historical data. Past 1000 investments were taken into account for each investment and an OLS regression was conducted to compute the relationship between each investment and its return

The sample statistics for the data are as follows:

Variable	Mean	Std. Dev.
ΔY	8.4	7.9
I_1	11.1	5.8
I_2	9.6	5.2

The OLS estimation results are as follows:

	Dependent Variable: ΔY	
I_1	0.5	(0.20)**
I_2	0.3	(0.05)**
Constant	-0.1	(0.15)
R^2	0.21	
n	1 000	

Standard errors in parentheses

** Significant at 95% confidence level

n is the number of observations

This means that both the investments are estimated to have positive and significant effects on the change in one's returns. Specifically, in the average, "Investment 1" is expected to generate a 50% increase and "Investment 2" is expected to generate a 30% increase over the invested amount.

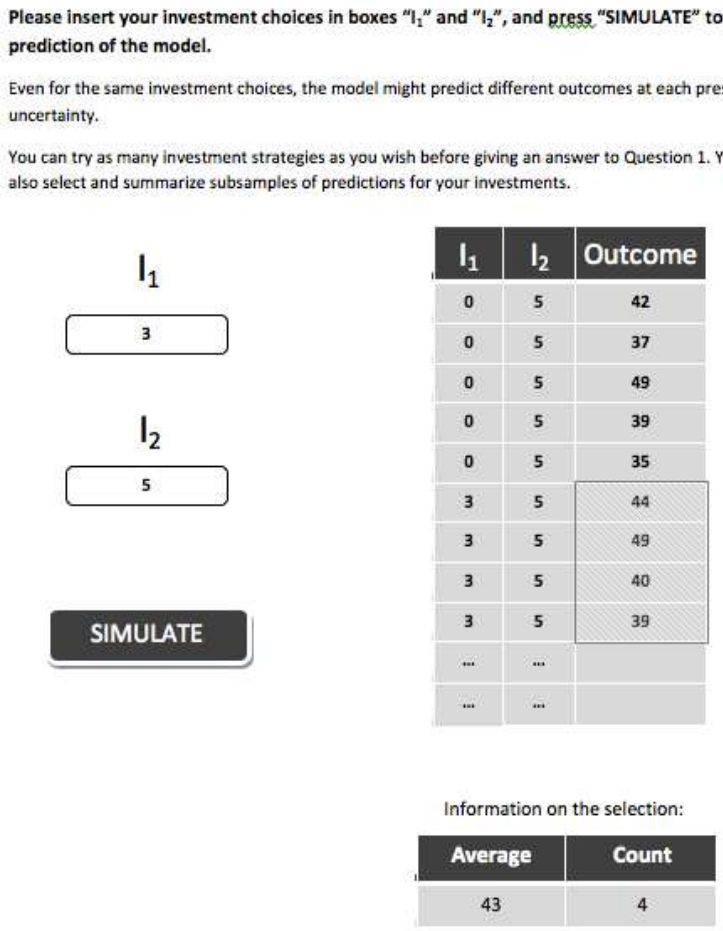


Figure 2.4. Experiment 2 Experience condition set-up. Functions are similar to the coin toss simulation shown in Figure 2. Each time the SIMULATE button is clicked, a predicted outcome is shown based on both the user's inputs and the parameters of the model. Participants in the experiment were free to sample as many outcomes as they wished and to obtain statistical summaries of subsamples they selected with the mouse.

b) Problem set-up

Table 2.4 provides the wording of the problem set-up for participants in the Analytic condition. As can be seen, the problem involves an investment situation, which requires allocating funds (40 credits) across three alternatives: “Investment 1”, “Investment 2”, and “no investment.” The profitability of the two investment opportunities is described by a regression model. The specific questions were:

1. How would you allocate your 40 credits in order to expect an increase of 5 credits (to expect to obtain 45 credits)? How much of 40 credits in Investment 1, how much in Investment 2, how much in N (no-investment)?

2. Given your investment decision in (1), what would you say is the probability of your obtaining a final total credit amount that is below 40 ($Y < 40$), i.e. less than what you started with?

3. Given your investment decision in (1), what would you say is the probability of your obtaining a final total credit amount that is below 45 ($Y < 45$)?

4. Given your investment decision in (1), what would you say is the probability that you will get a larger outcome with respect to a person who does not invest in Investments 1 and 2 (someone with $N = 40$)?

The statistical rationales for the answers are provided in Appendix 2.E.

c) Procedure

The experimental procedures differed necessarily between the Analytic and Experience conditions. For the former, the instructions sheet (see Table 2.4) and questions (see above) were given to 35 randomly selected graduate students in economics at Universitat Pompeu Fabra. When distributing the materials, the experimenter offered a bar of chocolate to each participant who was asked to return the answers, if possible within a few days, to a sealed mailbox in front of a University office. The participants were told that they could use any resources (e.g., calculators, books) they wished when answering the questions. The experimenter checked the mailbox everyday for two weeks and collected 26 responses in that period. No responses were received beyond ten days after the request to participate.

For the Experience group, participants in the Sophisticated subgroup were recruited in the same manner as those in the Analytic group (i.e., from graduate students in economics at Universitat Pompeu Fabra). For the Naïve subgroup, the experimenter contacted acquaintances (and acquaintances of acquaintances) outside Universitat Pompeu Fabra, all of whom had university degrees other than in economics. None of the Naïve subgroup participants were knowledgeable about regression analysis, whereas when asked at the end of experiment, the Sophisticated participants were all able to describe it correctly.

When conducting the experiment, the experimenter sat down one-by-one with the participants in the two Experience subgroups. These participants were given the description of the problem without any sample statistics or coefficient estimates. They were told both in writing and verbally that they possessed 40 credits and could choose to invest part of it in

Investment 1 and/or Investment 2. They were then introduced to the simulation tool (see Figure 2.4).

The experimenter explained briefly how the tool works, and then asked them to choose an investment plan so that they could expect to increase their 40 credits to 45 (the same as question 1 above). By using a tool that simulates the estimated model, we allowed them to enter a choice option, experience as many outcomes as they wished given that choice, and to repeat this for as many choice options as they wanted. Once they had made their decisions, we asked them to answer questions 2, 3, and 4 above (i.e., conditional on their investment allocations). Once again, we allowed them to use the simulation tool and they could experience the outcomes of their choices as many times as they desired. As in Experiment 1, we again made sure that participants could see all their choices as well as the outcome histories related to those choices and calculate and compare counts and averages of past outcomes. During the task, only information on the functions of the simulator was clarified by the experimenter. Appendix 2.B provides further details on the instructions provided to participants and the simulation tool.

d) Participants

The Analytic group consisted of 26 graduate students in economics at Universitat Pompeu Fabra in Barcelona who had taken at least one graduate course in econometrics and were knowledgeable about linear regression analysis. Participation was voluntary and anonymous. The average age was 25 and 30% were female. Of 35 surveys distributed, 26 were completed.

The Sophisticated participants in the Experience condition consisted of 28 graduate students in economics drawn from the same population as the

Analytic group. The Naïve subgroup consisted of 18 members of the general public having university degrees but no knowledge of regression analysis. Their mean age was 35 (range from 23 to 60) and 40% were female.

e) Results

Table 2.5 documents the means (standard deviations) of different variables –the decisions taken, and answers to the required probabilistic inferences– for the different experimental conditions. The first two rows of the table (labeled I_1 and I_2) indicate the mean amounts invested in Investments 1 and 2, respectively, by the different subgroups. According to the regression results, these two investments differ in the expected level and variability of their returns – Investment 1 having both greater expected return and more variability than Investment 2. On average, the Analytic participants adopt less risky strategies than their Sophisticated counterparts but all three subgroups select investment strategies that essentially meet the demands of the first question, that is, to achieve an expected target of 45.

Question 2 asks for the probability that the investment strategies will lead to outcomes of less than 40 (i.e., the starting amounts). The accuracy of each participant's response can be assessed by calculating the absolute values of the difference between the response and its normative counterpart (i.e., the correct response implied by the regression analysis). Using this measure, we note that respondents in the Analytic group seriously underestimate the probability that *Y* is less than 40 and their inferences are less accurate than those made in the two Experience groups (the difference between the Analytic and Experience conditions is significant ($t(48) = 5.4$, $p < .01$).

Table 2.5. Means <*SD*s> for conditions in Experiment 2

Condition:	Analytic	Experience	
		Sophisticated	Naive
(n=	26	28	18)
Decisions			
I_1	3.5	5.7	6.7
	<4.6>	<3.9>	<5.9>
I_2	12.3	7.8	9.8
	<7.0>	<4.5>	<7.7>
Expected outcome			
Y	45.5	45.2	46.3
	<0.9>	<1.6>	<2.1>
Prob (*Y*<40)			
Question 2: |Response – *Correct*|	17%	8%	8%
	<7%>	<7%>	<7%>
Proportion with negative deviation (Response < *Correct*)	88%	61%	22%
Prob (*Y*<45)			
Question 3: |Response – *Correct*|	2%	8%	8%
	<5%>	<8%>	<8%>
Prob($Y\|I_1,I_2$) > Prob(Y|no investment)			
Question 4: |Response - *Correct*|	24%	11%	6%
	<10%>	<9%>	<5%>
Proportion with positive deviation (Response > *Correct*)	100%	53%	67%

Question 3 asks for the probability that the investment strategies will lead to outcomes of less than 45 (i.e., the investment target). On average,

answers to this question are all quite accurate; more so for the Analytic condition than the Experience. In fact, these responses are consistent with answers to the first question that led to expectations of, on average, about 45, that is, with a symmetric predictive distribution there is as much chance of exceeding as falling short of the target.

Question 4 asks for the probability that the chosen investment strategy will lead to outcomes superior to a strategy of no investment. The Analytic group overestimates this probability and, as in Question 2, the responses are less accurate than those of the Experience groups. Again, the difference between the two conditions is significant ($t(41) = 6.7$, $p < .01$).

f) Discussion

Unlike the inference tasks of Experiment 1, Experiment 2 required participants to choose an investment plan and make probabilistic inferences based on their own idiosyncratic decisions. Also unlike several tasks of Experiment 1, it is unclear how one could have provided alternative descriptions of the questions in the form, say, of natural frequencies. Moreover, in the Experience condition, participants were not presented with non-transparent descriptions, as occurred in Experiment 1. However, like Experiment 1, participants in the Experience condition experienced data in the form of sequentially generated outcomes.

Experiment 2 only permitted between-subject comparisons. In brief, we found that participants gave more accurate probabilistic inferences when allowed to experience simulated outcomes. Second, there was little or no difference in accuracy of probabilistic inferences between the groups of Sophisticated and Naïve participants who experienced simulated

outcomes. These results are important. They suggest that the ability to encode frequencies in the form of sequentially experienced frequency data can be used to improve probabilistic inferences across a wide range of tasks.

We note also that the questions posed in Experiment 2 are important for many types of economic decisions where people would want to compare the probabilities of outcomes relative to starting points, goals, and/or the outcomes of others. Indeed, in a survey (Soyer & Hogarth, 2012a), we established that the majority of statistical analyses in the economics literature describe results in a way analogous to the description in our Analytic condition. Moreover, when we posed a simpler (univariate) version of this problem to economic scholars, the respondents made the same kinds of mistake as the Analytic group in Experiment 2.Clearly such presentation modes are far from transparent. An important implication of our work, therefore, is that simulated experience can make statistical analyses accessible to a wide range of decision makers.

Finally, for both of the Experience subgroups, we collected data on numbers of simulations. Before deciding on a final investment plan, the Sophisticated simulated an average of 7 different strategies some 19 times each. The Naïve simulated an average of 5 strategies about 9 times each. Thus, as in Experiment 1, we find that more statistically sophisticated participants choose to experience more outcomes than the less sophisticated.

2.6. General discussion

Our experiments demonstrate that probabilistic judgments can be more accurate when based on experience in the form of sequentially simulated outcomes as opposed to description. Moreover, this holds for participants

who vary in statistical sophistication. Indeed, although the statistically sophisticated outperformed naïve participants on some descriptive problems, there was little or no difference in performance after experience.

We interpret our results within the framework of characteristics of description and experience presented in Figure 2.1 that represents the latter as varying on a kind-wicked dimension and the former on transparency. Our experimental tasks all involved kind environments in that the samples of data that people observed were representative. At the same time, the tasks were not transparent to most respondents. Indeed, the results of Experiment 2 emphasize the value of experience for dealing with non-transparent tasks.

Given the human tendency to attend to the information that has actually been sampled (Fiedler, 2000), the quality of inferences based on experience should decline as tasks shift from being kind to wicked. At the limit, when tasks are wicked but transparent, description should be preferred to experience. However, it is uncertain what to choose in environments characterized by wicked experience and non-transparent description. An important task for future research is to understand the nature of this tradeoff.

In the present work, we deliberately limited our attention to kind environments because we wanted to observe how experiential information would affect answers to problems where people typically make erroneous judgments based on description. One critique of our approach is that by simplifying experience to the observation of a single bivariate relation – as well as providing optional memory aids – we were essentially telling

our participants the answers to the questions. We have several responses to this critique.

First, our participants were never told how much data to sample or whether their estimates were "correct." Second, if people had been told what to do, and had no doubts, then all responses would favor experience over description. This did not happen. Indeed, participants lost money by failing to select experience over description in all cases. Third, neither using nor creating a simulation tool for a specific process requires knowledge of the probabilities of the outcomes it produces. In fact, creating such a tool requires the same information as a description, that is, about the structure of the problem and its parameters. Finally, consider the coin toss exercise, which we used to familiarize our participants with the functions of simulation. Conceptually what we did was to make this "tossing" exercise available for all the other problems; that is, "toss" many groups of 25 people, many projects with seven parts, or many investment decisions given an estimated model.

Our work raises both theoretical and practical concerns that relate mainly to the boundary conditions of our findings. We consider five issues: (a) How much and what kind of experience do people need to make appropriate responses? (b) What do people learn from simulated experience? (c) Do people trust simulation mechanisms? (d) How general is the simulation methodology, that is, can models be easily constructed for all types of situations? (e) How does experience in the form of simulated outcomes solve the problem of understanding probabilities of unique events?

a) Amount and kinds of experience

In our experiments, we deliberately let participants determine the amount of information – in terms of number of trials – that they wanted to

experience. This raises the issue as to how much experience – that is number of trials – people need to reach conclusions with which they feel comfortable. Our participants generally experienced larger samples than those of Hertwig et al. (2004), possibly because our task was less taxing in terms of memory. Moreover, our data showed a relation between statistical sophistication and sample size with the more sophisticated requiring larger samples. Thus, we suspect that individual differences could play a role in the amount of information that people seek (see also Lejarraga, 2010).

A further important issue is whether being actively involved in the sampling process makes a difference compared to simply observing outcomes and this also demands further research. For example, one could conduct experiments varying both sample size and intervention in the sampling process and elicit measures of confidence as well as judgments of probability. One hypothesis, suggested by studies on inter-generational learning (e.g., Schotter & Sopher, 2003), is that people may not pay sufficient attention to the histories of trials that others experience but that they might be sensitive to advice offered by others.

b) Learning

It is unclear what our participants really learned from sampling experience other than probabilities of specific outcomes. As noted previously, past studies with the Monty Hall problem suggest that, after experiencing multiple trials, people do learn to make appropriate responses. However, there is no evidence that they achieve more insight into the problem (Franco-Watkins et al., 2003).

The simulated experience featured in this study only provided information on outcomes. Thus, it is unclear what insights our participants might have gained about the underlying processes generating the outcomes or if they could transfer knowledge to analogous tasks. We suspect that the experiences we provided our participants did not involve much transfer of knowledge. At the same time, however, we hypothesize that experience with outcomes of random processes can lead to greater understanding of uncertainty. Hence illuminating the factors that enable simulated experience to lead to greater insights is an important topic for future research.

For example, the use of simulation to confront people's first and erroneous judgments of probability in specific circumstances can, if accompanied by further explicit instruction, open the way to insights into the complexities of probability theory. Indeed, Sedlmeier's (1999) "flexible urn" concept is close to this suggestion in that it involves both perceiving simulated data dynamically and active involvement with a computer interface. However, most of Sedlmeier's work – and suggestions – have focused on how to present information in the form of *aggregate* natural frequencies as opposed to *sequentially* observed frequency data (see, e.g., Sedlmeier, 2000; Sedlmeier & Gigerenzer, 2001).

Our results encourage the belief that simulated experience could have an important role to play in teaching probability and statistics at all levels – from grade school through university and beyond. Nowadays, it is relatively simple to build simulation models and, with the widespread availability of personal computers, there is no reason why the ideas tested in this paper could not have wide application. Indeed, the Statistics Online Computational Resource (SOCR) website – www.socr.ucla.edu – provides

a repository of elegant simulations and applets for many probabilistic problems including several featured in Experiment 1. Moreover, Dinov, Sanchez, and Christou (2008) have shown that using the website while teaching statistics enhances students' understanding and retention of concepts.

c) Trust

When do people trust the implications of simulation models? We hypothesize that the key issue is the extent to which people understand the sampling mechanism. This may have several dimensions. One is the level of the participant's familiarity with the data generating process. For instance, whereas it is probably easy for the participants to understand the coin toss example, simulating outcomes related to different groups of 25 people in the birthday problem might seem odd as well as the fact that the experiential evidence typically runs counter to prior intuitions. At the same time, when people have little insight into the structure of a problem – as occurs in both the hospital and Monty Hall problems – living the experience of many outcomes can be helpful in stimulating further investigation.

Interestingly, if the participant already understands the structure of the problem – as happens in the conjunction problem – and recognizes that her capacity for calculation is deficient, she might welcome the simulation tool (Lejarraga, 2010). The investment problem in Experiment 2 is a good example of this. Participants clearly understood the goals of the exercise but lacked the ability to draw the appropriate inferences unless guided by the simulation model. In an intriguing parallel development, Goldstein, Johnson, and Sharpe (2008) have recently developed a simulation model that allows people to assess probabilities of different outcomes of pension

allocation decisions. The key idea is the same: people understand what decisions are to be made; what they do not understand are the implications; but the simulation allows them to see what would happen – in a probabilistic sense – if particular decisions were enacted many times. Further work by Haisley, Kaufmann, and Weber (2010) shows that experience sampling improves both comprehension and satisfaction with returns in investment decisions that involve risk.

Parenthetically, we note that many of the questions about trusting simulation models could also be raised about trusting the advice provided by experts. Under what conditions do people accept expert opinions and when would these be preferred to experience? We believe that this is also an important problem for future research.

Finally, it is easy to dismiss simulated experience as simply being the outcome of a "black box." We believe a more appropriate metaphor is that of a "grey box" where people experience outcomes generated by a computer as opposed to those arising from the naturally occurring environment. But much research is needed to determine what affects the different shades of grey and thus the conditions under which people do or do not feel comfortable in relying on outcomes of simulated experience.

d) Generality

Our fourth issue centers on limits to the generality of the simulation technique itself. At a conceptual level, and given sufficient ingenuity on the part of the investigator, there is almost no technical limit to the probabilistic situations that can be constructed. Whether they are meaningful, however, is another issue that can be viewed from two perspectives: the reality being modeled and the experience of the user.

For the latter, the critical issue is that already discussed above, namely the shade of grey of the box. For the former, it should be clear that the models are only as good as the fit of their assumptions to reality. As we see it, the goal of simulated experience is not necessarily to reach a precise probabilistic answer to a problem but more a means of understanding the implications of assumptions in reaching an *approximate* answer.

e) Understanding probability

Our fifth issue speaks to the meaning of probability. The main distinction is whether the concept applies to unique events (e.g., that a particular person has a certain disease) or classes of events (e.g., that people that belong to a particular group have the disease). This distinction has been given different names in the literature, for example, *epistemic* as opposed to *aleatory*, or singular versus distributional (Reeves & Lockhart, 1993). Although from the subjectivist perspective probability simply measures a degree of belief such that the distinction is irrelevant, there is much evidence that intuitions of probability are more clearly aligned with the distributional perspective (Gigerenzer & Hoffrage, 1995). People relate more easily to a statement that a fair coin tossed 100 times is expected to show heads roughly 50% of the time than the statement that the probability of heads on a single toss of the coin is 0.5. For the former, there is some informational "certainty" in the 50%. For the latter, 0.5 is a statement of total uncertainty. The experience of simulated outcomes clearly taps into people's distributional intuitions about the meaning of probability and this, in part, may explain why they find it meaningful.

From a theoretical viewpoint our approach can be seen as extending the work of Gigerenzer and his associates to its logical conclusion. As noted previously, Gigerenzer and Hoffrage (1985) emphasized the importance

of experience in the form of sequentially experienced outcomes and advocated presenting statistical information in the form of natural frequencies that summarize these outcomes. In other words, Gigerenzer's paradigm involves *descriptions of experience*. We have suggested mechanisms that allow experiencing sequential outcomes and that eliminate difficulties associated with description. That different problem representations can induce different psychological mechanisms and responses is well-established (Hogarth, 1982). Identifying simple means to induce accurate responses is important both theoretically and practically.

Appendix 2.A. Answers to the seven probabilistic inference problems in Experiment 1

1. Bayesian updating

$$p(Cancer) = 1\%$$
$$p(+ \mid Cancer) = 90\%$$
$$p(- \mid Cancer) = 9\%$$

$$p(Cancer \mid +) = \frac{p(Cancer) * p(+ \mid Cancer)}{p(Cancer) * p(+ \mid Cancer) + (1 - p(Cancer)) * p(- \mid Cancer)} \cong 10\%$$

Thus, the answer is:

c) Out of 10 women with a positive mammogram, about 1 has breast cancer.

2. Birthday problem

There are 365 days in a year.

The approximate probability of a birthday MATCH between any two specific people is 1/365. The probability of a NO MATCH is thus 364/365.

The probability of 2 NO MATCHES in a row is $(364/365)^2 = 0.9972$.

The probability of n NO MATCHES in a row is $(364/365)^n$

There are 300 different combinations of 2 people in a group of 25.

The probability of 300 NO MATCHES in a row is $(364/365)^{300} = 44\%$

The probability that there is at least one MATCH = 1 – 44% = 56%

The answer is approximately 56%

3. Conjunction problem

$$p(part_i) = 75\% \ , \quad i = 1,\ 2,\ ...,\ 7$$
$$p(success) = p(part_1) * p(part_2) * ... * p(part_7) = \left[p(part_i)\right]^7 \cong 13.3\%$$

The answer is approximately 13.3%

4. **Linda problem** (Tversky & Kahneman, 1983)

p(bank teller) > p(bank teller and feminist) by conjunction rule.

5. **Hospital problem** (Tversky & Kahneman, 1974)

The answer is "the smaller hospital"…
…because smaller sample sizes exhibit more variability.

6. Regression toward the mean

Consider the prediction of X_2 (the second TOEFL score) using X_1 (the first TOEFL score). The least squares regression of X_2 on X_1 would provide us with the two coefficients:

$$\hat{\beta} = \frac{\sum (X_{1,i} - \overline{X}_1)(X_{2,i} - \overline{X}_2)}{\sum (X_{1,i} - \overline{X}_1)^2} = \frac{Cov(X_1, X_2)}{Var(X_1)} = r_{X_1 X_2} \frac{s_{X_1}}{s_{X_2}}$$
$$\hat{\alpha} = \overline{X}_2 - \hat{\beta}\overline{X}_1$$

Where $-1 < r_{X_1 X_2} < 1$ is the sample correlation coefficient between X_1 and X_2 and *s* is the standard deviation.

The predicted values would be given by:

$$\hat{X}_2 = \hat{\alpha} + \hat{\beta}\hat{X}_1$$

Substituting the coefficients would give us:

$$\frac{\hat{X}_2 - \overline{X}_2}{s_{X_2}} = r_{X_1 X_2} \frac{X_1 - \overline{X}_1}{s_{X_1}}$$

Hence, given an absolute sample correlation coefficient lower than 1, there is regression toward the mean. That is, predicted standardized value of X_2 will be closer to its mean than the standardized value of its predictor.

Therefore the answer becomes:

b) It is more likely that the student in question now gets a worse ranking.

7. <u>Monty Hall problem</u>

The a priori probability that the prize is behind door *i* (D_i; *i* = 1, 2, 3) is:

$$p(D_i) = \frac{1}{3}$$

Assuming that the participant has selected door 1 (D_1), the probability that Monty opens door 2 (O_2) is

- if the prize were behind D_1; $p(O_2 \mid D_1) = 1 / 2$
- If the prize were behind D_2; $p(O_2 \mid D_2) = 0$
- if the prize were behind D3; $p(O_2 \mid D_3) = 1$

So, the probability that Monty opens door 2 is:

$$p(O_2) = \sum_{i=1}^{3} p(D_i) * p(O_2 \mid D_i) = \frac{1}{6} + 0 + \frac{1}{3} = \frac{1}{2}$$

Using Bayes Theorem, we have:

$$p(D_3 \mid O_2) = \frac{p(D_3) * p(O_2 \mid D_3)}{p(O_2)} = \frac{\frac{1}{3} * 1}{\frac{1}{2}} = \frac{2}{3}$$

and

$$p(D_1 \mid O_2) = \frac{p(D_1) * p(O_2 \mid D_1)}{p(O_2)} = \frac{\frac{1}{3} * \frac{1}{2}}{\frac{1}{2}} = \frac{1}{3}$$

Therefore, the probability of winning is higher if one changes the door, which implies that the optimal strategy is to change the initial choice, so the answer is:

b) Change to the other door

Appendix 2.B. Instructions for the experience tasks and information on simulators

The instructions provided to the participants in all groups before each problem during the experience tasks are detailed below. Also provided are details on the specific structure of the simulators for each problem which, of course, were not shared with the respondents.

Experiment 1: Tasks and simulators

1. Bayesian updating

Instructions to participants:
This program has the same functions as the coin toss program. However, it is designed for this particular problem. With each click on the SIMULATE button you will "meet" a different woman with a positive test result. For each of them it will show (1) if she really has cancer and (0) otherwise. Hence, with each click you will see the outcome associated with one randomly selected woman. You can click as many times you wish and select and obtain statistical summaries of a subset of outcomes that you have previously sampled by selecting it with the mouse.

Information on the simulator:
The simulator contained in its database a long column of 1s and 0s, unobservable to the user. Each entry was generated randomly, such that with 1% probability it was equal to 1 and with 99% probability it was equal to 0. For each entry, the simulator also generated a second information. If the entry was 1, the second information would be 1 with 90% probability and 0 with 10% probability. If the entry was 0, the second information would be 1 with 9% probability and 0 with 91% probability. The user could not observe these calculations. With each click of the user, the simulator located a 1 in the second information and looked at the entry associated with it. If it was 1, it reported 1 to the user. If it was 0, it reported 0.

2. Birthday problem

Instructions to participants:
This program has the same functions as the coin toss program. However, it is designed for this particular problem. With each click on the SIMULATE button you will "meet" with a different group of 25 people. For each group it will show (1) if two or more of them have the same birthday and (0) otherwise. Hence, with each click you will see the outcome associated with one group of 25 people. You can click as many times you wish and select and obtain statistical summaries of a

subset of outcomes that you have previously sampled by selecting it with the mouse.

Information on the simulator:
At each click, this simulator randomly generated 25 numbers between 1 and 365. Then it sorted them from minimum to maximum and took first differences. The user could not observe these calculations. If any of these first differences were 0, it reported 1 to the user, if not it reported 0.

3. Conjunction problem

Instructions to participants:
This program has the same functions as the coin toss program. However, it is designed for this particular problem. With each click on the SIMULATE button it will generate seven parts of a different project. For each project it will show (1) if all the parts are successful and (0) otherwise. Hence, with each click you will see the outcome associated with one project. You can click as many times you wish and select and summarize any subset of outcomes that you have previously sampled.

Information on the simulator:
With each click this simulator randomly generated a string of seven numbers. Each number was either 1 (with 75% probability) or 0 (with 25% probability). The user could not observe these calculations. Then if the sum of all the entries was 7, it reported 1 to the user, if it was less than 7, it reported 0.

4. Linda problem

Instructions to participants:
This program has the same functions as the coin toss program. However, it is designed for this particular problem. Here you see 7 SIMULATE buttons, each of them associated with a category that Jessica can be a member of. With each click on one of the SIMULATE buttons you will see the answer of a Jessica-like person to the question "are you a member of this category?" For each category it will show (1) if she said "yes" and (0) otherwise. Hence, with each click you will see the outcome associated with one answer of a Jessica-like person about that particular category. You can click as many times you wish and select and summarize any subset of outcomes that you have previously sampled.

Information on the simulator:
For this program the probabilities for each response category were exogenously determined by the experimenters but conformed to the restrictions implied by the conjunction rule. Hence given the probability for a given category is "p", when the user generated an answer for that category, the program produced 1 with probability p and 0 with probability 1-p. This kind of interaction was programmed for each category, with their respective, exogenously determined probabilities. The user had to simulate for each category separately and finally observed seven columns of 1s and 0s, one for each category.

5. The hospital problem

Instructions to participants:
This program has the same functions as the coin toss program. However, it is designed for this particular problem. With each click on the SIMULATE button it will generate one day, where 45 babies are born in the larger hospital and 15 in the smaller one. On the screen you see one column reserved for the larger hospital and one for the smaller one. For each day it will show (1) for each hospital if 60% or more of the babies born in that day were girls and (0) otherwise. Hence, with each click you will see the outcome associated with one day, for each hospital. You can click as many times you wish and select and obtain statistical summaries of a subset of outcomes that you have previously sampled by selecting it with the mouse.

Information on the simulator:
With each click this simulator randomly generated two columns of numbers. One contained 45 entries, the other contained 15, where each of the 60 entries were either 1 (with 50% probability) or 0 (with 50% probability). Then the program summed the numbers of the two columns. The user could not observe these calculations. If for the column with 45 entries, the sum was equal to or more than 24, it reported 1 in the column for the larger hospital and 0 otherwise. Analogously, if for the column with 15 entries, the sum was equal to or more than 9, it reported 1 in the column for the larger hospital and 0 otherwise.

6. Regression toward the mean problem

Instructions to participants:
This program has the same functions as the coin toss program. However, it is designed for this particular problem. With each click on the SIMULATE button the students will enter a different TOEFL exam. For each exam, it will show (1) if the student was ranked higher than before and (0) otherwise. Hence, with each click you will see the outcome associated with one TOEFL exam. You can click as many times you wish and select and summarize any subset of outcomes that you have previously sampled.

Information on the simulator:
With each click this simulator randomly generated a column with 1000 random numbers from a normal distribution with mean 50 and standard deviation 10. Second and third columns were calculated by adding to each entry of the first column a random number, drawn from a normal distribution with mean 0 and standard deviation 5. Consequently, the correlation between the second and third columns is approximately 0.8. Here, the second column represents the first test score, the third column represents the second test score and the first column is the main factor (e.g. ability) that affects the scores. The 90^{th} percentile in the first test was identified by locating the100^{th} largest entry in the second column. The corresponding entry in the third column was then also located. The user could not observe these calculations. If the rank of the entry identified in the third column was below 100, the program reported 1 to the user, and 0 if otherwise.

7. Monty Hall problem

Instructions to participants:
This program has the same functions as the coin toss program. However, it is designed for this particular problem. With each click on the SIMULATE button you will play the game: the program will generate three doors and it will eliminate one of the two doors that you have not selected and that does not lead to the car. On the screen you see one column reserved for the door you selected and one for the other remaining door. For each game it will show (1) for the door with the car behind it and (0) for the door that does not lead to the car. Hence, with each click you will see the outcome associated with one game. You can click as many times as you wish and select and obtain statistical summaries of a subset of outcomes that you have previously sampled by selecting it with the mouse.

Information on the simulator:
With each click, this simulator randomly generated a random number between 1 and 3. Depending on the generated number, it generated a column with three numbers, where one of them is (1) and the other two are (0)s, such that (1) corresponds in the sequence to the random number previously generated. Then the program eliminated one (0) from the first two entries of the sequence (hence it always considered the third entry as the selection of the user). The user could not observe these calculations. Then the program reported the third number in the sequence within the column reserved to the door selected by the user and the remaining number in the sequence within the other column.

Experiment 2: Experience task and simulator

Instructions to participants:
In this program, each click on the SIMULATE button will generate a new outcome, given your investment choices. On the screen you see two boxes where you can enter your investment choices. For the same inputs, you can get different outcomes with each click, as they would depend on the state of the world you find yourself in. For each set of inputs (investment choices) you can click as many times as you wish and select and obtain statistical summaries of a subset of outcomes that you have previously sampled by selecting it with the mouse. You can also copy and paste selected subsamples and compare them with other outcome samples that were generated through different investment choices.

Information on the simulator
With each click, the simulator predicted the outcomes using the model and the estimation results shown in Table 4, given the two inputs of the user. The error term was assumed to be normally distributed. Users could not observe these calculations. Only the outcome was reported in the last row of a column, which also stored all the previously predicted outcomes by the user. Next to each outcome, the inputs that were used to predict that outcome were also displayed. Similar to the coin toss simulator (see Figure 2.2), this simulator also provided the users with information on the count, sum and average of the subsets of outcomes that the user chose to select with the mouse. Moreover, the user was able to copy and paste subsets of data near another subset of data for a clearer visual inspection between the two subsets, say given two different sets of inputs.

Appendix 2.C. Statistical tests on differences between proportions of correct answers in Experiment 1

Table 2.C1. Difference between the proportions of correct answers in Experience and Analytic

	Sophisticated A-E		Sophisticated E-A		Naïve	
	Δ	*t*(df)	Δ	*t*(df)	Δ	*t*(df)
Bayesian updating	0.79	10.1(42)*	0.55	5.8(37)*	0.80	8.9(19)*
Birthday problem	0.52	5.4(37)*	0.48	4.6(53)*	0.65	6.1(19)*
Conjunction problem	0.20	1.6(59)	0.23	2.2(58)*	0.50	3.7(38)*
Linda problem	0.87	14.0(49)*	0.65	7.2(38)*	0.80	8.4(38)*
Hospital problem	0.58	6.2(37)*	0.36	3.8(37)*	0.75	7.7(19)*
Regression toward the mean	0.36	3.0(60)*	0.45	4.3(48)*	0.45	3.2(37)*
Monty Hall problem	0.58	6.3(55)*	0.48	5.1(37)*	0.80	8.6(31)*

(*) indicates significantly positive difference at 95% confidence level

Table 2.C2. Difference between the proportions of correct Analytic answers in Sophisticated E-A and A-E

	Δ	*t*(df)
Bayesian updating	0.25	2.2(55)*
Birthday problem	0.09	1.4(45)
Conjunction problem	-0.03	-0.3(60)
Linda problem	0.23	2.3(50)*
Hospital problem	0.23	1.8(60)
Regression toward the mean	0.13	1.1(59)
Monty Hall problem	0.19	1.4(59)

(*) indicates significantly positive difference at 95% confidence level

Table 2.C3. Difference between the proportions of correct answers in Sophisticated A-E and Naïve

	Analytic		Experience		Final	
	Δ	*t*(df)	Δ	*t*(df)	Δ	*t*(df)
Bayesian updating	-0.03	-0.2(39)	-0.03	-1.0(30)	0.09	0.7(37)
Birthday problem	0.03	1.0(30)	-0.10	-0.7(42)	0.06	0.4(42)
Conjunction problem	0.30	3.2(48)*	-0.01	-0.1(41)	0.02	0.2(39)
Linda problem	0.00	0.0(40)	0.07	0.9(27)	0.05	0.3(40)
Hospital problem	0.13	1.1(44)	-0.03	-1.0(30)	0.16	1.2(35)
Regression toward the mean	0.07	0.6(42)	-0.02	-0.2(41)	0.19	1.4(42)
Monty Hall problem	0.16	1.4(47)	-0.08	-1.0(49)	0.10	0.7(39)

(*) indicates significantly positive difference at 95% confidence level

Appendix 2.D. Histograms of responses given to six problems in Experiment 1

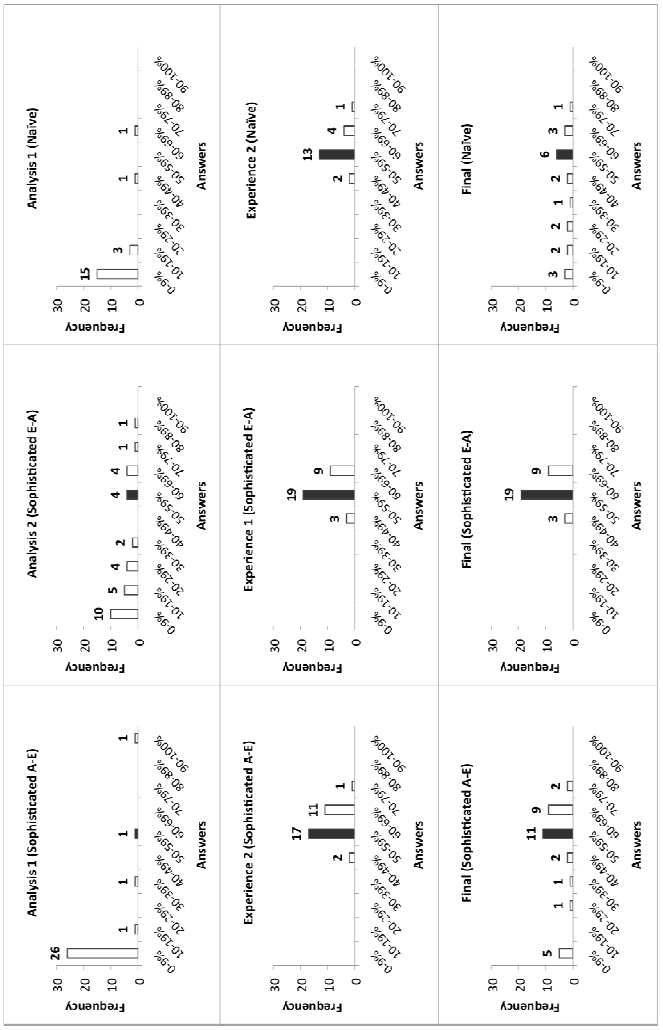


Figure 2.D1.Histograms of responses given to the birthday problem. From left to right, columns represent Sophisticated A-E, Sophisticated E-A and Naïve, with sample sizes 31, 31 and 20, respectively. From top to bottom, rows represent analytical responses, responses after simulated experience, and final responses. The numbers in the columns represent the numbers of answers. The dark column identifies the correct answer.

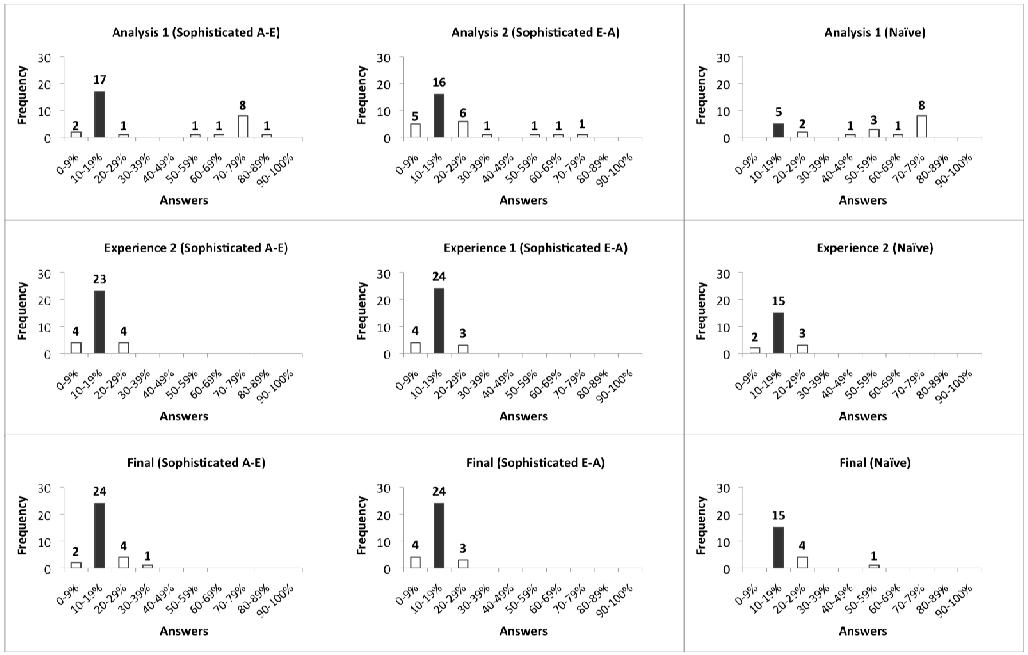


Figure 2.D2.Histograms of responses given to the conjunction problem. From left to right, columns represent Sophisticated A-E, Sophisticated E-A and Naïve, with sample sizes 31, 31 and 20, respectively. From top to bottom, rows represent analytical responses, responses after simulated experience, and final responses. The numbers in the columns represent the numbers of answers. The dark column identifies the correct answer.

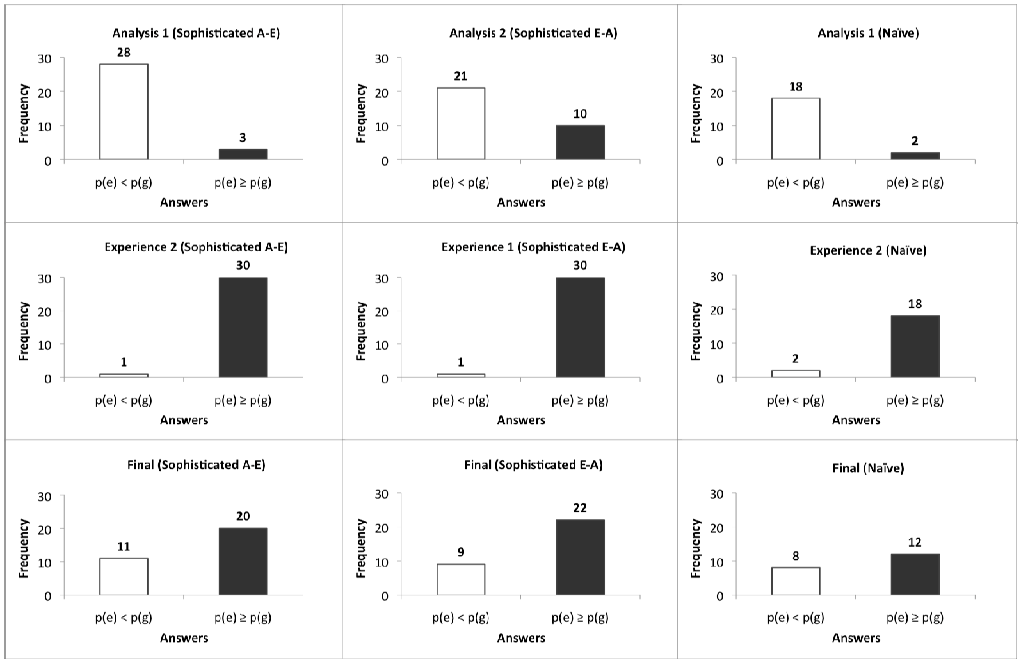


Figure 2.D3.Histograms of responses given to the Linda problem. From left to right, columns represent Sophisticated A-E, Sophisticated E-A and Naïve, with sample sizes 31, 31 and 20, respectively. From top to bottom, rows represent analytical responses, responses after simulated experience, and final responses. The numbers in the columns represent the numbers of answers. The dark column identifies the correct answer.

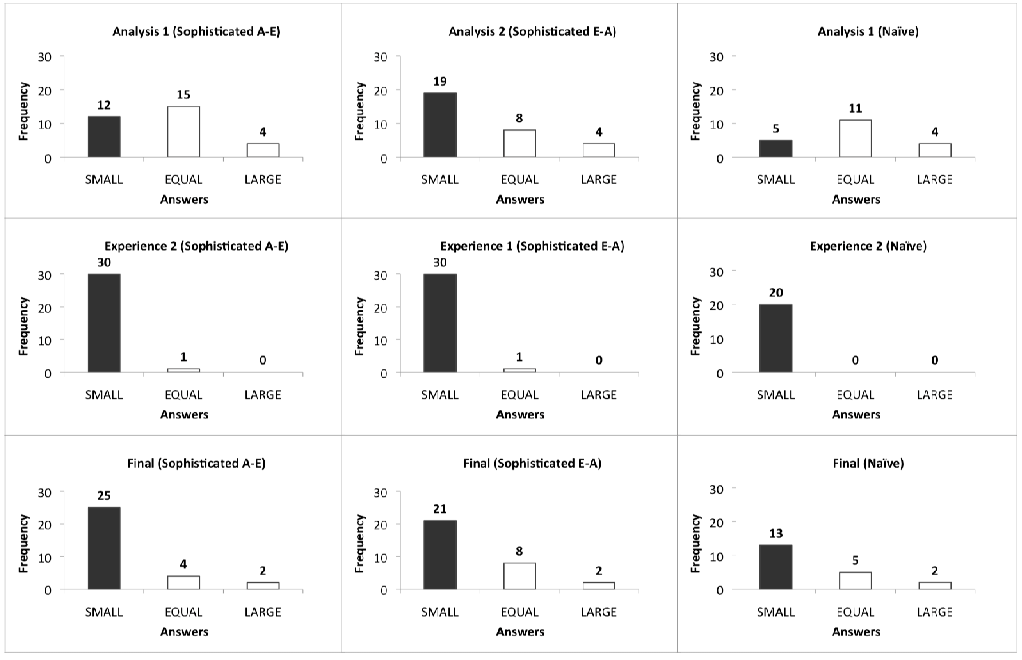


Figure 2.D4.Histograms of responses given to the hospital problem. From left to right, columns represent Sophisticated A-E, Sophisticated E-A and Naïve, with sample sizes 31, 31 and 20, respectively. From top to bottom, rows represent analytical responses, responses after simulated experience, and final responses. The numbers in the columns represent the numbers of answers. The dark column identifies the correct answer.

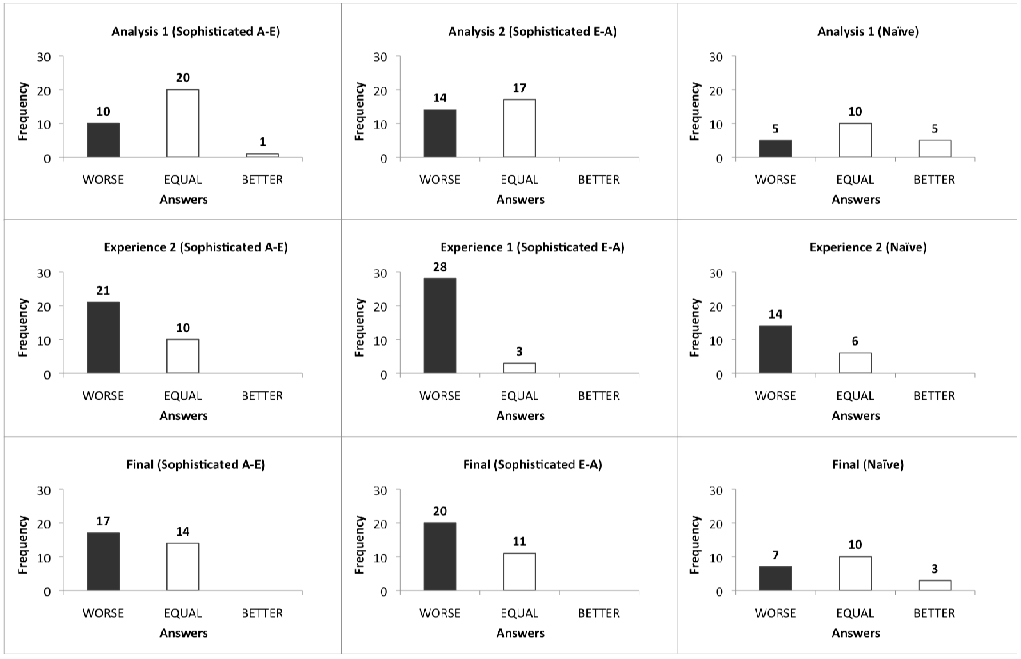


Figure 2.D5.Histograms of responses given to the regression toward the mean problem. From left to right, columns represent Sophisticated A-E, Sophisticated E-A and Naïve, with sample sizes 31, 31 and 20, respectively. From top to bottom, rows represent analytical responses, responses after simulated experience, and final responses. The numbers in the columns represent the numbers of answers. The dark column identifies the correct answer.

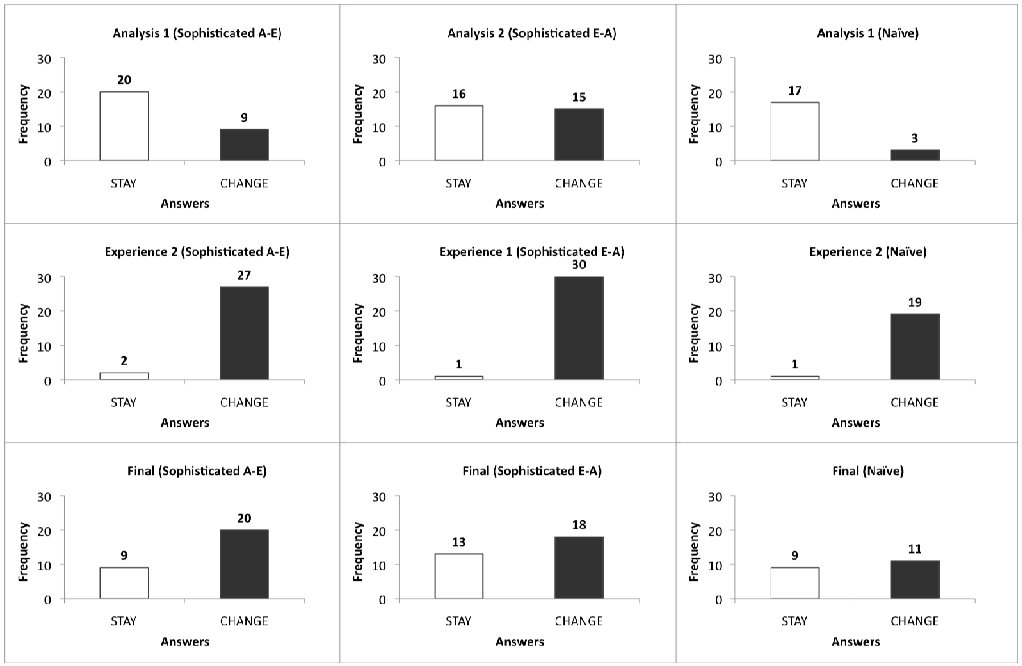


Figure 2.D6. Histograms of responses given to the Monty Hall problem. From left to right, columns represent Sophisticated A-E, Sophisticated E-A and Naïve, with sample sizes 31, 31 and 20, respectively. From top to bottom, rows represent analytical responses, responses after simulated experience, and final responses. The numbers in the columns represent the numbers of answers. The dark column identifies the correct answer.

Appendix 2.E. Rationale for answers to the four questions in Experiment 2

Question 1

This question was posed to elicit an answer from the participants. We wanted them to make an investment decision with a particular expectation about the results it would lead to. The answers given suggested that the participants in all groups identified average effects quite accurately.

Question 2

This question reflects the desire to obtain a positive outcome given any investment decision. The most popular answer for this question in the Analytic group was I_1=0 and I_2=16.7. We therefore base the calculations in this section on these particular values. Answers associated with other choices can be calculated analogously.

The answer to Question 2 depends mainly on the standard deviation of the estimated residuals (*SER*). In a linear regression analysis, SER^2 corresponds to the variance of the dependent variable that is unexplained by the independent variables and is captured by the statistic ($1-R^2$). In the set-up, this is given as 21%. One can compute the SDER using the ($1-R^2$) statistic and the variance of ΔY:

$$se(\hat{e}) = \sqrt{Var(\Delta Y)(1-R^2)} = \sqrt{(7.9^2)(1-0.21)} \cong 7 \qquad \text{(A1)}$$

Given I_1=0 and I_2=16.7 the answer to Question 2 is:

$$p(Y_i < 0 \mid I_{1,i} = 0, I_{2,i} = 16.7) = p(\hat{C} + \hat{\beta} I_i + \hat{e}_i < 0 \mid I_{1,i} = 0, I_{2,i} = 16.7) =$$

$$= p(\hat{e}_i < 0 - \hat{C} - \hat{\beta} I_i \mid I_{1,i} = 0, I_{2,i} = 16.7) = p(\frac{\hat{e}_i}{se(\hat{e}_i)} < \frac{0 - \hat{C} - \hat{\beta} I_i}{se(\hat{e}_i)} \mid I_{1,i} = 0, I_{2,i} = 16.7) =$$

$$= \Phi(\frac{0 + 0.1 - 0.3 * 16.7}{7}) = \Phi(-0.7) \cong 0.24 \qquad \text{(A2)}$$

Question 3

Here, one needs to make similar calculations as for the answer to Question 2. Given I_1=0 and I_2=16.7 the answer to Question 3 becomes:

$$p(Y_i < 5 \mid I_{1,i} = 0, I_{2,i} = 16.7) = p(\hat{C} + \hat{\beta} I_i + \hat{e}_i < 5 \mid I_{1,i} = 0, I_{2,i} = 16.7) =$$

$$= p(\hat{e}_i < 5 - \hat{C} - \hat{\beta} I_i \mid I_{1,i} = 0, I_{2,i} = 16.7) = p(\frac{\hat{e}_i}{se(\hat{e}_i)} < \frac{5 - \hat{C} - \hat{\beta} I_i}{se(\hat{e}_i)} \mid I_{1,i} = 0, I_{2,i} = 16.7) =$$

$$= \Phi(\frac{5 + 0.1 - 0.3 * 16.7}{7}) = \Phi(-0.01) \cong 0.50 \qquad \text{(A3)}$$

Question 4

This question reflects the desire to be better off with respect to an alternative of no-action in terms of Investment 1 and 2. Finding the answer requires making one additional calculation. Specifically, we need to know the standard deviation of the difference between two random variables, that is

$$(Y_i \mid I_{1,i} = x_1, I_{2,i} = x_2) - (Y_j \mid I_{1,j} = 0, I_{2,j} = 0), \qquad \text{(A4)}$$

where $x_1 > 0$ and/or $x_2 > 0$

We know that (Y_i | $I_{1,i}$=x_1, $I_{2,i}$=x_2) is an identically, independently and normally distributed random error with an estimated standard deviation of again 7. Given that a different and independent shock occurs for different people and actions, the standard deviation of becomes:

$$\sqrt{Var((Y_i \mid I_{1,i}=x_1, I_{2,i}=x_2)-(Y_j \mid I_{1,j}=0, I_{2,j}=0))} =$$
$$= \sqrt{Var((Y_i \mid I_{1,i}=x_1, I_{2,i}=x_2)+Var(Y_j \mid I_{1,j}=0, I_{2,j}=0))} = \sqrt{(7^2+7^2)} \cong 9.9 \quad \text{(A5)}$$

Given I_1=0 and I_2=16.7 the answer to Question 4 becomes:

$$p(Y_i \mid I_{1,i}=0, I_{2,i}=16.7 > Y_j \mid I_{1,j}=0, I_{2,j}=0) =$$
$$= p(\hat{C}+\hat{\beta}I_i+\hat{e}_i-\hat{C}-\hat{\beta}I_j-\hat{e}_j>0 \mid I_{1,i}=0, I_{2,i}=16.7, I_{1,j}=0, I_{2,j}=0) =$$
$$= p(\hat{e}_i-\hat{e}_j>0-\hat{\beta}I_i+\hat{\beta}I_j \mid I_{1,i}=0, I_{2,i}=16.7, I_{1,j}=0, I_{2,j}=0) =$$
$$= p(\frac{\hat{e}_i-\hat{e}_j}{se(\hat{e}_i-\hat{e}_j)}>\frac{0-\hat{\beta}I_i+\hat{\beta}I_j}{se(\hat{e}_i-\hat{e}_j)} \mid I_{1,i}=0, I_{2,i}=16.7, I_{1,j}=0, I_{2,j}=0) =$$
$$= 1-\Phi(\frac{0-0.3*16.7}{9.9}) = 1-\Phi(-0.51) \cong 0.69 \quad \text{(A6)}$$

Soyer, E., & Hogarth, R. M. (2011). The size and distribution of donations: Effects of number of recipients. *Judgment and Decision Making*, 616-628.

3. The size and distribution of donations: Effects of number of recipients

(Based on Soyer & Hogarth, 2011)

3.1. Introduction

Recently, much literature has highlighted the importance of numbers of alternatives in choice. This can be considered from two perspectives. In one, investigators have reported effects when people make unique selections from different numbers of alternatives (see, e.g., Iyengar & Lepper, 2000; Schwartz, 2004; Scheibehenne, Greifeneder & Todd, 2010). For example, studies have documented differential satisfaction with choice for decisions involving pens (Shah & Wolford, 2007), pension plans (Iyengar, Huberman & Jiang, 2004), gift boxes (Reutskaja & Hogarth, 2009), and wines (Bertini, Wathieu & Iyengar, 2010).Moreover, a recent meta-analysis suggests that the magnitude of effects depends on preconditions, choice moderators and the contexts in which decisions are made (Scheibehenne, Greifeneder & Todd, 2010).

The focus in the second perspective is on what happens when people allocate resources across different numbers of alternatives (see, e.g., Andreoni, 2007). This is the topic of the present paper. Specifically, we consider this issue in the context of charitable donations and investigate the effects of numbers of alternatives on the amount of total donations as well as their distribution across charitable organizations (NGOs) and specific campaigns. Both of these issues are important from theoretical and practical viewpoints. For example, when attempting to maximize donations, NGOs might consider whether donors perceive them as belonging to small or large subsets of potential recipients. At the same time, NGOs often seek funds for different campaigns and it is important to

know how the number and presentation of campaigns affect total donations.

We report two experiments. In the first, we explore effects when donors allocate funds across different numbers of NGOs. In the second, we investigate what happens when a single NGO solicits contributions for different numbers of campaigns. In short, we find two effects of increasing the number of alternatives: total contributions increase albeit at a decreasing rate; and distributions of donations are affected. Specifically, these tend to become less egalitarian in the case of NGOs but more so in the case of campaigns. In the second experiment, we also investigate the use of "drop down" menus in donation interfaces for soliciting donations to specific campaigns. When, as in current practice, choice is limited to one of several alternatives, contributions are lower than when this restriction does not apply. We conclude by discussing implications.

3.2. Relevant literature

Several recent studies have focused on different aspects of the donation process including determinants of donation decisions (Landry et al., 2006; Chang, 2005), the impact of presentation mode (Small, Loewenstein & Slovic, 2007), the effect of social interactions (Schweitzer & Mach, 2008), herding behavior among donors (Martin & Randal, 2008) and methodologies for measuring altruistic behavior (Bekkers, 2007).

Andreoni (2007) specifically examined the effects of numbers of recipients on donations in the context of an experimental economics game. He found that, as the number of recipients increased, participants gave more but that individual shares decreased. Specifically, for "the

average subject, a gift that results in one person receiving x is equivalent to one in which n people receive $x/n^{0.68}$each" (Andreoni, 2007, p. 1731).

A number of studies have shown that these kinds of results are sensitive to emotionally charged stimuli. For example, Hsee and Rottenstreich (2004) compared the effects of affect-rich as opposed to affect-poor stimuli to capture willingness to donate to saving from one to four endangered pandas. With affect-poor stimuli (dots), willingness to donate was greater for four endangered pandas than one. With affect-rich stimuli (cute pictures), however, there was no difference. Similar phenomena have been reported by Kogut and Ritov (2005a; 2005b). They have identified conditions under which people give more to help single individuals in need than to groups of individuals with the same needs. The key is providing specific information about the single individual (e.g., name and a picture) and eliciting judgments in separate as opposed to joint evaluation mode (Hsee et al., 1999).

The phenomenon that emotional responses are greater toward individual victims as opposed to aggregates has been termed the "collapse of compassion" and raises the issue of why and how it occurs. Cameron and Payne (2011) note that most studies demonstrating this phenomenon have been conducted within the context of donation decisions and they argue that the collapse is not because people lack feelings about larger numbers. Instead, large numbers cue people to regulate their emotions and particularly when they are motivated to do so (e.g., when money is at stake). Cameron and Payne (2011) go on to provide experimental evidence consistent with their hypothesis.

Two recent studies analyzed the effects of numbers of options on altruistic behavior without manipulating emotions. Scheibehenne, Greifeneder and

Todd (2009) conducted an experiment involving charitable institutions while studying possible moderators of choice overload. Specifically, participants (mainly students) were endowed with 1 Euro and had to decide either to donate it all to one institution they could choose from a specified list or to keep the money for themselves. Their findings suggest that more choices (represented by longer lists) increase the proportion of donors. In addition, people are more likely to give to charities that are better known. Note, however, that this study did not address the issue of allocating donations across alternative charities or multiple campaigns offered by one institution. Carroll, White and Pahl (2011) studied effects on people's choices of the number of alternative opportunities for volunteer work. They found adverse effects of more choice in that decisions to defer commitment were greater when there were more alternatives.

As in the above two studies, we do not make use of emotional stimuli in our work but (with one exception) we do not limit choices to one of several alternatives.

3.3. Hypotheses

In conceptualizing how donors' decisions are affected by numbers of potential recipients, we consider three issues. First are effects due to knowledge about the recipients.[6] Second, we consider the impact of numbers of potential recipients. And third, we speculate on how the number of alternatives changes the distributions of donations across recipients.

[6] By knowledge we mean how much a person is aware of the existence of the recipient, be it an NGO or campaign.

We hypothesize that donations made to specific NGOs or campaigns increase with knowledge about them (see also Scheibehenne, Greifeneder & Todd, 2009). This leads to:

H1. Recipients that are better known receive more donations.

When considering the impact of numbers of potential recipients, three points are important. First, donations are limited in that donors face budget constraints. Second, we assume that the utility donors obtain from giving increases with the size of donations but at a decreasing rate (Andreoni, 2007). Third, we hypothesize that decisions to make donations are sensitive to perceived needs of recipients. Thus, factors that signal perceived need are important. One such factor is the number of potential recipients. Our rationale is simple. If a single NGO is seeking funds for a specific cause, that cause might be seen as important and worthy of support. However, if several NGOs are seeking funds for the same (or similar) cause, the need will be perceived as greater. For campaigns, similar reasoning applies; the larger the number of campaigns offered by an NGO, the larger the perceived need.[7] These points can be summarized by our second hypothesis:

H2. Donations increase with the number of potential recipients, but at a decreasing rate.

To explore the relation between number of recipients, perceived need, and donations, we conducted a preliminary study with undergraduate students at Universitat Pompeu Fabra (our main experiments involve participants

[7] Saying that perceived need is a function of numbers of NGOs or campaigns begs the interesting question of how potential donors perceive specific sets of NGOs (or campaigns). This issue, however, is beyond the scope of the present research.

from the general public in Spain). In a survey, 40 participants were asked to imagine that they could distribute the resources of 100 NGOs to deal with four disasters. These disasters had different levels of devastation and each NGO could only deal with one disaster. For each of the four cases, the level of devastation was provided through information on casualties, homelessness and economic damages such that participants had a clear sense of the need for help. The participants assigned a higher number of NGOs to cases where the need was higher, consistent with the notion that perceived need is positively related to the number of NGOs. In a second survey, 35 participants hypothetically distributed 100 Euros across the same four disasters. The amounts donated to disasters increased with level of devastation (i.e., need), but at a decreasing rate.

Finally, how are donations distributed across potential recipients? We assume that donors seek to be fair, but in doing so they implicitly deal with two different concepts of fairness. In one, allocations reflect the relative merits of recipients. This is known as the "equity" rule. Second, although equity is sometimes assumed to guide judgments of fairness, people are also sensitive to considerations of "equality". That is, a rule whereby all recipients receive equal allocations (Sarbagh, Dar & Resh, 1994; Hertwig, Davis & Sulloway, 2002).

Indeed, Baron and Szymanska (2010) argue that if people know that one NGO makes more efficient use of its resources than all the others, then donors would be justified in allocating all their donations to that NGO. However, people are reluctant to do this and there is a diversification bias whereby donations are distributed more equally than is "rational".

How do donors reconcile the competing claims of equity and equality as the number of alternatives increases? We suggest that two factors are

important. One is that allocations reflect perceptions of differential merit. The second concerns the relative appeal of the equality principle as the number of alternatives increases and for this we envisage two possibilities: either the equality principle becomes less important as the number of alternatives increase; or, on the contrary, it becomes more important. A priori it is not clear which is correct. It may be that the equality principle is difficult to ignore when there are few alternatives. At the same time, the equality principle may be easier to implement when there are many alternatives. As a consequence, we state competing hypotheses:

H3a. The distribution of donations becomes less egalitarian across potential recipients as their numbers increase (i.e., the variability of donations increases).

H3b. The distribution of donations becomes more egalitarian across potential recipients as their numbers increase (i.e., the variability of donations decreases).

We have no objective measures of donors' judgments of merit. Thus in our experimental work we use knowledge of the NGOs and campaigns as a proxy assuming, in effect, that donors assess merit using the recognition heuristic (Goldstein & Gigerenzer, 2002).

Our two experiments aim to test the three hypotheses. The first involves conditions with varying numbers of NGOs; the second considers different numbers of campaigns offered by a single NGO.

3.4. Experiment 1: Number of NGOs

a) Participants, design and procedure

Participants were members of the general public in Spain enrolled in an online market research panel. Fifty-four percent of the 145 respondents were female and the mean age was 35 (median 34, minimum 15, and maximum 69). Most participants had at least a university degree.

At the beginning of a 40-minute market survey on an unrelated topic, they were informed that, in addition to the fixed remuneration for their participation, they had been entered in a lottery and had the chance of winning 50€ (expressed as 500 points) at the end of the session. Once the survey ended, they were notified that, if they wished, they could donate as much as they wanted of their lottery winnings (from 0 to 500 points) to certain specified NGOs, split between recipients in any way they desired. The online setup guaranteed anonymity of responses. After making their choices, one person was to be chosen at random and given the extra 50€, less the amount of her/his donations. Thus, if the winner of the lottery donated 0, s/he would get to keep 50€; if s/he donated, say, 30€, s/he would get to keep 20€. The money donated would go to precisely those NGOs specified by the winner.

The names of the NGOs were provided along with the information that their common agenda is to aid underprivileged children. The respondents were allocated at random to three groups where they faced an alphabetical list of:

- 3 NGOs (condition NGO_3 with 54 respondents)
- 8 NGOs (condition NGO_8 with 43 respondents)
- 16 NGOs (condition NGO_16 with 48 respondents)

The specific NGOs were selected after searching in the internet and popular media for international organizations with a charity agenda involving underprivileged children. The names of NGOs presented in these three conditions are shown in Table 3.1.

After making their decisions, respondents rated all 16 NGOs by indicating how much they knew about each prior to the experiment as follows: "0" implied that they had not heard of it, "1" that they had heard of it, "2" that they knew it, and "3" that the NGO is "very famous". Only six respondents claimed to have heard of all 16 NGOs and four of the 16 NGOs received average ratings greater than 1 on what we call the "knowledge score". These data suggest that 16 NGOs represented a large choice set.

Table 3.1. NGO options across conditions in Experiment 1

NGO_3	NGO_8	NGO_16
Mercy Corps	Children's Network International	Care
Oxfam	Every Child	Children in Crisis
Unicef	Global Fund for Children	Children's Network International
	Mercy Corps	EveryChild
	Oxfam	Global Fund for Children
	Stop Child Poverty	Médecins Sans Frontières
	Unicef	Mercy Corps
	United Children's Fund	Oxfam
		Plan International
		Serving Our World
		Save the Children
		SOS Kinderdorf International
		Stop Child Poverty
		Unicef
		United Children's Fund
		World Emergency Relief

b) Results

Table 3.2 lists the different NGOs in the order of their mean popularity scores that are indicated on the right hand side of the table. Here we also report the proportions of participants who stated that they had never heard of the respective NGOs. Four NGOs are quite well known whereas the other twelve are largely unknown. These results make sense within the Spanish context of the study. Unicef, for example, has a sponsorship deal with the Barcelona Football Cub that is very popular in the region where the study took place. Mercy Corps, on the other hand, is not well known within Spain.

The intermediate columns of Table 3.2 show the mean donations in points in the three experimental conditions.

Results in Table 3.2 support hypothesis *H1* at an aggregate level. Mean knowledge scores of the NGOs correlate (in an ordinal sense) with mean donations (the better known NGOs receiving larger contributions). Spearman's rho is 1.00 for NGO_3; 0.64 (p = .10) for NGO_8; and 0.47 (p = .07) for NGO_16.

To estimate the effect of knowledge at the level of individual donations, we regressed individual donation decisions (n = 1274) on knowledge scores. Controlling for individual NGO effects, number of alternatives and adjusting the standard errors for clusters of 145 different donors, we obtain a statistically significant coefficient of 17.1 (*s.e.* = 2.9, p = .001) for the knowledge score. The F-ratio of the analysis is $F(16, 144) = 18.6$, with p = .001, R^2 =.25 and *root-MSE* = 71.7. These results suggest that both at the aggregate and individual levels, better known recipients obtain larger contributions.

Table 3.2. Donation decisions by knowledge and number of alternatives in Experiment 1

NGOs	**Mean donations in points (stdev)**			**Mean knowledge score**	**Knowledge score = 0 (%)**
Condition	NGO_3	NGO_8	NGO_16		
N	54	43	48		
No. of NGOs	3	8	16		
Unicef	100 (97)	128 (163)	142 (181)	2.59	3
Médicins Sans Frontières	x	x	79 (157)	2.30	8
Oxfam	83 (80)	67 (118)	52 (102)	2.01	14
Save the Children	x	x	29 (53)	1.32	34
Global Fund for Children	x	26 (46)	0 (2)	0.44	75
Mercy Corps	53 (65)	16 (25)	0 (2)	0.39	78
Plan International	x	x	0 (2)	0.39	77
United Children's Fund	x	18 (28)	2 (14)	0.37	76
SOS Kinderdorf International	x	x	9 (39)	0.24	84
Children's Network International	x	17 (27)	1 (7)	0.21	84
Serving Our World	x	x	3 (15)	0.21	86
Stop Child Poverty	x	25 (51)	3 (15)	0.20	87
EveryChild	x	18 (28)	1 (7)	0.19	88
Care	x	x	0 (2)	0.17	86
World Emergency Relief	x	x	3 (15)	0.17	88
Children in Crisis	x	x	1 (7)	0.16	87
Total	236	314	326		

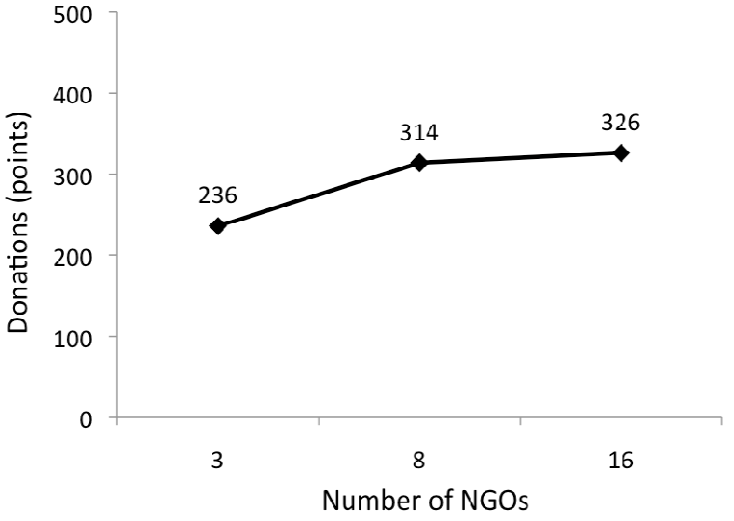


Figure 3.1. Mean donations in the three conditions in Experiment 1.

Our second hypothesis (*H2*) is that, overall, donations increase with the number of recipients but at decreasing rate. Figure 3.1 shows mean donations as a function of experimental conditions. An analysis of variance indicates that the effect of number of alternatives on donations is significant ($F(2, 142) = 2.98$, $p = .05$). When we look at pairwise contrasts and effect sizes between the mean donations, we find that the mean in condition NGO_8 is greater than in condition NGO_3 (314 vs. 236, $z = 1.91$, $p = .06$, Cohen's $d = .52$); and the mean in condition NGO_16 at 326 is also greater than in condition NGO_3 ($z = 2.23$, $p = .03$, Cohen's $d = .54$). Finally, the difference between the means for condition NGO_16 and NGO_8 is not statistically significant with a medium effect size (326 vs. 314, $z = 0.3$, $p = .78$, Cohen's $d = .42$). Post-hoc multiple comparisons through Tukey's HSD test find only a

difference between the means for NGO_16 and NGO_3 (q = 3.08, p = .08).

Further evidence that donations increase with the number of potential recipients can be seen in Table 3.3 where we provide data characterizing individual contributions. As the number of potential recipients rises, so does the proportion of participants who donate their total endowment of 500 points – from 24% (NGO_3) to 37% (NGO_8) to 50% (NGO_16). (The difference between NGO_16 and NGO_3 is significant, z = 2.8, p = .01). Moreover, note that whereas 30% of participants donate nothing when there are only three NGOs, this figure drops to 19% for the cases with 8 and 16 alternatives.

Table 3.3. Proportions of donation behavior in Experiment 1

	NGO_3	NGO_8	NGO_16
% of participants giving equal non-zero amounts	24	23	0
% of participants giving away 0 points	30	19	19
% of participants giving away all 500 points	24	37	50

Hypotheses *H3a* and *H3b* make contrary predictions – increasing as opposed to decreasing variability in donations as the number of alternatives increases. At the aggregate level, the variances of the contributions to the different NGOs are 582, 1556 and 1549 in conditions NGO_3, NGO_8, and NGO_16, respectively. The F-tests for the difference in variances between NGO_3 and NGO_16 (F(15, 2) = 2.67, p = .30) and between NGO_3 and NGO_8 (F(7, 2) = 2.67, p = .30) indicate

that the change in the variability of donations is not significant. Moreover, an analysis of variance on variances of donations by individuals shows that the effect of number of available NGOs on the variance of donations is again not significant ($F(2, 142) = 1.54$, $p = .22$).

On the other hand, in terms of the distribution of donations, in condition NGO_3, all potential recipients receive substantial donations. In condition NGO_8, four (or 50%) receive 76% of the contributions, and in condition NGO_16, four (or 25%) receive 92% of the contributions. These overall trends are also supported by the data summarized in Table 3.3; whereas 24% of participants adopt the strategy of giving the same non-zero amounts to all participants when there are three NGOs, this figure is zero for the case with 16 NGOs. These latter results are consistent with the hypothesis that the variability of donations is positively related to the number of NGOs.

Finally, it is of interest to note how changes in the number of alternatives affect the fortunes of different NGOs. When there are only three NGOs, Mercy Corps receives a large average donation despite being unknown. However, this changes dramatically as the number of alternatives increases. Unicef, on the other hand, retains its leading position, its relative share and its donation in absolute terms as the number of alternatives increases. Oxfam sees reductions in donations as the number of alternatives increases. However, being known appears to save Oxfam from the extreme reductions from which Mercy Corps suffers as the number of alternatives increases.

3.5. Experiment 2: Number of campaigns

Experiment 2 was designed to replicate the results of Experiment 1. However, it involved varying numbers of campaigns instead of varying numbers of NGOs.

a) Participants, design and procedure

The design and procedure of this second study were analogous to Experiment 1. The respondents, who were entered in a 50€ lottery (expressed as 500 points) after participating in an unrelated survey, were notified that they could make a donation (of between 0 and 500 points) if they wished at the end of the session. The participants were again members of the general public in Spain enrolled in a market research panel. Fifty percent of the 505 respondents were female and the mean age was 38 (median 38, minimum 18, and maximum 74). Most participants had at least a university degree.

Unlike participants in Experiment 1, who had to decide among charitable institutions, participants in this study faced different numbers of campaigns offered by a single, well known NGO: Unicef. The study had a between-subject design involving five conditions to which respondents were allocated at random. Three conditions involved different numbers of campaigns (1, 7, and 13) and the two further conditions varied the number of options that could be chosen when there were 7 and 13 campaigns. Specifically, in the former respondents could only donate to one of several options (from 7 or 13), whereas in the latter they could distribute their contributions across several options (out of 7 or 13).

In summary, there were five groups, each with 101 respondents, facing lists of:

- 1 campaign (condition Only_1)

- 7 campaigns (condition Single_7; campaigns were listed in a drop down menu, where donations could only be made to a single option)

- 13 campaigns (condition Single_13; campaigns were listed in a drop down menu, where donations could only be made to a single option)

- 7 campaigns (condition Multiple_7; campaigns were listed in an open menu where donations could be distributed across multiple options)

- 13 campaigns (condition Multiple_13; campaigns were listed in an open menu where donations could be distributed across multiple options)

The difference between conditions Single_7 and Multiple_7, and conditions Single_13 and Multiple_13, lies in how the options are displayed. In all the online sites of Unicef and the majority of NGOs featuring multiple campaigns, the alternatives are exclusively listed in a drop down menu (analogous to conditions Single_7 and Single_13). Hence contributors are constrained to make a selection from a list and to donate to a single recipient, that is, without being able to distribute their donations across alternatives (unless they revisit the site). We included Multiple_7 and Multiple_13 in order to observe whether the elimination of this constraint would encourage donors to distribute their contributions over multiple campaigns and thus change the distribution and, more importantly, the amount of contributions. As will be shown below, this change does have an impact.

Table 3.4. Unicef campaigns across conditions in Experiment 2

Only_1	Single_7 & Multiple_7	Single_13 & Multiple_13
Unicef (where most needed)	Where most needed	Where most needed
	Haiti, after one year	Haiti, after one year
	Emergency fund	Emergency fund
	Floods in Pakistan	Floods in Pakistan
	Libyan crisis	Libyan crisis
	Earthquake and tsunami in Japan	Earthquake and tsunami in Japan
	Water for Niger	Water for Niger
		United against hunger
		Fight against malaria
		Clean drinking water
		Children's education
		Humanitarian aid for Sudan
		Promotion of Unicef

The specific campaigns were selected following a survey of Unicef's campaigns in its 36 national websites in April 2011 (campaign compositions change depending on the occurrence of disasters). The campaigns presented in these five conditions are shown in Table 3.4. In condition Only_1, participants were asked if they would consider donating to Unicef (without mentioning a specific campaign), who then would decide how to use the contributions. In all the other conditions, the option "where most needed" was featured at the top of the options list, whereas the remainder of campaigns were displayed in a random order (this structure mimics donation sites that feature multiple options). The campaigns in conditions Single_7 and Multiple_7 were the ones available in Unicef's Spanish site (www.unicef.es) in April 2011, whereas the six additional campaigns featured in Single_13 and Multiple_13 are among those that are frequently featured in Unicef's other national sites.

The number of alternatives featured in different conditions is consistent with the current available numbers of options offered by Unicef and many other NGOs. Specifically, as of April 2011, across all websites where one can make a one-time donation to Unicef, the mean number of campaigns from which to choose is 7 (*SD* = 12.8). When the German site is excluded (this site offers an unusually large number of 72 alternative campaigns), this figure drops to 5 (*SD* = 4.5). One third of these sites offer only one alternative (denoted as "where most needed"), and only 15% feature more than 10. Hence, while condition Only_1 mimics the majority of situations encountered in online environments, conditions Single_7 and Multiple_7 represent average situations across all sites. Given current practice, conditions with 13 choices (e.g. conditions Single_13 and Multiple_13) constitute a valid analogy for large sets of alternatives.

As in Experiment 1, after making their donation decisions, respondents rated all 12 campaigns (excluding "where most needed") by indicating how much they knew about each prior to the experiment as follows: "0" implied that they had not heard of it, "1" that they had heard of it, "2" that they knew it, and "3" that the campaign is "very well known".

b) Results

In Table 3.5, the different Unicef campaigns are listed in the order of their mean knowledge scores that are indicated in the column on the right hand side of the table. Here, we again report the proportions of participants who stated that they had never heard of the respective campaigns. The campaign "where most needed" has been assigned the knowledge score of Unicef from Experiment 1. Among other campaigns, our participants were relatively more knowledgeable about two recent (as of April 2011) and highly publicized specific disasters (Japan and Haiti) and two general causes (eradication of hunger and malaria).

The intermediate columns of Table 3.5 show the mean donations in points for the five experimental conditions.

Results support *H1* at an aggregate level. As in Experiment 1, the mean knowledge scores of the campaigns correlate (in an ordinal sense) with mean donations (the better known campaigns receiving larger overall contributions). Spearman's rho is .79 (p = .05) for both Single_7 and Multiple_7; .49 (p = .08) for Single_13; and .63 (p = .03) for Multiple_13.

To identify the effect of knowledge at an individual level, we regressed individual donation decisions (excluding the ones made for "where most needed", which lacks the individual knowledge score) on knowledge scores (n = 3636). Controlling for campaign effects, number of alternatives and presence of a drop down menu, and adjusting the standard errors for clusters of 404 different donors, we obtain a statistically significant coefficient of 7.1 (s.e. = 1.15, p = .001) for the knowledge score. The F-ratio of the analysis is $F(14, 403) = 12.0$, with p = .001, R^2 =.04 and *root-MSE* = 50.3. These results suggest that both at the aggregate and individual levels, better known campaigns obtain larger contributions.

The results are in line with *H2*. Figure 3.2 shows mean donations as a function of experimental conditions. Visually, this suggests a main effect for the Multiple as opposed to the Single conditions (the means for the former being larger than those of the latter). A two-way factorial analysis of variance shows that both number of alternatives and drop down menus have significant impacts on donations ($F(4, 500) = 6.52$, p = .001). The effect of number of alternatives yields an F-ratio of $F(2, 500) = 10.78$ with p = .001, and the ratio for the effect of drop down menu is $F(1, 500)$ = 12.01 with p = .001. The interaction effect is not significant. Post-hoc

multiple comparisons using Tukey's HSD test reveal that the donations to Multiple_7 and Multiple_13 are higher than Only_1 (q = 4.8 with p = .001 and q = 6.5 with p = .001, respectively).

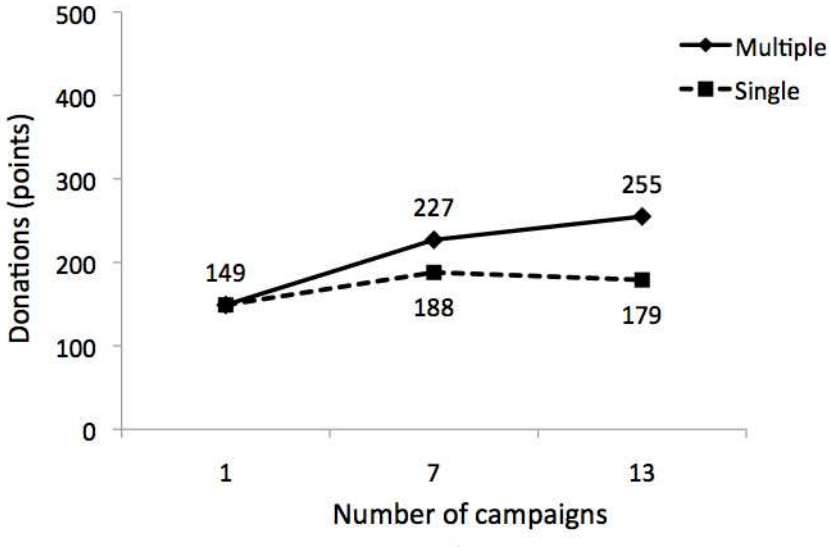


Figure 3.2. Mean donations in the five conditions in Experiment 2.

Unicef Campaigns	Mean donations in points (stdev)					Mean knowledge score	Knowledge score = 0 (%)
Condition	Only_1	Single_7	Single_13	Multiple_7	Multiple_13		
N	101	101	101	101	101		
No. of Campaigns	1	7	13	7	13		
Where most needed	149 (148)	102 (149)	74 (142)	120 (158)	93 (146)	2.59*	3
Earthquake and tsunami in Japan	x	19 (64)	13 (56)	28 (67)	18 (52)	1.55	13
Haiti, after one year	x	35 (103)	5 (39)	27 (65)	16 (45)	1.24	20
United against hunger	x	x	20 (85)	x	23 (53)	1.15	23
Fight against malaria	x	x	0 (0)	x	11 (25)	1.11	23
Children's education	x	x	22 (73)	x	20 (53)	0.99	33
Libyan crisis	x	5 (50)	0 (0)	10 (23)	4 (13)	0.93	41
Clean drinking water	x	x	19 (77)	x	18 (38)	0.84	42
Promotion of Unicef	x	x	19 (85)	x	27 (94)	0.70	53
Emergency fund	x	16 (55)	1 (10)	20 (61)	9 (23)	0.67	52
Humanitarian aid for Sudan	x	x	0 (0)	x	6 (17)	0.59	55
Floods in Pakistan	x	0 (0)	0 (0)	11 (24)	5 (12)	0.53	60
Water for Niger	x	12 (62)	5 (33)	11 (23)	5 (17)	0.28	79
Total	149	188	179	227	255		

* The knowledge score for the option "where most needed" was taken from Experiment 1, knowledge score of Unicef.

Table 3.5. Donation decisions by knowledge and number of alternatives in Experiment 2

In terms of specific pairwise contrasts and effect sizes, we find that when participants were constrained to select a single option, the mean donation in condition Single_7 is greater than in condition Only_1 (188 vs. 149, z = 1.9, p = .06, Cohen's d = .27). The mean for condition Single_13 at 179 is not statistically different than those for Only_1 and Single_7. However, when the mean for condition Only_1 is compared with those for conditions Multiple_7 (227) and Multiple_13 (255), the differences are significant with larger effect sizes (z = 3.3, p = .001, Cohen's d = .47 and z = 4.7, p = .001 Cohen's d = .67 respectively). Given the structural similarity of these conditions to NGO_3, NGO_8 and NGO_16, these last results echo the findings of Experiment 1.

The effect of allowing donors to distribute their contributions over the available options can be further observed in Table 3.4 where we provide data characterizing individual contributions. Similar to Experiment 1, as the number of potential recipients rises, so does the proportion of participants who donate their total endowment of 500 points – from 8% (Only_1) to 22% (Multiple_7) and 18% (Multiple_13). (The difference between Multiple_7 and Only_1 is significant, z = 2.8, p = .01 and so is the difference between Mutiple_13 and Only_1, z = 2.2, p = .03). Moreover, note that whereas 31% of participants donate nothing when there is only one option, this figure drops to 20% and 15% for the cases with 7 and 13 Multiple alternatives (the difference is significant for conditions Only_1 and Multiple_13, z = 2.73, p = .01).

Table 3.6. Unicef campaigns across conditions in Experiment 2

	Only 1	Single 7	Single 13	Multiple 7	Multiple 13
% of participants giving equal non-zero amounts	x	x	x	6	4
% of participants giving away 0 points	31	22	26	20	15
% of participants giving away all 500 points	8	9	10	22	18

The data of Experiment 2 appears to reject *H3a*, the hypothesis that the variability of donations increases with numbers of alternatives. At the aggregate level, the variances of the contributions to the different campaigns are 1209, 403, 1548 and 538 in conditions Single_7, Single_13, Multiple_7 and Multiple_13, respectively. The F-tests for the difference in variances between Single_7 and Single_13 ($F(6, 12) = 3.00$, $p = .05$) and between Multiple_7 and Multiple_13 ($F(6, 12) = 2.88$, $p = .06$) indicate that the variability of donations decreases as the number of alternatives increases thereby supporting *H3b*. Moreover, a two-way factorial analysis of variance on variances of individuals' donations shows that the negative effect of number of available campaigns on the variance of donations is again significant ($F(1, 400) = 15.84$, $p = .001$), whereas neither the effect of using a drop down menu, nor the effect of the interaction term is significant.

In terms of the distribution of donations, each campaign, including the option "where most needed", suffers reductions in both absolute terms and in shares within total donations as the number of alternatives increases.

3.6. Discussion

We conducted two experiments that investigated effects on charitable donations when these are allocated to varying numbers of recipients. The tasks in our experiments differed in two ways. In one, recipients were different NGOs; in the other, recipients were different campaigns of the same NGO. Unlike the former, the latter also involved conditions that limited donors to allocating their whole donation to one of several recipients.

We hypothesized that better known recipients would receive more donations than lesser-known recipients (*H1*). Both in Experiment 1 and 2, we showed this to be the case at both the aggregate and individual levels.

To measure knowledge of NGOs and campaigns, we explicitly adopted a simple strategy of only asking our respondents whether they had heard of these (on a scale from "not having heard" to "well known"). We did not inquire about the nature of respondents' knowledge or attitudes. Moreover, we used knowledge scores as a proxy for respondents' assessments of the merits of NGOs and campaigns (appealing to the recognition heuristic, Goldstein & Gigerenzer, 2002). Clearly, however, the fact that a respondent is knowledgeable about an NGO does not necessarily imply a positive attitude. It would be appropriate to elicit knowledge in a more complete manner in future research.

One intriguing finding was the apparent interaction between knowledge and number of potential recipients as the latter increases. Consider the donations made in Experiment 1 to the three NGOs in condition NGO_3, namely Unicef, Oxfam, and Mercy Corps. In condition NGO_3, two well-

known NGOs, Unicef and Oxfam, receive large mean donations (100 and 83), and even the little known Mercy Corps receives 53. As the numbers of recipients increase, Unicef – the best known NGO – maintains its share of total donations (some 40%) and so benefits in absolute terms as overall donations grow. On the other hand, both Oxfam and Mercy Corps see reductions. In the case of Mercy Corps, the drop-off is dramatic: from 53 (NGO_3) to 16 (NGO_8) to 0 (NGO_16).

The data of both experiments support our second hypothesis that donations increase with the number of potential recipients, but at a decreasing rate. In Experiment 1, there is a 33% increase in mean donations as the number of recipients increases from three to eight (236 to 314), and a 38% increase from three to 16 (236 to 326). In the Multiple condition of Experiment 2, the increase from a single recipient to seven is 52% (149 to 227), and 71% from the single to 13 recipients (149 to 255). These are important results from both theoretical and practical perspectives.

One of the rationales underlying *H2* is that the presence of recipients is a cue to need and that respondents are sensitive to this. Indeed, the results of our two surveys with undergraduate students suggested that there is a relation in people's minds between need and numbers of NGOs. However, we neither measured nor manipulated need independently in our experiments and thus cannot rule out the possibility that some other explanation drives the increases in donations that we observed. On the other hand, our assumption that people gain more utility from being more generous is similar to that of Andreoni (2007) who – subject to one exception – observed behavior similar to our results in the setting of an experimental economics game.

Andreoni's (2007) model predicts that, when the number of recipients increases, those recipients who are present in the different conditions each receive smaller donations (even though total donations increase). This is precisely the pattern of results we observed in Experiment 2. However, in Experiment 1, Unicef (the best-known NGO) was an exception to the rule in that, as the number of recipients increased, so did the donations it received. It is possible that respondents view donating to NGOs differently from donating to campaigns and this possibility should be investigated in future research.

Although not explicitly related to *H2*, the finding in Experiment 2 that donations were greater when respondents could give to several recipients as opposed to being limited to a single option is important. In particular, it suggests that NGOs should consider revising the current design of the drop down menus of their online sites. Of course, one difference between our experimental set up and the online sites of NGOs is that in the Single conditions we did not allow respondents to access the list of potential recipients more than once. It is an open empirical question as to whether the procedures used by NGOs do in fact discourage potential donors from engaging in repeated interactions with drop down menus.

Hypothesis 3 considers the possibility that as the number of recipients increases so does the variability in donations. We framed this question as involving the extent to which respondents – in attempting to be fair – place more or less weight on considerations of equality as opposed to equity as numbers of recipients change. The results of Experiment 1 are ambiguous in that whereas some measures support more variance as numbers of alternatives increase, others suggest no difference. On the other hand, in Experiment 2 variance in donations decreases as the number of alternatives increases. Once again, we are led to suspect that

people think differently about donations to NGOs and donations to campaigns.

Figure 3.3 summarizes our results by showing donation amounts across the eight experimental conditions of our two experiments. The innovation of the present work is to consider how the number of potential recipients affects donation decisions in terms of both amounts and distributions across alternatives. That there are such effects is important from both theoretical and practical viewpoints. From a theoretical perspective our approach can be described as cognitive in nature. It does not account for emotional considerations that have been shown to be important in donation decisions (Dickert, Sagara & Slovic, in press). Thus extending our work to incorporate the effects of emotional influences is an important task for future research.

At a practical level, our results emphasize the importance of the reputation of NGOs and the size of the markets in which they compete for funds. If market size is captured by the number of potential recipients, then it pays for leading NGOs to seek large, competitive "markets". Lesser known NGOs, however, should avoid competition. On the other hand, featuring multiple campaigns is beneficial for resource generation, so long as donors are not constrained to a single option when making a contribution. Given that almost all NGOs employ such limitations in their current online sites and donation interfaces, our results have implications for improving processes of resource generation.

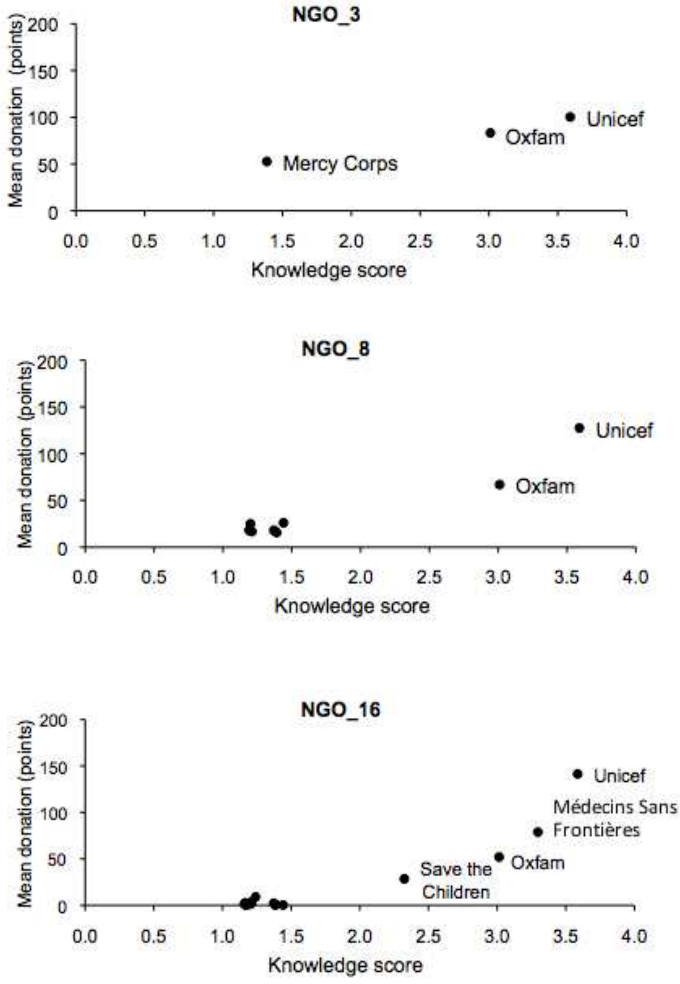


Figure 3.3. Visualization of donations made to recipients across all eight experimental conditions (cont'd on the next page).

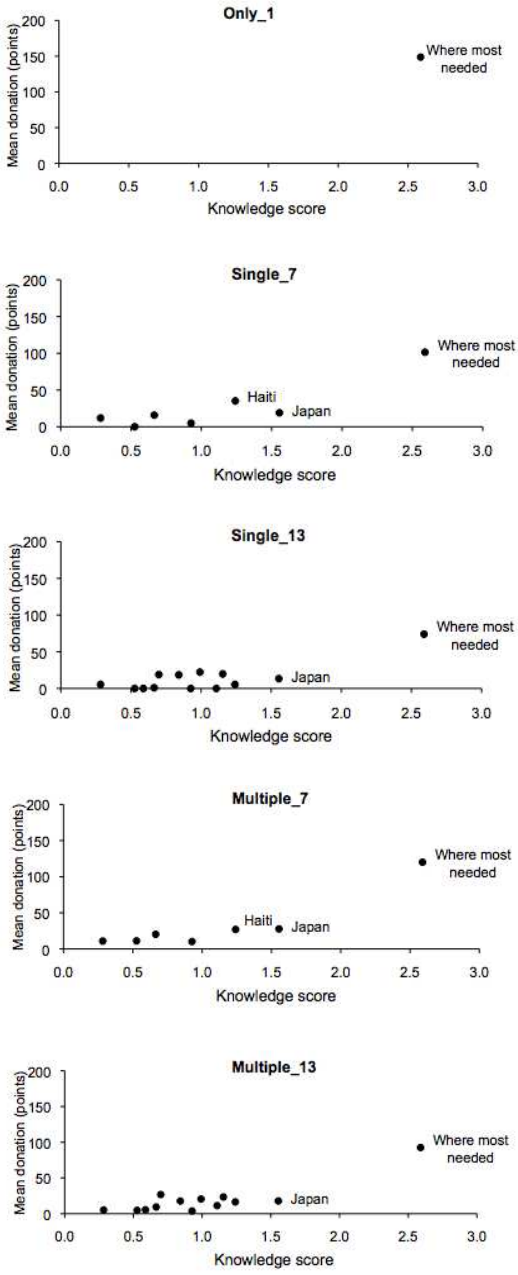


Figure 3.3. Continued

REFERENCES

Andreoni, J. (2007). Giving gifts to groups: How altruism depends on the number of recipients. *Journal of Public Economics*, 91, 1731-1749.

Armstrong, J. S. (2007). Significance tests harm progress in forecasting. *International Journal of Forecasting*, 23, 321-327.

Armstrong, J. S. (2012). Illusions in regression analysis. *International Journal of Forecasting* (in press).

Baltagi, B. H. (2007). Worldwide econometrics rankings: 1989-2005. *Econometric Theory*, 23(5), 952-1012.

Baron, J., & Szymanska, E. (2010). Heuristics and biases in charity. In D. Oppenheimer & C. Olivola (Eds). *The science of giving: Experimental approaches to the study of charity* (pp. 215-236). New York: Taylor and Francis.

Bekkers, R. (2007). Measuring altruistic behavior in surveys: The all-or-nothing dictator game. *Survey Research Methods*, 1, 139-144.

Bertini, M., Wathieu, L., & Iyengar, S. S. (2010). *The discriminating consumer: Product proliferation and willingness to pay for quality*. Working paper, London Business School.

Betsch, T., Biel, G.M., Eddelbüttel, C., & Mock, A. (1998). Natural sampling and base-rate neglect. *European Journal of Social Psychology*, 28, 269-273.

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, 15, 284-289.

Camerer, C. F. & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty*, 19, 7-42.

Camerer, C. F. (2000). Prospect theory in the wild: Evidence from the field. In D. Kahneman & A. Tversky (Eds.), *Choice, Values, and Frames* (pp. 288-300). New York, NY: Russell Sage Foundation & Cambridge University Press.

Cameron, C. D., & Payne, B. K. (2011). Escaping affect: How motivated emotion regulation creates insensitivity to mass suffering. *Journal of Personality and Social Psychology*, 100, 1-15.

Carhart, M. (1997). On persistence in mutual fund performance. *Journal of Finance*, 52(1), 57-82.

Carroll, L. S., White, M. P., & Pahl, S. (2011). The impact of excess choice on deferment of decisions to volunteer. *Judgment and Decision Making*, 629-637.

Chang, W. (2005). Determinants of donations: Empirical evidence from Taiwan. The Developing Economics, 43, 217-234.

Christensen-Szalanski, J. J. J., & Beach, L.R. (1982). Experience and the base-rate fallacy. *Organizational Behavior and Human Performance*,29, 270-278.

Cobos, P. L., Almaraz, J., & García-Madruga, J. A. (2003). An associative framework for probability judgment: An application to biases. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,29, 80-96.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, 58, 1-73.

Dickert, S., Sagara, N., & Slovic, P.. Affective motivations to help others: A two-stage model of donation decisions. *Journal of Behavioral Decision Making* (in press).

Dinov, I. D., Sanchez, J.,& Christou, N. (2008). Pedagogical utilization and assessment of the statistic online computational resource in introductory probability and statistics courses. *Computers and Education*, 50, 284-300.

Edgell, S. E., Harbison, J. I., Neace, W. P., Nahinsky, I. D., & Lajoie, A. S. (2004). What is learned from experience in a probabilistic environment? *Journal of Behavioral Decision Making*, 17, 213-229.

Einhorn, H. J., & Hogarth, R. M. (1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology*, 32, 53-88.

Elwin, E., Juslin, P., Olsson, H., & Enkvist, T. (2007). Constructivist coding: Learning from selective feedback. *Psychological Science*, 18, 105-110.

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological Research*, 50, 123-129.

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

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.

Fiedler, K., & Juslin, P. (2006). (Eds.) *Information sampling and adaptive cognition.* New York, NY: Cambridge University Press.

Fiedler, K., & Unkelbach, C. (2011). Lottery attractiveness and presentation mode of probability and value information. *Journal of Behavioral Decision Making*, 24, 99-115.

Fox, C. R., & Hadar, L. (2006). "Decisions from experience" = sampling error + prospect theory: Reconstructing Hertwig, Barron, Weber, & Erev (2004). *Judgment and Decision Making*, 1, 159-161.

Franco-Watkins, A. M., Derks, P. L., & Dougherty, M. R. P. (2003). Reasoning in the Monty Hall problem: Examining choice behaviour and probability judgements. *Thinking and Reasoning*, 9, 67-90.

Friedman, D. (1998). Monty Hall's three doors: Construction and deconstruction of a choice anomaly. *American Economic Review*, 88, 933-946.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instructions: Frequency formats. *Psychological Review*, 102, 684-704.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke. E., Schwartz, L. M, & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8(2), 53-96.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, 109, 75-90.

Goldstein, D. G. & Taleb, N. N. (2007). We don't quite know what we are talking about when we talk about volatility. *Journal of Portfolio Management*, 33(4), 84-86.

Goldstein, D. G., Johnson, E. J., & Sharpe, W. F. (2008). Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *Journal of Consumer Research*, 35, 440-456.

Granberg, D., & Brown, T. A. (1995). The Monty Hall dilemma. *Personality and Social Psychology Bulletin*, 21, 711-723.

Granberg, D., & Dorr, N. (1998). Further exploration of two-stage decision making in the Monty Hall dilemma. *American Journal of Psychology*, 111, 561-579.

Greene, W. H. (2003). *Econometric Analysis*, 5th edition. Upper Saddle River NJ: Prentice Hall.

Gujarati, D. N. & Porter, D. (2009). *Basic Econometrics*. McGraw-Hill Irwin, New York.

Haisley, E., Kaufmann, C., & Weber, M. (2010). *The role of experience sampling and graphical displays on one's investment risk appetite*. New Haven, CT: Yale School of Management working paper.

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General*, 108, 356-358.

Hasher, L., & Zacks, R. T. (1984), Automatic processing of fundamental information: The case of frequency occurrence. *American Psychologist*, 39, 1372-1388.

Hertwig, R., & Gigerenzer, G. (1999). The "conjunction fallacy" revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12, 275-305.

Hertwig, R., Davis, J. N., & Sulloway, F. J. (2002). Parental investment: How an equity motive can produce inequality. *Psychological Bulletin*, 128, 728-745.

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.

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, 73, 538-540.

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, 290, 2261-2262.

Hogarth, R. M. (Ed.) (1982). *Question framing and response consistency: New directions for methodology of social and behavioral science.* San Francisco, CA: Jossey-Bass.

Hogarth, R. M. (2001). *Educating intuition.* Chicago, IL: The University of Chicago Press.

Hogarth, R. M., & Soyer, E. (2011). Sequentially simulated outcomes: Kind experience vs. non-transparent description. *Journal of Experimental Psychology: General*, 140, 434-463.

Hogarth, R. M., Mukherjee, K., & Soyer, E. (2011). Assessing the chances of success: Overconfident or just confused? *Journal of Experimental Psychology: Learning, Memory and Cognition* (in press).

Holyoak, K. J., & Spellman, B. A. (1993). Thinking. *Annual Review of Psychology*, 44, 265-315.

Hsee, C. K., & Rottenstreich, Y. (2004). Music, pandas, and muggers: On the affective psychology of value. *Journal of Experimental Psychology: General*, 133, 23-30.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, 125, 576-590.

Hubbard, R., & Armstrong, J. S. (1994). Replications and extensions in marketing - Rarely published but quite contrary. *International Journal of Research in Marketing*, 11, 233-248.

Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79, 995-1006

Iyengar, S. S., Huberman, G., & Jiang, W. (2004). How much choice is too much? Contributions to 401(k) retirement plans. In O. S. Mitchell & S. Utkus (Eds.), *Pension design and structure: New lessons from behavioral finance* (pp. 83-95). Oxford, UK: Oxford University Press.

Jensen, M. C. (1968). The performance of mutual funds in the period 1945-1964. *Journal of Finance*, 23(2), 389-416.

Judge, G. G., Griffiths, W., Hill, C. R., & Lee T. C. (1985). *Theory and Practice in Econometrics*. Wiley, New York.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47, 263-291.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19, 1-53.

Kogut, T., & Ritov, I. (2005a). The singularity effect of identified victims in separate and joint evaluations. *Organizational Behavior and Human Decision Processes*, 97, 106-116.

Kogut, T., & Ritov, I. (2005b). The "identified victim" effect: An identified group, or just a single individual? *Journal of Behavioral Decision Making*, 18, 157-167.

Kunreuther, H. & Michel-Kerjan, E. (2010). Market and Government Failure in Insuring and Mitigating Natural Catastrophes: How Long-Term Contracts Can Help in *Public Insurance and Private Markets*, Jeffrey R, Brown, (ed.) AEI Press.

Krauss, S., & Wang, X. T. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General*, 132, 3-22.

Landry, C. E., Lange, A., List, J. A., Price, M. K., & Rupp, N. G. (2006). Toward an understanding of the economics of charity: Evidence from a field experiment. *The Quarterly Journal of Economics*, 121, 747-782.

Lawrence, M. & Makridakis, S. (1989). Factors affecting judgmental forecasts and confidence intervals. *Organizational Behavior and Human Decision Processes*, 42, 172-187.

Lejarraga, T. (2010). When experience is better than description: Time delays and complexity. *Journal of Behavioral Decision Making*, 23, 100-116.

Martin, R., & Randall, J. (2008). How is donation behavior affected by the donations of others? *Journal of Economic Behavior & Organization*, 67, 228-238.

McCloskey, D. N., & Ziliak, S. T. (1996). The standard error of regressions. *Journal of Economic Literature*, 34, 97-114.

Nilsson, H. (2008). Exploring the conjunction fallacy within a category learning framework. *Journal of Behavioral Decision Making*, 21, 471-490.

Ord, K. (2012). The illusion of predictability: A call to action. *International Journal of Forecasting* (in press).

Real, L. A. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, 253, 980-986.

Real, L. A. (1996). Paradox, performance, and the architecture of decision-making in animals. *American Zoologist*, 36, 518-529.

Reeves, T., & Lockhart, R. S. (1993). Distributional versus singular approaches to probability and errors in probabilistic reasoning. *Journal of Experimental Psychology: General*, 122, 207-226.

Reutskaja, E., & Hogarth, R. M. (2009). Satisfaction in choice as a function of the number of alternatives: When "goods satiate". *Psychology & Marketing*, 26, 197-203.

Sarbagh, C., Dar, Y., & Resh, N. (1994). The structure of social justice judgments. Social Psychology Quarterly, 57, 244-261.

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2009). What moderates the too-much-choice effect? *Psychology & Marketing*, 26, 229-253

Scheibehenne, B., Greifeneder, R., & Todd, P. M. (2010). Can there ever be too many options? A meta-analytic review of choice overload. *Journal of Consumer Research*, 37, 409-425.

Schotter, A., & Sopher, B. (2003). Social learning and coordination conventions in intergenerational games: An experimental study. *Journal of Political Economy*, 111, 498-529.

Schwab, A., & Starbuck, W. H. (2009). Null-hypothesis significance testing in behavioral and management research: we can do better. In D. Bergh & D. Ketchen (Eds.), *Research Methodology in Strategy and Management*, 5, 29-54. Oxford, UK: Elsevier.

Schwartz, B. (2004). *The paradox of choice: Why more is less*. New York: Eco/Harper-Collins Publishers.

Schweitzer, F. & Mach, R. (2008). The epidemics of donations: Logistic growth and power-laws. PLoS ONE, 3: e1458. *doi:10.1371/journal.pone.0001458*

Sedlmeier, P. (1998). The distribution matters: Two types of sample-size tasks. *Journal of Behavioral Decision Making*, 11, 281-301.

Sedlmeier, P. (1999). *Improving statistical reasoning: Theoretical models and practical applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sedlmeier, P. (2000). How to improve statistical thinking: Choose the task representation wisely and learn by doing. *Instructional Science*, 28, 227-262.

Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, 130, 380-400.

Shah, A. M., & Wolford, G. (2007). Buying behavior as a function of parametric variation of number of choices. *Psychological Science*, 18, 369-370.

Shanks, D. R. (1991). On similarities between causal judgments in experienced and described situations. *Psychological Science*, 2, 341-350.

Simon, H. A. (1978). Rationality as process and product of thought. *American Economic Review*, 68(2), 1-16.

Small, D. A., Loewenstein, G. & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes*, 102, 143-153.

Soyer, E., & Hogarth, R. M. (2011). The size and distribution of donations: Effects of number of recipients. *Judgment and Decision Making*, 616-628.

Soyer, E., & Hogarth R. M. (2012a). The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting* (in press).

Soyer, E. & Hogarth, R. M. (2012b). Response to commentaries on "The illusion of predictability: How regression statistics mislead experts." *International Journal of Forecasting* (in press).

Taleb, N. N., & Goldstein, D.G. (2012). The problem is beyond psychology: The real world is more random than regression analyses. *International Journal of Forecasting* (in press).

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. New Haven, CT: Yale University Press.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-1131.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90, 293-315.

Tversky, A., & Kahneman, D. (1986). Rational choice and the framing of decisions. *Journal of Business*, 59, Part 2, S251-S278.

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

Wooldridge, J. M. (2008). *Introductory Econometrics: A Modern Approach.* International Student Edition (3rd), Thomson, South Western.

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Bestch (Eds.) *Etc.frequency processing and cognition* (21-36), New York, NY: Oxford University Press.

Zellner, A. (1984). Posterior odds ratios for regression hypotheses: General considerations and some specific results. In A. Zellner (Ed.), *Basic Issues in Econometrics*, (pp. 275-305). Chicago, IL: The University of Chicago Press.

Zellner, A. (2004). To test or not to test and if so, how? Comments on "size matters". *Journal of Socio-Economics*, 33, 581-586.

Ziliak, S. T., & McCloskey, D. N. (2004). Size matters: the standard error of regressions in the American Economic Review. *Journal of Socio-Economics*, 33, 527-546.

Ziliak, S. T., & McCloskey, D. N. (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives.* Ann Arbor: University of Michigan Press.

Ziliak, S. T. (2012). Visualizing uncertainty: On Soyer's and Hogarth's The illusion of predictability: how regression statistics mislead experts. *International Journal of Forecasting* (in press).