On the Significance of Information

DOCTORAL THESIS

to acquire the academic degree of

doctor rerum politicarum
(Doctor of Economics and Management Science)

submitted to the

School of Business and Economics of
Humboldt-Universität zu Berlin

by

Perke Jacobs, M.Sc.

President of Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dean of the School of Business and Economics:
Prof. Dr. Daniel Klapper

Reviewers: 1. Prof. Dr. Dirk Engelmann
2. Prof. Dr. Gerd Gigerenzer

Date of Colloquium: May 26, 2021


HUMBOLDT-UNIVERSITÄT ZU BERLIN

Perke Jacobs

On the Significance of Information

May 2021

to my family

Forrest Gump' mama said: Life is like a box of chocolates;
my mama told me: Go to school, get your doctorate,
something to fall back on, you could profit with;
but still supported me when I did the opposite.

— from a song

Contents

Acknowledgements

[removed for privacy]

Summary

This doctoral thesis emphasizes the significance of informational conditions in studying economic decisions. The first chapter concerns access to accurate probabilistic information in the domain of medical interventions. We estimate that, in many developed countries, there appears to be demand for governments to grant citizens free access to impartial reviews of medical evidence, as provided in Cochrane Reviews. For these countries, we estimate that this demand can be met at low costs. The second chapter concerns the communication of such information and examines the facilitating effect of natural frequencies on the derivation of posterior probabilities, as delineated by Gigerenzer and Hoffrage (1995). In a meta-analysis, we clarify concepts and disentangle the effects of 15 study characteristics. We find that in the simplest study design, 4 percent correct solutions when presented with conditional probabilities and 24 percent when presented with natural frequencies. The final two chapters examine the satisficing class of strategies for uncertain decision environments in which agents lack a full probabilistic description of the decision problem. According to Simon (1955), satisficing strategies use aspiration levels to terminate search for suitable alternatives. The third chapter describes how in economics, satisficing is modeled as a preference structure or as a decision rule that yields choices inferior to utility maximization, whereas in cognitive science, satisficing strategies use aspiration levels to solve inference problems. We explain the divergence, noting that they refer to risky and uncertain environments, respectively. The final chapter examines satisficing in an applied setting, studying taxi drivers' shift termination behavior. We find that drivers' hourly earnings are very uncertain and drivers' behavior is best predicted by simple satisficing strategies that terminate shifts when reaching an aspired shift income or shift duration.

Zusammenfassung

Diese Dissertation unterstreicht die Rolle von Informationen bei der Erforschung ökonomischer Entscheidungen. Das erste Kapitel beschäftigt sich mit dem Zugang zu akkuraten Informationen über die Wirksamkeit medizinischer Interventionen. Unserem Model zufolge besteht in vielen entwickelten Ländern Nachfrage für Zugang zu unabhängigen medizinischen Informationen wie Cochrane Reviews. Wir schätzen, dass für viele Länder diese Nachfrage zu moderaten oder geringen Kosten erfüllt werden kann. Das zweite Kapitel beschäftigt sich mit der Kommunikation solcher Informationen und untersucht den unterstützenden Effekt natürlicher Häufigkeiten bei der Berechnung von A-posteriori Wahrscheinlichkeiten (Gigerenzer & Hoffrage, 1995). Durch eine Meta-Analyse erklären wir Konzepte und entflechten die Effekte von 15 Studienmerkmalen. Im einfachsten Studiendesign führen natürliche Häufigkeiten zu 24 Prozent korrekten Antworten verglichen mit 4 Prozent bei konditionellen Wahrscheinlichkeiten. Die finalen beiden Kapitel analysieren Satisficing-Strategien für unsichere Entscheidungsumgebungen in denen Agenten eine vollumfassende, probabilistische Beschreibung des Entscheidungsproblems fehlt. Simon (1955) zufolge nutzen Satisficing-Strategien ein Anspruchsniveau um die Suche nach weiteren Entscheidungsalternativen zu beenden. Das dritte Kapitel beschreibt wie solche Strategien in der ökonomischen Literatur als Präferenz modelliert werden um Entscheidungen zu erklären die der Nutzenmaximierung unterlegen sind während die Kognitionswissenschaften diese Strategien als Lösungen für Inferenzprobleme betrachten. Wir erklären die Divergenz mit unterschiedlichen Annahmen über die vorliegenden Informationen der Agenten. Das letzte Kapitel untersucht Satisficing-Strategien unter Taxifahrern. Wir stellen fest, dass die Stundenlöhne von Taxifahrern kaum vorhersagbar sind und ihre Entscheidungen Schichten zu beenden am besten durch einfache Satisficing-Strategien vorhergesagt werden können.

Introduction

This text is an appreciation of information. As a doctoral thesis, it concludes more than twenty years of formal education. The privilege of having such access to information cannot be overstated. The title of this thesis is an acknowledgement of this privilege while also synthesizing the four independent chapters of this text. Jointly, these chapters emphasize the significance of information in studying economic decisions.

Information has been studied broadly in economics and touches on many of the discipline's fundamental topics. For example, decisions under risk are fundamentally characterized by the absence of information. Instead of knowing future states of the world with certainty, agents know their probabilities under each available path of action. For this reason, theories of risky choice, including expected utility theory (Bernoulli, 1738/1954), can plausibly be called theories of imperfect information. Later, Frank Knight (1921) introduced uncertainty as a distinct, more severe category of information shortage. In contrast to risk, uncertainty refers to agents lacking a complete description of the probabilistic structure of the environment. For example, Herbert Simon (1955, 1956) modeled agents searching for suitable paths of action with limited knowledge of possible alternatives, let alone their consequences. In the second half the twentieth century, the effects of imperfect information were studied in more diverse settings. In individual choice, George Stigler (1961) picked up Simon's focus on search and modeled a situation in which a consumer is aware of the distribution of prices for cars but needs to find the seller offering the best car price. In macroeconomics, George Akerlof (1970) famously pointed out that used cars are usually "lemons", as

information of the quality of the car is private to sellers. Consequently, consumers are ultimately unwilling to pay more than the price of a low-quality car, driving high-quality cars out of the market. In labor economics, Michael Spence (1973) described how employers lack information on prospective employees' skills. Employees can therefore use education as a costly signal of their skills. Finally, in a similar case of asymmetric information, Joseph Stiglitz and Andrew Weiss (1987) describe a "screening" process to overcome the problem of sellers who are unaware of customers' willingness to pay. By this process, firms set prices and conditions in such a way so as to induce customers to reveal their true willingness to pay. This brief summary of the work by some of the twentieth century Nobel laureates does not do justice to the range of informational conditions addressed by contemporary economic theory. Nonetheless, its diversity serves to demonstrate that the topic of information has moved out of the "slum dwelling in the town of economics" Stigler had seen it occupying in 1961.

At the same time, building on this analogy, information does not yet reside on the hillside where every suburb of the discipline can see it. In an attempt to broaden the consideration of information within economics, the focus of this text is on the *quality* of information, in contrast to most existing work that has focused more dichotomously on the *presence* of specific pieces of information. To be effective, information must not only be available but also useful. This aspect is often overlooked in existing research, particularly in much of behavioral economics. I argue that the quality of information affects behavior in different ways, which I address within the four chapters of this text. These chapters are organized around Knight's two categories of information shortage, with the first two chapters concerning decisions under risk and the final two concerning decisions under uncertainty.

Risky environments lack information about future states of the world but offer agents a full probabilistic description of the decision problem. The textbook example is a game of chance, such as a coin that is tossed or a dice that is thrown. In such simple cases, agents are informed that the probability of a specific event is one in two or one in six and agents can integrate this information with their valuation of possible outcomes to reach a decision. More applied problems include the probability of rain or the probability of a specific side effect of some medication. In these problems, all relevant probabilities

are reported[1]. However, many applied problems are more complex.

Consider, for example, a screening test to detect a disease. Often, screening tests use a marker to detect a disease before a patient becomes symptomatic. Screening tests are not without error and both a negative and a positive test result can be incorrect, referred to as a miss and a false-alarm, respectively. Consider an agent who has taken such a screening test. To interpret the result, the agent needs to derive the relevant posterior probability of having the disease from the test's specifications. Accordingly, the agent consults information on the test's miss and false-alarm rates. To be useful, this information needs to fulfill at least two conditions, addressed in chapters one and two.

First, information about the test's miss and false-alarm rates needs to be accurate and accessible. Particularly in the medical domain, probabilities of side effects or test outcomes are estimated from randomized controlled trials and often change as new data become available. In addition, the structures of many markets for health services and products do not incentivize providers to inform patients neutrally and based on evidence. Therefore, consumers often find themselves faced with selective and distorted information. To address both of these problems, the Cochrane Collaboration publishes and updates meta-analyses and lay summaries of the effects of several thousand health interventions. These are written by some 30,000 medical scholars and epidemiologists to inform doctors and patients alike. Downloads of Cochrane Reviews are costly at about $38 per review, unless they are accessed from a country with a national subscription to the Cochrane Library. To date, few countries have subscribed to the Cochrane Library, possibly owing to uncertainty about whether it would be used.

The first chapter estimates demand for Cochrane Reviews as if they were free in countries that have not already subscribed nationally. To this end, we aggregate web-traffic data from Cochrane websites to obtain the number of downloads per country. We use conventional OLS regression to disentangle the effect of a national subscription from that of alternative factors, including those related to language and development. Using these estimates, we calculate for all members of the Organization for Economic Cooperation and Development

[1]Note, however, the long and ongoing debate about the application of relative frequencies to single events (compare e.g., Savage, 1954; von Mises, 1957)

(OECD) the expected number of review downloads if they were to offer a national subscription. We find that, for the vast majority of member states, the number of review downloads would increase under a national subscription, and for some they would more than double. Next, we calculate estimates of the effective price per additional download. Using these estimates, we identify three groups of countries: (1) About half a dozen countries for which a national subscription is unlikely to be worthwhile, (2) the majority of countries for whom additional downloads would cost less than $2 each, and (3) three countries that would save money under a national subscription. Thus, for many countries a national subscription appears to offer a promising means to disseminate medical evidence and empower agents to base their decisions on the best information available.

Second, information about the test's specifications needs to be intuitively understandable. Interpreting a test result from the test's miss and false-alarm rates can be difficult, even for those familiar with Bayes' theorem, including economists (Berg, Biele, & Gigerenzer, 2016). Apart from the miss rate, p(negative result|disease) and false-alarm rate p(positive result|no disease), the disease's base rate p(disease) is required to obtain the predictive value of the positive or negative result p(disease|result) using Bayes' equation. Gerd Gigerenzer and Ulrich Hoffrage (1995) reported that only 16 percent of their participants were able to solve a problem when given information in conditional probabilities. The authors hypothesized that if test specifications were reported as *natural frequencies*, more agents would be able to correctly derive the relevant posterior probability. Natural frequencies report frequencies analogously to their natural occurrence. In the case of a screening test, the miss rate could be reported as "x in y people with the disease receive a negative test result". If x and y are chosen such that y is the number of people out of 100 who have the disease, this format maintains the base rate and aides the decision. Indeed, Gigerenzer and Hoffrage (1995) reported that under this format, 46 percent of their participants were able to solve the problem correctly, a finding referred to as the natural frequency facilitation effect.

The second chapter comprises a meta-analysis of the natural frequency facilitation effect. To account for the effect, various explanations and experimental designs were put forth, resulting in a fragmented body of evidence. For example, some studies used stricter criteria than others in scoring correct solutions, and some studied lay

persons whereas others studied participants with above-average statistical literacy. In addition, some authors mistook natural frequencies to mean any frequency representation. This chapter reviews this literature, clarifies concepts, and analyzes 226 estimates of performance under conditional probability and the natural frequency formats. In addition, we characterized each study along 15 dimensions pertaining to methodology, participants, and experimental setup. We then use a meta-analytic model to disentangle the effects of the various moderators on performance. We find that, in its simplest form we would, on average, expect 24 percent of those presented with natural frequencies and 4 percent of those presented with conditional probabilities to solve such problems correctly. This implies an odds ratio of 7.10, which is a sizable effect. Further, methodological choices such as showing participants multiple problems of the same format or of different formats are estimated to affect the performance under both formats. The strongest moderator, overall and conceptually, was the presentation of a visual aid, which increased performance under both formats by 22 and 23 percentage points, on average. Both the confirmation of the natural-frequency facilitation effect and the additional effect of the visual aid illustrate how probabilistic choice requires comprehensible information and how adequate communication can assist such decisions.

To summarize, the first two chapters maintain that good choices are not only a function of the decision strategy but also of the decision environment. By this account, risky choice often requires probabilistic information to be accurate and available in a form that allows agents to intuitively extract the relevant pieces. These conclusions, although fairly straightforward, are difficult to draw from conventional economic theory. Expected utility theory explains differences in choices by differences in preferences, assuming that agents are capable of obtaining the necessary information without loss of accuracy. Much of the work in behavioral economics has challenged this assumption, demonstrating that human judgements are not necessarily flawless (e.g., Kahneman & Tversky, 1972). However, this literature concludes that flaws are rooted in the human mind. The view advocated in these two chapters is that such conclusions may or may not be warranted, but studies of this issue need to go beyond probability trick questions and consider the environmental factors that shape the performance of the mind.

Next, we turn our attention to uncertain decision environments.

In contrast to risky environments, uncertain decision environments do not supply agents with the information necessary for an exhaustive probabilistic description of the decision problem. Under such circumstances, agents need to use the available information as best as they can to reach a decision. Unfortunately, similar to the arguments made in the previous two chapters, the best use of the available information depends on its quality. Consider, for example, a situation where the agent is aware of all possible alternatives and their consequences but unaware of the probabilities of these consequences. With sufficient information available, the agent can estimate the probabilities and treat the uncertainty as risk. In contrast, suppose there exists a subset of consequences that the agent is unaware of. In this situation, viewing the decision problem as a risky choice task and applying some form of optimization on the known portion of the environment is not guaranteed to be the most profitable decision strategy. Instead, it is possible that alternative decision strategies that ignore part of the information, *heuristics*, yield better decisions. In his Nobel Prize acceptance speech, Herbert Simon (1979, p.489) succinctly pointed out that uncertain agents can decide "either by finding optimal solutions for a simplified world, or by finding satisfactory solutions for a more realistic world. Neither approach, in general, dominates the other [...]."

The third chapter elaborates on one particular class of strategies for "finding satisfactory solutions for a more realistic world". Simon (1955) described a satisficing strategy that uses aspiration levels to find a suitable alternative. In this model, agents search for alternatives until one meets all aspiration levels formulated by the agent. Aspiration levels are defined directly on the relevant decision attributes rather than a composite, such as a utility function. Consider, for example, an agent looking for a new car. One possible satisficing strategy would be to formulate aspiration levels on trunk size, fuel consumption, and annual costs, and then examine cars one by one until one meets all three aspiration levels. Despite its simplicity, two different literatures have evolved around the concept of satisficing.

On the one hand, the cognitive science literature studies strategies for inference problems. In this type of problem, the agent tries to predict an objective criterion from available information. For example, an agent may guess which of two cars has higher fuel consumption based on information about horse power and weight. In this literature, satisficing strategies use an aspiration level to make

inferences. One possible strategy uses an aspiration level to determine which information about the two cars to consider: If the difference in weight exceeds some aspiration level, the heavier car is predicted to have higher fuel consumption, but if the weight is sufficiently similar, the more powerful car is predicted to consume more fuel (Luan, Schooler, & Gigerenzer, 2014). Analysts can compare the decision of this strategy to the true level of fuel consumption. Therefore, empirical studies of satisficing strategies in this literature examine two questions — the descriptive question of how agents behave and the prescriptive question of which strategy yields the best decision for a particular problem.

On the other hand, the economics literature studies strategies for preference problems. In these problems, agents are confronted with choices between certain or risky alternatives. For example, an agent may decide between two cars, one small but fast and the other large but comparatively slow. In these settings, decisions cannot be right or wrong, as the correct choice depends on the agent's preferences. Therefore, empirical studies of such problems examine only the descriptive question of how agents behave. Variation in behavior between agents is modeled as differences in the preferences of these agents. This literature interprets satisficing as a particular preference structure that is often contrasted with expected utility theory. For example, an agent may be modeled by a utility function defined over the speed and size of a car. A satisficing model can then be distinguished from the neoclassical model by a kink of the utility function, such that marginal utilities (and marginal rates of substitution) change discontinuously at specific points along the utility function. Rational choice in these problems is defined by consistency criteria and attributed to expected utility theory. In contrast, deviations from this ideal, including satisficing, are viewed as inferior.

Depending on the stream of literature, the concept of satisficing refers to different classes of models, methodologies and conclusions about the rationality of satisficing. We review both these literatures and explain how their seemingly opposing approaches and conclusions follow from their assumed degrees of information shortage: Whereas the economics literature focuses on risky conditions, the cognitive science literature is concerned with uncertain environments. We clarify this distinction and describe the informational conditions under which satisficing can be rational.

The final chapter of this thesis combines elements from both streams of the literature and studies preferences under uncertainty. It adds to the existing literature on the wage elasticity of taxi drivers, which was initially found to be negative and taken as evidence of the income-target hypothesis (Camerer, Babcock, Loewenstein, & Thaler, 1997). According to this model, drivers set daily income targets and terminate their shifts as soon as they reach their target. The resulting negative correlation between a shift's length and average hourly wage contradicts the intuition that higher wages should be associated with longer labor supply. Following the approach for risky choice, the existing literature has interpreted this finding in terms of differences in agents' preferences, focusing on comparisons of theories of expected and reference-dependent utility. However, this approach confuses choices under risk and uncertainty.

The chapter presents two analyses. The first examines taxi drivers' decision environment using the Hamburg taxi data, which we acquired for this study. To understand the degree of uncertainty drivers face, we used a range of statistical models to predict drivers' hourly earnings from observables and assessed the quality of these predictions. We find that taxi drivers face severe uncertainty, with root-mean squared errors of the best predicting model only slightly lower than the standard deviation of hourly earnings. Under these circumstances, using a strategy that requires accurate expected values cannot be expected to yield good decisions. In the second analysis, we predict drivers' shift ends using behavioral models. In contrast to earlier studies, we compare both utility models and satisficing heuristics. These heuristics set an aspiration level on shift income, shift duration, clock hour, or the hiatus between trips and terminates the shift as soon as that aspiration level is satisfied. We interpret the choice of the aspiration variable as an expression of which goal the driver ranks highest. In contrast, we understand the aspiration level as the threshold that offers the driver an acceptable balance of time and income. We find that the majority of drivers are best predicted by one of these heuristic models, whereas both utility models predict only a fraction of drivers best.

Parenthetically, the four chapters of this thesis stand in the tradition of Herbert Simon's aspirations for behavioral economics. These aspirations are often described as an empirical test of the assumptions underlying various forms of utility theory. Alternatively, I have tried to make the point that Simon's research agenda can be viewed

as an invitation to analysts to step into the agent's shoes and take seriously their informational conditions. Such a step would offer economics several opportunities, two of which are emphasized in this text. First, there exist many situations beyond lotteries and other games that closely resemble risky choice and present agents with a probabilistic description of their decision environment. Focusing empirical scrutiny on this environment can offer insights that are relevant for improving decisions. After all, any decision can only be as good as the environment affords. Second, there remain many uncertain situations where agents have only partial knowledge of the decision problem. Engaging with these situations, rather than failing to distinguish them from risky ones, offers the opportunity to expand the rich theory of preferences to uncertainty. However, doing so requires additions to the methodological toolbox. Chapters three and four describe how some tools can be borrowed from cognitive science, however, to address preferences under uncertainty more generally, more innovation may be needed. In a letter to John Conlisk in response to his 1996 article on bounded rationality, Simon (1996) remarked that methods are slow to change and also require changes in graduate teaching, but can improve in the long run. He concluded noting that "Keynes had something to say about the long run, and I would very much like still to be alive when this all happens on a major scale. I'm counting on you of the next generation to bring it about" (p.3). Here is my modest contribution.

Chapter 1

What Would be Demand for Cochrane Reviews if Access Was Free?

This chapter is forthcoming as: Jacobs, P. & Gigerenzer, G. (2020), Using Variation Between Countries to Estimate Demand for Cochrane Reviews When Access Is Free: A Cost-Benefit Analysis. *British Medical Journal open*, doi: 10.1136/ bmjopen-2019-033310.

Medical research accounts for a substantial proportion of research and development (R&D) expenditures. In the United States (U.S.), total spending on medical and health R&D increased between 2013 and 2016 to $172 billion, led by industry with 67 percent and the federal government with 22 percent (Research America, 2017). Worldwide, biomedical publications are increasing year by year: for instance, about one million articles are added to this literature annually (Khare, Leaman, & Lu, 2014). Faced with this large volume of articles, no healthcare worker is able to stay fully informed about recent research. The problem of quantity is amplified by one of quality; many of the clinical trials published are unreliable or of uncertain reliability, and most healthcare professionals, including physicians and nurses, do not have the time and/or training to evaluate the quality of a research article (Ioannidis, Stuart, Brownlee, & Strite, 2017). Additionally, direct-to-consumer ads, websites, and television shows compete for the attention of healthcare professionals and patients, disseminating a mix of evidence and unwarranted claims based on commercial interests or personal opinion (Gigerenzer & Goldstein, 2011). In the U.S., an estimated 20 to 50 percent of health care service use is inappropriate, wasteful, or harmful for patients (Ioannidis et al., 2017).

To address these issues, over 10,000 medical researchers have built an international network, named Cochrane after the British epidemiologist Archie Cochrane, to assist healthcare professionals and patients

in making well-informed decisions about healthcare interventions. This network produces systematic reviews of the available evidence on the benefits and harms of medical interventions and tests, such as measles, mumps, and rubella vaccination, check-ups, prostate cancer screening, and statins. Since 1992, these Cochrane Reviews have been written by some 30,000 medical researchers and are generally recognized as the gold standard of medical evidence (Schünemann et al., 2008; Useem et al., 2015). The reviews are intended to be regularly updated as new findings become available. Cochrane Reviews provide three important services for healthcare professionals (Jefferson, Demicheli, Rivetti, & Deeks, 1999). First, they offer an overall assessment of the available evidence by evaluating individual studies according to the quality of their evidence and statistically integrating their results, which often vary due to their small sample sizes. Second, in contrast to a self-survey of the literature, systematic reviews allow professionals to absorb the relevant information about the benefits and harms of specific treatments under the typical conditions of time pressure. Finally, Cochrane Reviews offer plain-language summaries and summary-of-findings tables that highlight key findings and can be easily understood by persons without statistical training, which makes them suitable for both professionals and lay people alike. For these reasons, many professionals consult the Cochrane Reviews regarding interventions. Yet here is where the problem arises.

Whereas plain-language summaries are openly available online, access to the full-text reviews is often restricted, despite their containing large amounts of relevant information for patients and healthcare professionals. Institutions in many low and middle income countries are granted free or inexpensive access through the World Health Organization's Hinari Access to Research for Health Programme (see also `www.who.org/hinari`), but healthcare professionals or patients outside of an institutional context are excluded. Most countries in North America and Europe (including the U.S. and Germany), by contrast, are not eligible and fall into one of two groups: those with and those without a national subscription. The latter group far exceeds the former, with only eight countries subscribing nationally in 2014, six of which are members of the OECD[1]. Specifically, Australia,

[1]The 34 OECD member states are Australia (AUS), Austria (AUT), Belgium (BEL), Canada (CAN), The Czech Republic (CZE), Denmark (DNK), Estonia (EST), Finland (FIN), France (FRA), Germany (DEU), Greece (GRC), Hungary (HUN), Iceland

Denmark, Ireland, Norway, New Zealand, and Great Britain offered free access nationwide, as did Egypt and India, which are not OECD member states. In addition, one U.S. state, Wyoming, and three Canadian provinces, New Brunswick, Nova Scotia, and Saskatchewan, had statewide subscriptions in 2014. Given their small shares of the country's total population, we treated the U.S. and Canada as having no subscription. Whereas a national subscription grants all domestic internet users free access to Cochrane Reviews, users in countries without a national subscription need to pay for access[2].

This article examines the expected demand for full-text reviews and plain-language summaries under free access for countries that have no national subscription. Absent institutional access, many healthcare professionals and patients may be unwilling or unable to purchase alternative access but would use reviews if access was free. Governments in countries without a national subscription, however, may be reluctant to subscribe nationally without knowing the expected benefit of such a policy.

In this article, we define the benefit of a national subscription as the increase in the downloads of Cochrane Reviews. This benefit depends on the elasticity of demand, that is, users' responsiveness to changes in the price of review downloads. National subscriptions reduce the marginal cost a user incurs for download of a review to zero. Using the standard model of supply and demand, we would expect review downloads to increase as more users can afford to download. When access is restricted, these potential users are either unable or unwilling to pay for review downloads and resort to summaries, which provide less detailed information, or other — potentially misleading — sources of information. Free access would attract downloads from both these users and those who learned about the service through its growing popularity.

(ISL), Ireland (IRL), Israel (ISR), Italy (ITA), Japan (JPN), Republic of Korea (KOR), Luxembourg (LUX), Mexico (MEX), Netherlands (NDL), New Zealand (NZL), Norway (NOR), Poland (POL), Portugal (POR), Slovak Republic (SVK), Slovenia (SVN), Spain (ESP), Sweden (SWE), Switzerland (CHE), Turkey (TUR), United Kingdom (GBR), and United States (USA).

[2]Individual users can read reviews at $6 each, download reviews at $38 each, or obtain a personal subscription at $365 annually. In addition, academic and corporate institutions with fewer than 1,001 employees can obtain licenses at annual prices of $2,582 and $3,812, respectively. All prices are given in U.S. dollars and retrieved from `www.cochranelibrary.com/help/how-to-order` and links therein on April 5, 2020.

An increase in review downloads can be expected to have a converse effect on its (imperfect) substitutes. On the one hand, this would be desirable if increased reviews manifested in reduced use of misleading sources of information. For example, misleading information, such as exaggerating benefits and downplaying harms of drugs or cancer screening, is the norm on (commercial) websites and in patient brochures (Gigerenzer & Hoffrage, 2007; Yeung & Mortensen, 2012). On the other hand, an increase in review downloads may also subtract from plain-language summary views; ignoring this substitution effect would overestimate the effect of a national subscription. We expect this effect to be limited because some users may prefer or need the detail of the reviews whereas others may prefer the conciseness and availability of plain-language summaries, particularly when summaries are translated into their native language. Translations from English into other national languages primarily address a lay audience (or healthcare professionals who do not understand statistics) with little or no command of English. We therefore expect that translating additional plain-language summaries can counteract the drop in summary views under free access, as they attract additional users who were previously unable to use the service.

To test these hypotheses, the goal of this article is to estimate the impact of national subscriptions on the number of downloads and views of (translated or untranslated) online summaries for individual OECD countries.

1.1 Method

The data used for the analysis were drawn from both Cochrane and publicly available databases. We obtained from Cochrane data of web traffic data on their websites in 2014, including the Cochrane Library hosted by Wiley and third-party sites such as EBSCO and OVID. From these data, we derived our two variables of interest for this study: the number of review downloads and the number of summary views, stratified by country. Each of these variables captures one way in which Cochrane reviews can be used. Full-text reviews are likely, but not exclusively, downloaded by healthcare professionals who understand technical details. Naturally, these professionals often function as multipliers who pass on information to patients. In contrast, patients without medical training are more

Table 1.1
Overview of Variables

variable	measure	type	source
`free`	open access in 2014	binary	Cochrane
`HINARI`	access possibility through HINARI	binary	Cochrane
`OECD`	OECD member state	binary	OECD
`english`	majority or official language is English	binary	CIA
`subscriptions`	number of subscriptions (absent national subscription)	intervals	Cochrane
`translations`	number of summaries in majority language (in 100)	continuous	Cochrane
`GDP`	gross domestic product per capita in 2016 US Dollars	continuous	World Bank
`population`	total population size	continuous	World Bank
`life`	average life expectancy	continuous	World Bank
`research`	number of scientific articles published in all fields	continuous	World Bank
`internet`	percentage of population with internet access	continuous	World Bank
`physicians`	number of physicians per 1000 persons	continuous	World Bank
`downloads`	number of full-text downloads in 2014	continuous	Cochrane
`views`	number of summary views in 2014	continuous	Cochrane

likely to consult plain-language or other summaries available on different Cochrane websites. These summaries are intended for a lay audience and are sometimes translated for this purpose. Jointly, the number of downloads and summary views give a comprehensive picture of how Cochrane Reviews are accessed.

Our analysis exploited the variation in the use of Cochrane Reviews across a range of countries to estimate the effect of different subscription schemes. Specifically, we compared the groups of countries with and without free access on their number of downloads and used the difference to calculate the expected effect of a national subscription on countries without one. Taking into consideration that each country's use of Cochrane Reviews is not exclusively affected by their subscription scheme, we collected data on additional determinants of review downloads and summary views. For example, we expected that more populous countries download, all else being equal, more reviews than less populous countries. Our analysis hence needed to isolate the effect of subscription type from that of

population size and other country characteristics.

Table 1.1 lists all variables considered in the analysis. The number of review downloads, number of summary views, and subscription status refer to 2014, whereas supplementary data (Central Intelligence Agency, 2020; World Bank, 2020) are as recent as 2016 but may go back as far as 2008, especially in less-developed countries. One variable, subscriptions, was available only as intervals of the form 0–50, 50–100, etc. For the analysis, we used the center of each interval as an estimate of each country's number of subscriptions. For some countries, the available data were incomplete. Excluding these countries, we obtained a total set of 158 countries for the analysis. Binary variables are coded as zero and one for no and yes, respectively.

We used two linear models to isolate the effects of a national subscription on review downloads and summary views, respectively. The first model, DOWN, decomposes the number of downloads into the effects of the different country characteristics listed in Table 1.1. Formally, the number of review downloads of country i is given by

$$\begin{aligned} ln[\texttt{downloads}_i] = \alpha_0 &+ \alpha_1\, ln[\texttt{GDP}_i] + \alpha_2\, ln[\texttt{population}_i] \\ &+ \alpha_3\, ln[\texttt{research}_i] + \alpha_4\, ln[\texttt{internet}_i] \\ &+ \alpha_5\, ln[\texttt{life}_i] + \alpha_6\, ln[\texttt{physicians}_i] \\ &+ \alpha_7\, ln[\texttt{subscriptions}_i] + \alpha_9\, \texttt{HINARI}_i \\ &+ \alpha_{10}\, \texttt{OECD}_i + \alpha_{11}\, \texttt{english}_i \\ &+ \alpha_{12}\, \texttt{free}_i + \epsilon_i, \end{aligned}$$

where α_0 denotes the intercept, α_j denotes the partial effect of variable j, and ϵ_i denotes an error term that is assumed to be independently, identically, and normally distributed. The purpose of the analysis was to estimate the parameters α_0 to α_{12} chief among them was α_{12}, the effect of a national subscription.

In addition to estimating the model shown here, we also estimated an augmented model that includes interaction effects of `free` with `english` and `population`. Likewise, we estimated three different nonlinear models that predict downloads by combining a set of regression trees such as random forests (Breiman, 2001a). Because many variables were not normally distributed but included considerable outliers, all five models were tested with and without logarithmic transformation of all continuous variables, yielding a total set of 10 models that were tested. These models were compared on the qual-

ity of their out-of-sample predictions using 17-fold cross-validation, where the test set was restricted to the 34 OECD member states. The model presented above, with logarithmic transformation, produced a root-mean squared error (RMSE) of 182,662 downloads, whereas the closest competitor exhibited an RSME of 187,298 downloads. A sensitivity check using the model with the next lowest out-of-sample error yielded comparable results. Further, a visual check of the model assumptions revealed no irregularities.

The second model, VIEW, decomposed summary views into the effects of the different country characteristics listed in Table 1.1. Unlike reviews, summaries are sometimes translated into other national languages, but the number of translated summaries varies across countries. To separate the effect of language from that of national subscriptions, we used the same linear model as before to estimate the number of summary views based on country characteristics but replaced the binary variable `english` with `translations`, which gives the number of plain-language summaries available in the national language. Formally, summary views are then described as follows:

$$\begin{aligned} ln[\texttt{views}_i] = \beta_0 &+ \beta_1\, ln[\texttt{GDP}_i] + \beta_2\, ln[\texttt{population}_i] \\ &+ \beta_3\, ln[\texttt{research}_i] + \beta_4\, ln[\texttt{internet}_i] \\ &+ \beta_5\, ln[\texttt{life}_i] + \beta_6\, ln[\texttt{physicians}_i] \\ &+ \beta_7\, ln[\texttt{subscriptions}_i] + \beta_8\, ln[\texttt{translations}_i] \\ &+ \beta_9\, \texttt{HINARI}_i + \beta_{10}\, \texttt{OECD}_i \\ &+ \beta_{12}\, \texttt{free}_i + \epsilon_i, \end{aligned}$$

where β_0 denotes the intercept, β_j denotes the partial effect of variable j, and ϵ_i denotes an error term that is assumed to be independently, identically, and normally distributed. Again, the purpose of the analysis was to estimate the parameters β_1 to β_{12}, with particular interest in variables β_8 and β_{12}.

As before, we chose this model from a set of 10, including six random-forest and four linear models. Two of the linear models slightly outperformed the selected model in 17-fold cross-validation, with test sets restricted to OECD countries. These models used non-logarithmic versions of the variables included and yielded root mean squared errors around 301,000 views whereas the chosen model yielded an error around 317,000 views. Nonetheless, we chose the selected model because the logarithmic versions seemed

more adequate, particularly because the model led to slightly better estimates for the majority of countries, although predictions for a few countries were less precise. A sensitivity check showed that this choice was conservative in the sense that the combined effects of `free` and `translations`, which are most relevant to our argument, are somewhat smaller in the model chosen than in the model with the lowest out-of-sample error.

1.2 Results

In this section, we compare countries with and without a national subscription on their review downloads and summary views. We present the results of our two statistical models and use these models to calculate the expected number of reviews for all OECD countries. Finally, we provide rough estimates of the monetary costs of a national subscription.

1.2.1 Review Downloads

The black and gray circles in Figure 1.1 show the total number of review downloads in 2014 for all OECD member states. The position of each circle on the x-axis indicates the number of downloads per 1,000 persons and the size of the circle indicates the total number of downloads. Among countries without free access, shown by the black circles, the Netherlands, Sweden, and Switzerland had the highest and Mexico, Slovakia, and the Czech Republic the lowest number of downloads per capita. Although there was a tendency for more prosperous countries to have more downloads per capita, exceptions can be found. Most notably, there were seven downloads per 1,000 persons in Chile, but only 0.25 per 1,000 in Japan. On average, countries without a national subscription downloaded 2.33 reviews per 1,000 persons.

The Netherlands had 10 downloads per 1,000 persons, making it the country with by far the highest download rate among those without free access. For countries with a national subscription, the gray circles show downloads per capita. Each of these countries had more downloads per capita than the Netherlands, on average 19.20 reviews per 1,000 persons. Download rates were particularly high for anglophone countries, suggesting a linguistic advantage.

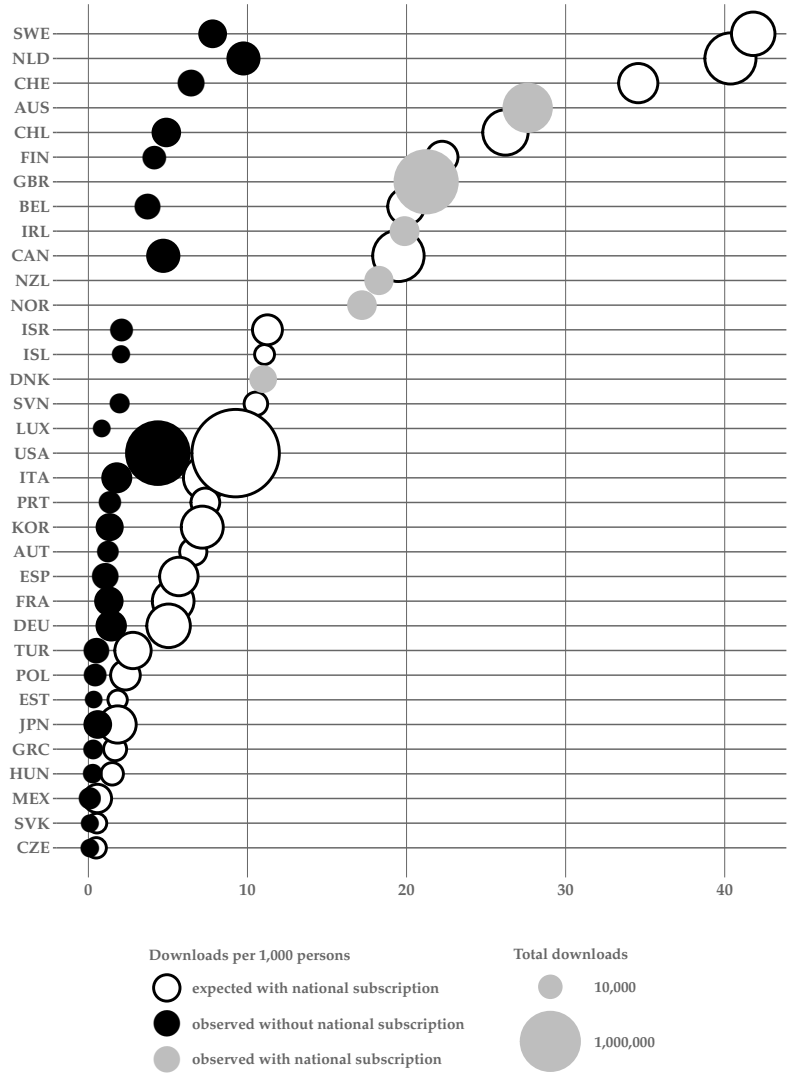

Figure 1.1: *Observed and expected annual review downloads per 1,000 persons*

To illustrate the effect of a national subscription, it can be instructive to compare countries that differ in their subscription status but are similar in many other respects. For example, Denmark and Norway, with free access, had roughly twice as many downloads as Finland, which was without free access. Likewise, United Kingdom, with free access, had roughly the same total as the U.S., without free access, despite its population being only a fifth of the latter's.

Although these comparisons provide a first indication that more reviews were downloaded when access was free, the DOWN model offers a more rigorous estimation of the effect of a national subscription when it is isolated from other factors. Comparing the standard deviation of ln(`downloads`) with the prediction error for OECD countries indicates that the model yielded fairly accurate predictions. However, a visual inspection of the predictions per country (not shown) revealed that the models underestimated downloads for Chile and the Netherlands, whose downloads appeared to be driven by idiosyncratic factors omitted here. At the same time, the countries that were best predicted appear to be those with free access.

Columns 5 and 6 of Table 1.2 present the estimates of α_1 to α_{12}, which indicate the approximate percentage increase in review downloads associated with a one-percent increase in the variable of interest[3].. For example, a gross domestic product (GDP) increase of one percent is associated with a rise in review downloads of half a percent on average. As expected, most variables are positively associated with review downloads. The only exceptions are OECD membership, which appears to have no discernible effect beyond the effect of GDP, and the number of physicians. Increasing the number of physicians by 10 percent is estimated to reduce downloads by about three percent on average. This negative effect may appear surprising. One possible explanation is that an increase in the number of physicians per capita implies fiercer competition among them, given that the number of patients is fixed. Such increased competition may incentivize physicians to favor profitable over effective treatments, lowering demand for medical evidence (Mulley & Wennberg, 2011). A second possible explanation is that countries with more physicians

[3]Coefficients of this type are often referred to as *elasticities*. In $ln[y] = \alpha \times ln[x]$, elasticity α gives the approximate percentage change in y associated with a one percent change in x. To see this, recall that $1 + \Delta \approx e^{\Delta}$ for small values of Δ, so raising x by one percent increases $ln[x]$ to $ln[x \times 1.01] \approx ln[x \times e^{0.01}] = ln[x] + ln[e^{0.01}] = ln[x] + 0.01$. Correspondingly, α gives the increase in $ln[y]$ and percentage increase in y.

Table 1.2
Estimated Coefficients and Diagnostics of OLS-Regression Models

		OECD		review downloads		summary views	
No	variable	min	max	$\hat{\alpha}$	SE_α	$\hat{\beta}$	SE_β
0	intercept	—	—	−28.33	5.48	−33.86	8.35
1	ln[GDP]	9.24	11.67	0.48	0.17	0.46	0.26
2	ln[population]	12.70	19.58	0.64	0.10	0.90	0.16
3	ln[research]	5.32	12.25	0.34	0.09	0.25	0.14
4	ln[internet]	3.79	4.59	0.41	0.21	0.73	0.32
5	ln[life]	4.32	4.42	3.94	1.20	4.36	1.82
6	ln[physicians]	0.03	1.82	−0.28	0.13	−0.38	0.19
7	ln[subscriptions]	0	7.13	0.24	0.09	0.06	0.14
8	ln[translations]	−2.30	8.69			0.12	0.03
9	HINARI	0	1	1.64	0.34	0.27	0.52
10	OECD	0	1	−0.09	0.35	−0.38	0.53
11	english	0	1	0.76	0.26		
12	free	0	1	2.46	0.53	0.30	0.80

estimated variable	ln[downloads]	ln[views]
standard deviation (all 158 countries)	2.68	3.14
standard deviation (all 34 OECD countries)	2.14	1.70
fitting: share of variance explained	0.85	0.74
cross-validation RMSE (158 countries)	1.12	1.76
cross-validation RMSE (34 OECD countries)	0.88	1.06
cross-validation RMSE (8 free-access countries)	0.21	0.94

are more likely to have alternative resources available such as national guidelines for professional practice including those by the US Preventive Services Task Force (USPSTF). A comparison of guidelines in the UK and the US points in this direction (Weisz et al., 2007).

For the present purpose, interest lies in the estimated effect of a national subscription. All else equal, the model estimated that the number of review downloads increased, on average, to $e^{2.46} \times 100 \approx 1{,}166$ percent when access was free. However, concluding that a national subscription increases downloads tenfold would be premature. Under a national subscription, institutional and individual subscriptions are no longer needed and should no longer be considered in the model. We therefore need to subtract the estimated effect of those subscriptions from that of a national subscription to obtain the incremental effect. The model then estimates that for countries with 25 or 150 subscriptions, the number of downloads would increase to $e^{2.46-0.24\times ln[25]} \times 100 \approx 540$ percent and to $e^{2.46-0.24\times ln[150]} \times 100 \approx 352$

percent, respectively. As usual, these estimates indicate the average increase in the number of downloads, and observed increases may vary for countries that are dissimilar to those that had a national subscription in our data.

The estimated coefficient for a national subscription exhibits a large standard error of 0.53 (see Table 1.2). Although we can reject the null that $\alpha_{12} = 0$, we suspect that the lack of precision is due to the fact that only eight countries are currently subscribed whereas 150 countries are not. Given this imbalance, a large standard error is not surprising. For an alternative assessment of the accuracy of the estimated coefficient $\hat{\alpha_{12}} = 2.46$, we calculated the root-mean squared error in cross-validation specifically for those countries with free access. To this end, we predicted downloads for each country separately, based on parameters estimated from the data of all other countries. Across the resulting 158 models, the estimated effect of free access varied only slightly between 2.33 and 2.56. Using these estimates, the bottom of Table 1.2 shows that the model's predictions were considerably more precise among countries with free access than among those without. These findings indicate that the estimated effect of free access is closer to its true value than its standard error may suggest.

Using the estimated coefficients and the data on existing subscriptions, we can calculate for each OECD country the number of expected downloads under a national subscription. These projections are shown by the white circles in Figure 1.1. The logarithmic nature of the model implies that the number of additional downloads generated by a national subscription is driven by the existing download volume: countries with larger download volumes (e.g., anglophone, populous, and prosperous) are expected to profit more from their introduction.

Consider two cases that illustrate the expected effects of a national subscription. First, recall the case of the U.S. with as many downloads as the United Kingdom (around 1.4 million), despite having a population that is five times larger. The results of our analysis showed that a national subscription would be expected to generate an additional 1.56 million downloads per year, doubling the national total. Second, among non-anglophone countries, Germany had a download level of only 116,000 reviews, less than twice as many as Denmark despite its population being around 13 times larger. A national subscription would be estimated to increase national totals in Germany to 408,000,

approaching Denmark's rate of downloads per person.

1.2.2 Summary Views

Our second analysis concerned the effect of a national subscription on plain-language summary views. The black and gray circles in Figure 1.2 show the number of plain-language summary views in 2014 per 1,000 persons for all OECD member states. Among countries without a national subscription, France, Canada, and Spain had the highest number of views per capita, and Turkey, South Korea, and Japan had the lowest. The average number of summary views for countries without a national subscription was 2.18 summaries per 1,000 persons.

In contrast, there were on average 5.39 summaries per 1,000 persons in countries with a national subscription, indicating an effect of such a subscription. Although the levels for countries with and without national subscription overlap, the highest level (9.11 views per 1,000 persons) was reached by Australia, which held a national subscription. Within this group of subscribing countries, Denmark had the fewest views per capita, in keeping with the level of structurally similar countries such as Finland and Sweden. Among national subscribers, anglophone countries appear to have consumed more: not only were more reviews downloaded, as noted before, but also more summaries were viewed.

The VIEW model offers a more detailed examination of the effects of a national subscription and of language. Although the model diagnostics indicated that the model yielded acceptable predictions, predicting the number of summary views was apparently more difficult than predicting downloads. Most notably, the model overestimated the number of views from Japan, Germany, and South Korea, where there appeared to have been constraining factors omitted from the model. Columns 7 and 8 of Table 1.2 report the estimated model parameters. Whereas most variables had their expected positive effect on the number of summary views, a higher density of physicians and OECD membership decreased the number, although this latter effect is imprecisely estimated. We were particularly interested in the estimated effects of free access and translations.

Given the substituting nature of full-text downloads and summaries, we had expected a negative effect of a national subscription on summary views. Surprisingly, the estimated effect was positive,

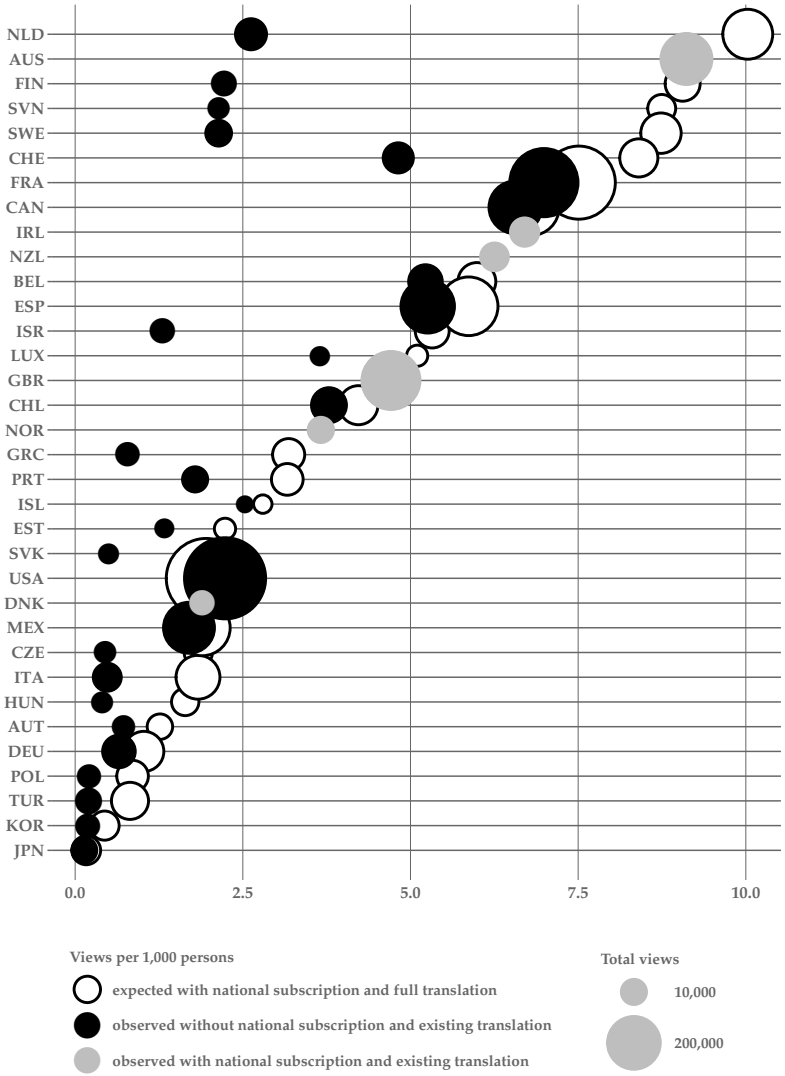

Figure 1.2: *Current and expected annual summary views per 1,000 persons; note that scales differ from Figure 1.1.*

indicating at first sight that the additional popularity of the service compensates for summary views supplanted by review downloads. However, there are two caveats to this conclusion. First, the effect was imprecisely estimated so that the degree of compensation cannot be firmly established to be positive or negative. More importantly, subtracting the effects of existing subscriptions can lead to a negative net effect for countries with more than $e^{0.30/0.06} \approx 136$ existing subscriptions. Generally, we conclude that the negative effect of a national subscription on summary views appears to be small, if at all present.

In contrast, the effect of translations was precisely estimated and positive. The point estimate indicates that increasing the number of summary translations by 100 percent increases views by approximately $e^{0.12} \times 100 - 100 \approx 13$ percent. Although the magnitude of this effect appears small, it is worth pointing out that some countries had only few translations. For example, only 128 of 5,952 summaries have been translated into German. A translation of all summaries is then estimated to increase summary views to 158 percent.

To illustrate the interaction of the effects of free access to full-text reviews and summary translations, we used the model estimates to calculate for all OECD countries the number of expected summary views under a national subscription and full translation. These projections are shown by the white circles in Figure 1.2 and vary considerably for two reasons. First, as before, the logarithmic nature of the model implies that those with many views benefit more strongly than those with few views. Second, countries vary in their progress on summary translations, and those with few translations have more room for improvement than those with many translations.

1.2.3 Implied Costs

Like all policy instruments, national subscriptions to Cochrane Reviews ought to be subjected to a cost-benefit analysis. We have seen above that the benefits in terms of additional full-text downloads and summary views vary across countries but can be substantial in some cases. Here, we set these benefits in relation to the monetary costs of a national subscription. These costs depend on the price of a national subscription and the amount spent on existing subscriptions that would be obsolete under a national subscription. We will discuss each of these factors in turn.

Although Cochrane does not publish rates for national subscriptions, the annual rate is believed to be around \$0.01 per capita (Antes, n.d.). On the basis of this estimate, the first column of Table 1.3 lists the total costs of a national subscription for each country according to its population size. For example, a national subscription for small countries such as Finland or Austria would cost less than \$100,000 annually while larger countries such as Germany or Japan would require around one million dollars per year.

At the same time, a national subscription implies that existing subscription holders no longer need their subscriptions. This may further lower the cost of a national subscription. Unfortunately, we do not know each country's total spending on individual downloads, personal licenses, or institutional subscriptions. However, our data include an interval of the total number of subscriptions, which can be used to estimate existing total spending. For this purpose, we assumed that observed downloads increase linearly within each subscription interval and estimated for each country i the number of subscriptions, $\hat{s}_i$, from the number of review downloads, d_i, using

$$\hat{s}_i = r_i + (t_i - r_i) \times \frac{d_i - c_i}{e_i - c_i},$$

where r_i and t_i denote the lowest and highest possible number of subscriptions in the interval of country i, and c_i and e_i denote the minimum and maximum number of downloads for countries with the same interval. The approximated number of subscriptions is then multiplied by \$2,582 , which is the price of the least expensive institutional subscription. For the country with the fewest downloads in the interval, $\hat{s}_i$ is set at the lower bound of the interval plus ten percent of its range, to avoid inconsistencies at the interval bounds. Conversely, for the country with the most downloads, $\hat{s}_i$ is set at the upper interval bound minus ten percent of its range. For example, if three countries in the 50−100 subscriptions interval had 100, 1,100, and 350 downloads, they would be assumed to have 55, 95, and $55 + (95 - 55) \times 250/1000 = 65$ subscriptions, respectively. With only one country per interval, $\hat{s}_i$ is set at the center of the interval.

The second column of Table 1.3 lists the approximate existing total spending for each country. It shows that some countries without a national subscription, such as Germany and Japan, have spent large amounts of money on Cochrane licenses for research institutions or medical organizations. Under a national subscription, these

Table 1.3
Estimated Costs

	estimated total costs		
	with open access	without open access	estimated costs per additional download
Australia	$234,907	—	—
Austria	$85,345	$24,590	$1.33
Belgium	$112,252	$59,016	$0.29
Canada	$355,404	$233,605	$0.23
Chile	$177,626	$110,655	$0.18
Czech Republic	$105,106	$12,295	$22.17
Denmark	$56,396	—	—
Estonia	$13,136	$12,295	$0.43
Finland	$54,636	$36,885	$0.18
France	$662,069	$135,245	$1.97
Germany	$808,895	$368,850	$1.51
Greece	$109,577	$14,754	$6.34
Hungary	$98,617	$14,754	$6.99
Iceland	$3,276	$12,295	−$3.06
Ireland	$46,127	—	—
Israel	$82,153	$31,967	$0.67
Italy	$613,364	$164,753	$1.31
Japan	$1,271,318	$614,750	$4.14
Luxembourg	$5,561	$0	$1.11
Mexico	$1,253,858	$27,049	$21.25
Netherlands	$168,542	$231,146	−$0.12
New Zealand	$45,097	—	—
Norway	$51,365	—	—
Poland	$379,955	$29,508	$4.87
Portugal	$103,974	$27,049	$1.24
Slovakia	$54,185	$12,295	$17.70
Slovenia	$20,622	$17,213	$0.19
South Korea	$504,240	$88,524	$1.42
Spain	$464,046	$68,852	$1.84
Sweden	$96,896	$98,360	−$0.00
Switzerland	$81,902	$71,311	$0.05
Turkey	$759,323	$56,557	$4.06
United Kingdom	$645,104	—	—
United States	$3,188,571	$3,073,750	$0.07

individual licenses would become obsolete. However, to determine the actual financial burden of a national subscription, it may be important to consider the mix of private and public institutions among existing subscribers. Unlike potential savings by public institutions, which may be subtracted from the total costs, savings by private institutions would in fact raise costs to governments through

foregone sales taxes. However, we suspect that the large majority of existing subscribers are publicly funded, implying that omitting the need for existing spending on Cochrane licenses would reduce the effective cost of a national subscription.

The third column of Table 1.3 subtracts the estimated existing costs from the estimated total and divides it by the estimated increase in review downloads shown in Figure 1.1. Integrating costs and benefits, this column can be used to separate the countries into three groups. First, three countries, Czech Republic, Mexico, and Slovakia, would pay around $20 per additional download, a sum that falls short of the price of an individual download but exceeds the cost of merely viewing a review. Similarly, Greece, Hungary, Poland, and Turkey would pay $4 to $7 , considerably more per additional download than most other countries. Second, three countries, Iceland, Sweden and the Netherlands, are predicted to save money through a national subscription. The majority of countries, including Canada, France, Germany, Italy, and the U.S., fall in between these extreme groups, with costs per additional download ranging between $0.05 and $2

1.3 Discussion

Cochrane Reviews are currently of limited use, as many healthcare professionals and most patients do not have free access to them.. In spite of efforts to promote informed healthcare professionals and patients, governments have been reluctant to purchase national subscriptions. We calculated estimates of the increase in full-text downloads and summary views of Cochrane Reviews in OECD countries if they were to purchase a national subscription. We then integrated these estimated benefits with the estimated costs of a national subscription and provided a measure of the effective costs.

Our findings are encouraging. Although the estimated increases in downloads vary between countries, Figure 1.1 shows that considerable improvements are possible. Indeed, the majority of countries is projected to multiply their downloads by a factor above two, including countries with few downloads in the absence of a national subscriptions.

In addition, our analysis of summary views showed that a national subscription is not associated with a reduction in summary views. Instead, the effect of a national subscription could be both

positive and negative, depending on the country. However, Figure 1.2 illustrates that translations of summaries into the national language can attenuate possible negative effects and offer a second avenue for disseminating Cochrane evidence. As we used each country's national language to determine the number of available translations, the model did not control for national differences in English proficiency. Therefore, the model may have overestimated the effect of additional translations for countries in which English is widely and well understood, such as Scandinavian countries or the Netherlands. Nonetheless, the results indicate that translations have the potential to increase summary views in many countries, including some without exceptional English proficiency. For example, Slovenia, Greece, Italy, and Germany hold the potential for considerable improvements through comprehensive translation of existing summaries. We therefore conclude that translations of Cochrane summaries offer an additional tool for disseminating Cochrane evidence that can be used independently of a national subscription.

Integrating these estimated benefits with the costs of a national subscriptions, we find that for most OECD members, the net costs would be small. Whereas seven countries can expect to face insufficient demand to justify the purchase of a national subscription, according to our estimates in Table 1.3, many countries would pay less than $1 for each additional download and three countries would save money under a national subscription. Thus, for most countries, national subscriptions to Cochrane Reviews present an inexpensive way of disseminating medical evidence. The question of to what degree this evidence will be used cannot be answered by the present study, although Cochrane Reviews have in the past had direct impact on policy making (Bunn et al., 2014) and, when translated into fact boxes and other understandable forms, can foster physicians' and patients' understanding and decision making (McDowell, Gigerenzer, Wegwarth, & Rebitschek, 2019; Wegwarth & Gigerenzer, 2018; Wegwarth, Wagner, & Gigerenzer, 2017).

The estimates of our analysis are based on an OLS regression model of observational data and are not without caveats. Most importantly, observational data are ill-suited to establish causal relationships. That is, our analysis cannot formally answer the question whether national subscriptions lead to increases in downloads, whether the reverse is true, or whether both variables have a common cause. Instead, we have found that it is more plausible that a

national subscription leads to a given number of downloads than vice versa because subscriptions have a causal effect on the costliness of a download. However, it is important to note that there remains the possibility that both subscription and downloads are caused by a third variable that we have not accounted for in our models. Despite all efforts to control for potential confounds such as economic strength or research activity, comparisons across countries retain the possibility that relevant differences between countries remain unnoticed or unobserved. To corroborate our findings, we therefore encourage studies that examine the effects of national subscriptions by comparing downloads before and after its introduction within the same country.

A second limitation of this study concerns the uncertainty of our cost estimates. When calculating the expected costs per additional download, both the numerator and the denominator were based on estimates. The costs in the numerator were based on estimates of the costs of existing subscriptions and the denominator was based on our model estimates. Although we could compute confidence intervals for the denominator, we cannot quantify the uncertainty of the numerator, which is based on the number subscriptions and an estimate of their costs. These estimates are conservative but their uncertainty remains unclear until more detailed data on subscriptions becomes available. Our cost-benefit analysis provides estimates of the effective costs per download gained through a national subscription. The analysis remains agnostic as to how highly additional downloads are valued and leaves such judgements to policy makers. However, we emphasize the importance of evidence for directing healthcare resources to where they are most effective. This is especially true in healthcare systems where various actors are incentivized to overstate the effectiveness of different health interventions. In these environments it is key that professionals and patients are empowered to base their decisions on evidence instead of advertisements. To be effective, good evidence requires not only high-quality studies but also easy access to their conclusions.

Chapter 2

Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning

This chapter was published as McDowell, M. & Jacobs, P. (2017), Meta-Analysis of the Effect of Natural Frequencies on Bayesian Reasoning, *Psychological Bulletin 143*(12), 1273–1312.

The ability to make sound probabilistic inferences has long been considered essential to human rationality. Assessing whether human inferences adhere to the rules of probability theory has therefore a long tradition, not only in the study of judgement and decision making, but also in economics and philosophy (Gigerenzer et al., 1989; Hacking, 2006; Savage, 1954; von Neumann & Morgenstern, 1944). The conclusions from early work in psychology suggested that "man is an intuitive statistician", albeit a slightly conservative one (Peterson & Beach, 1967; Phillips & Edwards, 1966, p. 39). That is, studies suggested that human inference followed or approximated the rules of probability theory.

Subsequent work led to the contradictory view that the human mind was not built to work according to the rules of probability (Kahneman & Tversky, 1972, 1973). For example, based largely on laboratory studies using textbook problems, the heuristics-and-biases program documented a long list of cognitive errors or fallacies where human judgements about probabilities deviated from the normative standards of probability theory (for an overview see Gilovich, Griffin, & Kahneman, 2002). For example, one prominent finding was that participants tended to overweight or ignore base rates (e.g., the prevalence of breast cancer in a population) in probabilistic inference,

phenomena referred to as the base-rate fallacy or base-rate neglect. In contrast, earlier work suggested that participants almost always take the base rate into consideration (see Koehler, 1996, for a discussion of theoretical and methodological criticisms of prior empirical work on the base-rate fallacy). Nevertheless, the apparent robustness of the base-rate fallacy was considered a demonstration of cognitive error (Bar-Hillel, 1980, 1984). This finding, among others, led to the conclusion that "[i]n his evaluation of evidence, man is apparently not a conservative Bayesian: he is not Bayesian at all" (Kahneman & Tversky, 1972, p. 450). These findings underlie the view of some behavioral economists that we can *nudge* people into making better decisions by exploiting their cognitive biases — a view often referred to as libertarian paternalism, as people retain choice but are steered towards decisions that governments and institutions deem welfare enhancing (Gigerenzer, 2015; Grüne-Yanoff & Hertwig, 2016; Thaler & Sunstein, 2008).

A number of explanations have been offered to reconcile these conflicting views. Some authors have argued that many findings from the heuristics-and-biases program use an inadequate normative standard, or that in many everyday situations it can be ecologically rational (e.g., adaptive in a natural ecology) to ignore information or contradict the axioms of rational choice theory (see, e.g., Gigerenzer, 1996a; Kahneman & Tversky, 1982; Stanovich & West, 2000). Critiques have also compared the research methodologies used within different programs (Hertwig & Erev, 2009; Schulze & Hertwig, 2016) or have observed that prior studies have lacked ecological validity (Fiedler & von Sydow, 2015; Koehler, 1996). For instance, the apparent base-rate neglect phenomenon from research on textbook problems contradicts findings from related work on probabilistic reasoning using experience-based paradigms where participants learn the associations or co-occurrences between events. Animals, children, and adults in prenumerate indigenous populations are found to be capable of making probabilistic inferences in line with the statistical properties of environments (Biernaskie, Walker, & Gegear, 2009; Fontanari, Gonzalez, Vallortigara, & Girotto, 2014; Gopnik, Sobel, Schulz, & Glymour, 2001; Rakoczy et al., 2014; Real, 1991; Real & Caraco, 1986; Sobel & Munro, 2009; Sobel, Tenenbaum, & Gopnik, 2004).

The present meta-analysis focuses on one of these criticisms that relates to the tendency for textbook tasks on Bayesian inference to ignore the connection between external information representations

(e.g., numerical representations) and cognitive processing. In these textbook tasks, probabilities are summarised and one must make an inference based on a description of the relevant statistics. In their seminal paper, Gigerenzer and Hoffrage (1995) argued that there are different ways to summarise statistical information that are mathematically equivalent but not necessarily computationally equivalent, so that the choice of representation format affects the performance of a given cognitive process. For instance, although different numerical representations (e.g., Arabic or Roman numerals, or binary systems) can be mapped onto one another, the (cognitive) algorithms that operate on these representations may require different computations. A common analogy is that of the pocket calculator, designed to operate on Arabic numerals: can one infer that it has an algorithm for multiplication when it is fed information in binary numbers? In relation to an elementary Bayesian inference textbook task, often used in demonstrations of the base-rate fallacy, Gigerenzer and Hoffrage demonstrated that presenting statistical information in the form of joint frequencies resulting from natural samples, known as *natural frequencies* (defined in detail, below), yielded substantial improvements in Bayesian reasoning. The present meta-analysis focuses on research that has sought to account for or explain this facilitation effect.

Over the past 20 years, there has been some debate as to why natural frequencies can facilitate Bayesian inference in textbook problems. Although there is now a general consensus that natural frequencies can facilitate Bayesian inference (Brase & Hill, 2015; Johnson & Tubau, 2015), studies in this area report substantial variations in performance rates, ranging from 0 to 90 percent correct solutions recorded across studies. As such, it is unclear exactly how much natural frequencies facilitate performance and it is difficult to quantify the conditions under which facilitation effects are most likely to occur. For example, we know that natural frequencies can boost probabilistic inferences but to what extent and how high can performance be bolstered, given which features of the problem and in what contexts? More recent work has turned its attention to identifying the features of the person, problem, or methodological context that account for differences in performance rates across studies or that can shed light on underlying mechanisms. However, the field lacks a systematic examination of these factors. Rather, much of the prior work has focused on debating which theoretical perspective, the *ecological rationality framework* or

nested sets theory, offers the most coherent account for why facilitation effects occur (Brase & Hill, 2015). Early work was plagued by misinterpretations of the natural frequency format, yet the theoretical accounts that were proposed on the basis of these misinterpretations actually converge on many common concepts (e.g., the importance of the subset problem structure). We move beyond these theoretical debates to examine why, when, and for whom natural frequencies facilitate Bayesian inference in an attempt to offer some closure or focus to the debate, and to provide guidance for future studies.

Accordingly, the current review and meta-analysis has three broad aims. First, we clarify what are natural frequencies (and what they are not) and highlight where theoretical accounts converge and diverge. Second, we identify and examine potential moderators for when, why, and for whom natural frequencies facilitate Bayesian inference. We report results from a meta-analysis of 35 studies representing 9,611 participants on the relative effect of natural frequencies in comparison to conditional probabilities (normalised formats, defined below), and report how individual, methodological, and problem representation factors moderate this effect. Third, we emphasise current research gaps and suggest how to stimulate and progress research in this area.

2.1 Bayesian Inference and Natural Frequencies

Bayesian inference refers to the process of updating a prior probability of some hypothesis in response to new data. The normative benchmark for this process is provided by Bayes' theorem, described in detail below. A broad range of phenomena across many different research areas has been modeled using Bayes' theorem, from the study of perception to human cognition (Chater, Oaksford, Hahn, & Heit, 2010; Chater, Tenenbaum, & Yuille, 2006). In relation to human cognition, the question is how well Bayes' theorem describes human inferences, for example, how the probability of some (unobserved) event changes in the presence of another (observed) event.

One approach to study Bayesian inference is the textbook paradigm where probabilities are summarised and one must make an inference based on a description of the relevant statistics. Consider, for example, an elementary textbook version of a Bayesian inference task presented using conditional probabilities (Eddy, 1982; Gigerenzer & Hoffrage, 1995, p.685):

> The probability of breast cancer is 1% for a woman at age forty who participates in routine screening. If a woman has breast cancer, the probability is 80% that she will get a positive mammography. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? _____%.

In order to solve this task, let H denote the hypothesis (here, the presence of breast cancer), D a specific data outcome (here, a positive mammogram), and $\neg H$ and $\neg D$ their negations (here, the absence of breast cancer and a negative mammogram, respectively). The probability of the hypothesis after learning outcome D is given by

$$p(H|D) = \frac{p(D \cap H)}{p(D)} = \frac{p(D \cap H)}{p(D \cap H) + p(D \cap \neg H)}, \tag{2.1}$$

that is, by dividing the probability of having both breast cancer and a positive mammogram by the overall probability of having a positive mammogram (both with and without breast cancer). As the problem does not state the *joint probabilities* $p(D \cap H)$ and $p(D \cap \neg H)$, they need to be calculated. First, $p(D \cap H)$ can be obtained from multiplying $p(H)$, the base rate of breast cancer (that is, the prevalence of breast cancer in the reference class), with $p(D|H)$, the hit rate of the mammography test (the probability of a positive mammogram given that one has breast cancer). Similarly, $p(D \cap \neg H)$ is given by $p(\neg H) \times p(D|\neg H)$. Filling these probabilities into equation (2.1) yields Bayes' theorem:

$$p(H|D) = \frac{p(H) \times p(D|H)}{p(H) \times p(D|H) + p(\neg H) \times p(D|\neg H)}. \tag{2.2}$$

Generally speaking, Bayes' theorem describes how the *prior probability* $p(H)$, that is the probability of the hypothesis without additional information, is combined with a *likelihood* $p(D|H)$, that is the probability of outcome D if hypothesis H was true, to obtain the *posterior probability* $p(H|D)$, that is the probability of the hypothesis after obtaining additional information. Using the numbers given in the

problem and equation (2.2), one can compute the solution:

$$p(H|D) = \frac{.01 \times .80}{.01 \times .80 + .99 \times .096}$$

$$\approx .078.$$

There is overwhelming evidence demonstrating that participants have difficulty solving such problems when presented in conditional probability formats, as shown above. Gigerenzer and Hoffrage (1995) reported that only 16 percent of participants in their study could provide the correct Bayesian solution to such problems, a finding consistent with many subsequent studies in the field (e.g., Chapman & Liu, 2009; Ferguson & Starmer, 2013; Mellers & McGraw, 1999). One interpretation of the poor performance on the above textbook problem is the base-rate neglect phenomena mentioned previously: participants neglect the base rate in their calculations and erroneously focus on specific case data (Barbey & Sloman, 2007).

An alternative view was offered by Gigerenzer and Hoffrage (1995). Drawing on an evolutionary perspective, they proposed that, as probabilities are a relatively recent information representation, probabilistic information would likely have been acquired and updated sequentially in reference to the event's frequency in natural environments — that is, a process of *natural sampling* (Kleiter, 1994). Gigerenzer and Hoffrage argued that cognitive algorithms for statistical inference would likely have evolved to operate on this type of representation. Consider, by analogy, a physician who acquires information sequentially about patients who have a symptom and a disease, and those who have the symptom but do not have the disease. The physician can then use these joint frequencies to calculate specific probabilities for a newly presented patient.

As we will see, the Bayesian algorithm is computationally simpler if probabilities are represented as using joint frequencies, which should allow more participants to find the correct solution. To test this assumption, Gigerenzer and Hoffrage (1995) presented participants with the same elementary Bayesian inference task described earlier but presented information in natural frequencies (p.688):

> 10 out of every 1,000 women at age forty who participate in routine screening have breast cancer. 8 of every 10

> women with breast cancer will get a positive mammography. 95 out of every 990 women without breast cancer will also get a positive mammography. Here is a new representative sample of women at age forty who got a positive mammography in routine screening. How many of these women do you expect to actually have breast cancer? _____ out of _____.

Here, the *joint frequencies* are given in the problem and can be combined to obtain the posterior probability directly, analogously to equation (2.1),

$$p(H|D) = \frac{n(D \cap H)}{n(D \cap H) + n(D \cap \neg H)} \tag{2.3}$$

$$= \frac{8}{8 + 95}$$

$$\approx .078,$$

where $n(\cdot)$ denotes the frequencies. When these frequencies result from a natural sampling process, these are referred to as natural frequencies. Related work on experience-based probabilistic inference allows participants to experience this type of natural sampling process, whereas this version in the textbook paradigm presents participants only with the outcome of this process (Hoffrage, Krauss, Martignon, & Gigerenzer, 2015; Schulze & Hertwig, 2016).

As natural sampling preserves information about the base rate, which is contained in the joint frequencies, the base rate can be ignored, simplifying the calculation of the correct solution: stating the hit rate as 8 in every 10 women preserves the information that only 10 in every 1,000 women have breast cancer; the information is not normalised[1] (e.g., compare Figures 2.1A and 2.1B to 2.1C and 2.1D, where the information is normalised in the former but not in the latter). When presented with information in the natural frequency format, Gigerenzer and Hoffrage (1995) found that 46 percent of

[1]For consistency and clarity, in the present paper we use the term *conditional probabilities* to refer to any format that has a normalised structure, although we acknowledge that chances, percentages, and frequency formats with a normalised structure are not strictly conditional *probabilities*.

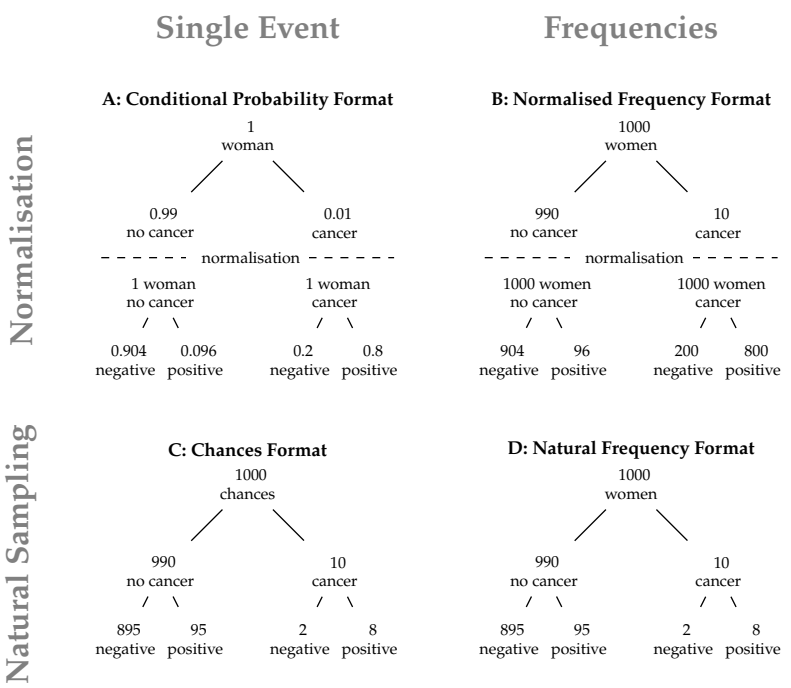

Figure 2.1: *Taxonomy of representation formats (adapted from Gigerenzer & Hoffrage, 2007).*

participants were able to provide the correct Bayesian solution for these types of problems. Compared to the 16 percent who were able to solve the conditional probability problems in their study, this represents a considerable improvement, an effect we refer to as the *natural frequency facilitation effect*.

2.1.1 Clarifying the Natural Frequency Facilitation Effect

Following Gigerenzer and Hoffrage's (1995) study, a number of authors critiqued the notion that natural frequency formats facilitated Bayesian inference based on a misinterpretation of the format. Initial misconceptions related to the information structure, that is the distinction between naturally sampled frequencies, drawn from the concept of natural sampling by Kleiter (1994), and frequencies that are

normalised or standardised. Natural sampling refers to the process by which one would naturally acquire information about events and their classes from experience, a sequential process where information is collated without fixing the marginal frequencies a priori (Gigerenzer & Hoffrage, 1995). In contrast, information formats that standardise or normalise information (see Figure 2.1A and 2.1B), do not preserve information about the base rates, which therefore require additional computation to incorporate them back into the calculation, as seen in Equation (2.2).

The natural sampling component was overlooked, rediscovered, and relabelled a number of times, often in order to incorporate the information structure within existing theoretical perspectives or to propose new theories (Gigerenzer & Hoffrage, 2007). For example, the natural sampling structure was introduced as the subset principle in the context of mental models theory (Johnson-Laird, Legrenzi, Girotto, Legrenzi, & Caverni, 1999); the conjunctive information structure was rediscovered as a partitioned information structure[2] by Girotto and Gonzalez (2001) and Macchi (2000); and the observation that the numerator is a subset of the denominator was described as facilitating the construction of a set-inclusion mental model by Evans, Handley, Perham, Over, and Thompson (2000). This latter point was followed up by Sloman, Over, Slovak, and Stibel (2003) in support of a more general nested-sets hypothesis, or the argument that any representation that makes the nested (subset) relations transparent will facilitate performance. The nested-set observation is evident in Gigerenzer and Hoffrage's (1995) Equations 2 and 3, or the observation that the numerator is a subset of the denominator (Gigerenzer & Hoffrage, 1995, 2007).

It is now generally understood that information structure is central to the facilitation effect (that is, the natural sampling structures in Figure 2.1C and 2.1D compared to the normalised structures of 2.1A and 2.1B). Nevertheless, nested-sets theory and the ecological rationality framework have been pitted against one another to debate which theoretical account provides the most homogenous explanation

[2] According to Girotto and Gonzalez (2001, p.250), a partitioned structure is one where "the problem statement partitions a set of units into exhaustive subsets (e.g., a set of 100 people is partitioned into two subsets: 4 infected people versus 96 uninfected people; these in turn are divided into two subsets: 3 persons with a positive versus 1 person with a negative test result, and 12 persons with a positive versus 84 persons with a negative test result)."

of findings to date. In the following section, we provide a brief overview of the ecological rationality framework and nested-sets theory to highlight points of distinction and to set the context for the different moderators of the natural frequency facilitation effect that have been proposed and that we review in our meta-analysis.

2.2 Theoretical Perspectives on Natural Frequencies

One aim of our meta-analysis was to review and quantify the predictions made by the ecological rationality framework and nested-sets theory in relation to the natural frequency facilitation effect. However, the many similarities in the predictions made by proponents of the two frameworks make it difficult to tease apart those that clearly differentiate the perspectives (see also, e.g., Brase and Hill, 2015, Johnson and Tubau, 2015, and McNair, 2015 for related observations). Admittedly, this problem was exacerbated by the fact that there is heterogeneity within the theoretical perspectives themselves and proponents differ as to the emphasis they place on certain concepts. In some cases, properties attributed to the theories have been imposed by others, often to the disagreement of the theory's proponents (see, e.g., comments on Barbey & Sloman, 2007). We review these theories with the general aim to highlight their contributions to the literature, summarise points of theoretical divergence, and to provide context for arguments made as to the relevance of different moderators. The meta-analysis ultimately moves beyond these theoretical debates and reviews the rich literature on the natural frequency facilitation effect to provide quantitative estimates and identify problem features or influential studies that can account for some of the variability in performance rates reported across studies.

2.2.1 Ecological Rationality Framework

The ecological rationality framework is a broad theoretical approach to the study of human cognition and decision-making (Gigerenzer, Todd, & the ABC Research Group, 1999) and has been applied to study of a range of topics including inference, choice, and group decision-making (Gigerenzer, Todd, & The ABC Research Group, 2012; Hertwig, Hoffrage, & the ABC Research Group, 2013; Todd & Brighton, 2015; Todd & Gigerenzer, 2007). Central to the framework is the concept of ecological rationality that emphasises the importance

of considering the match between the human mind and the structure of the environment as a fundamental unit of analysis (Gigerenzer, 1998; Gigerenzer et al., 1999). On the one hand, this involves the study of cognitive mechanisms (e.g., the *adaptive toolbox* of decision strategies) and on the other hand it involves the study of the environmental structures that determine which of these mechanisms will be successful (*ecological rationality*; Todd & Gigerenzer, 2007). For example, the framework has been applied to study the role of recognition knowledge in inference, showing that simple recognition knowledge can promote accurate inferences specifically in environments where recognition is a valid cue (e.g., the recognition heuristic; Gigerenzer & Goldstein, 2011; Marewski & Schooler, 2011). With respect to Bayesian inference with textbook problems, the ecological rationality framework analysed different types of information representations (i.e., conditional probabilities and natural frequencies) from a computational perspective: The proponents showed that, although these informations formats are informationally equivalent, they do not entail computationally equivalent cognitive algorithms (i.e., Bayesian computations).

In the context of Bayesian inference, the ecological rationality framework has been attributed to Gigerenzer and Hoffrage (1995) and the study by Cosmides and Tooby (1996). Proponents converge on the idea that an evolutionary perspective is useful for identifying the match between cognitive strategies and environmental structures, offering a framework for investigating the adaptiveness of strategies to environments (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 2007). The proponents agree that external representations (e.g., probability information formats) influence internal representations (e.g., cognitive algorithms). However, proponents differ as to their views on the modularity of cognitive mechanisms. According to Cosmides and Tooby (1996; 2008), an independent, domain-specific cognitive mechanism evolved to operate on frequency representations and, when given the appropriate content (e.g., frequencies), this frequency module would "allow certain computations to proceed automatically or 'intuitively' and with enhanced efficiency over what a more general reasoning process could achieve given the same input" (Cosmides & Tooby, 2008, p.66). As such, their perspective suggests an advantage for frequency representations more generally (e.g., also for normalised frequencies). In fact, in their widely cited paper on Bayesian facilitation often associated with the natural frequency

facilitation effect, Cosmides and Tooby (1996) did not test natural frequency formats but rather variations on normalised frequency formats (e.g., formats similar in structure to 2.1B). Specifically, the authors investigated the strength of different numerical format manipulations (e.g., presenting a problem using normalised frequencies but asking for a response as a conditional probability and vice versa) on Bayesian performance. Girotto and Gonzalez (2001) argued that the high performance achieved across studies in their paper is a consequence of the correct Bayesian response (2 percent) also potentially capturing participants who falsely divided the base rate by the total number of false positives, $1/50$ or $1/51$.

In contrast, claims about the modularity of cognitive mechanisms are generally not held in broader applications of the ecological rationality framework (Todd, Hertwig, & Hoffrage, 2005). Rather, Gigerenzer and Hoffrage's (1995) analysis of cognitive algorithms and information structures was computational and generated a priori theory-driven hypotheses about how cognitive processes map onto informational structures (the *ecological* aspect of the framework). In our view, the alignment between the theoretical perspective espoused by Gigerenzer and colleagues (Gigerenzer & Hoffrage, 2007) with the modularity view of Cosmides and Tooby (1996, 2008) has contributed to some of the critique of the ecological rationality framework from proponents of nested-sets theory.

Nested-Sets Theory

The most dominant alternative framework to account for the facilitative effect of natural frequencies on Bayesian inference is nested-sets theory. A general premise of the theory is that any representation that makes the nested-set structure of problems transparent will facilitate computations (Barbey & Sloman, 2007; Mandel, 2007). For example, one could use verbal description or visual representation to make the partitions and relations between relevant subsets clearer (e.g., a Venn diagram showing the relation between or nesting of subsets; Mandel, 2007). Nested-sets theory is founded in work on set relations in probability judgement and extensional reasoning. As such, its origins have been attributed to a variety of authors and theories (e.g., Barbey & Sloman, 2007; Girotto & Gonzalez, 2001; Johnson-Laird et al., 1999; Sloman et al., 2003; A. Tversky & Kahneman, 1983) and it claims to have broad applications to a variety of reasoning tasks

(Barbey & Sloman, 2007; Mandel, 2007). Sloman and colleagues (2003) were the first to use the term *nested-sets hypothesis* and outline the theory in relation to probability problems. According to them, the most elementary relations are those related to set inclusion and set membership, and should these be made transparent in a representation, the arithmetic operations that follow are easier to perform. In this regard, proponents of the theory have made a number of predictions about alternative nested-set manipulations that can facilitate performance on Bayesian inference tasks, such as enhancing the transparency of nested sets through visualisations (Yamagishi, 2003), and modifications to problem wording, such as improving the wording of text to show causal relations (Krynski & Tenenbaum, 2007; Sloman et al., 2003).

As nested-sets theory emerged, in part, in response to initial misconceptions about the natural sampling component of natural frequencies (outlined in the previous section), many of the current claims of the theory converge with the computational arguments initially put forward by Gigerenzer and Hoffrage (1995). In fact, at the time of Barbey and Sloman's (2007) review of research on base-rate neglect that incorporated work on natural frequencies, Mandel (2007) regarded nested-sets theory as "an assemblage of hypotheses, empirical findings, and rebuttals to theorists proposing some form of the 'frequentist mind' perspective" (p.275). Proponents also disagree as to whether the theory should be situated within a dual-process framework (Barbey & Sloman, 2007; Evans & Elqayam, 2007; Lagnado & Shanks, 2007; Mandel, 2007; Samuels, 2007).

Nevertheless, despite recent calls for theory integration (Brase & Hill, 2015; Johnson & Tubau, 2015; McNair, 2015), nested-sets theory continues to be pitted against the ecological rationality framework on the basis of a few general claims. Specifically, in an effort to differentiate domain general reasoning processes from the strong modular domain-specific processes attributed to Cosmides and Tooby (1996, 2008) and the modularity principle, the theory has explored the effects of individual difference measures such as general intelligence (e.g., education, cognitive abilities) and motivation (e.g., incentivised performance; Barbey & Sloman, 2007; Lesage, Navarrete, & De Neys, 2013; Lesage et al., 2013; Sirota & Juanchich, 2011; Sirota, Juanchich, & Hagmayer, 2014). However, the finding that general intelligence or motivation is relevant to problem solving is not surprising to some authors who argue against the use of such data to infer cognitive

structures or abilities (Brase, 2007; Brase & Hill, 2015; Trafimow, 2007). If evolutionary and dual-process claims were put aside, nested-sets theory and the ecological rationality framework appear to converge on the argument as to why natural frequencies support Bayesian inference: natural frequencies are a representation that provides a transparent information structure to simplify computations.

In summary, researchers from both nested-sets theory and the ecological rationality framework are interested in understanding the interaction or relation between external and internal representations of information (e.g., information formats and cognitive processes; Brase & Hill, 2015; Johnson & Tubau, 2015). Where divergence remains it relates to arguments as to how these external representations influence internal ones (e.g., domain-general or specific cognitive mechanisms). A positive side of this debate has been the proposition of a variety of potential moderators to the natural frequency facilitation effect, which we discuss in detail below. The present meta-analysis quantifies the effect of these moderators with the aim to identify those features of the problem representation, methodology, and the individual that are most influential and can account for the substantial variability in performance across studies.

2.3 When, Why and for Whom is the Natural Frequency Facilitation Effect Most Likely to Occur?

Potential moderators of the effect are those that manipulate the problem representation (e.g., inclusion of visual aids), alter the methodological procedure (e.g., use of incentives), or account for aspects of the individual (e.g., expert or non-expert participants). Moderators related to problem representation can provide an understanding of the features of the format most influential to the effect. Methodological moderators are generally overlooked in debates about the facilitation effect, but nonetheless can contribute to the variation in the size of the effects reported across studies. Individual difference moderators can improve our understanding of the basic competencies of individuals that underlie effects and can help to tailor problem representations to different audiences. Previous reviews of this work have been qualitative, and although these have sought to provide a comprehensive overview of the literature to date, the conclusions of some of these reviews tend to support opposite theoretical viewpoints

(Barbey & Sloman, 2007; Brase & Hill, 2015; Johnson & Tubau, 2015). A quantitative estimate of the magnitude of effects from different manipulations is currently lacking.

The current meta-analysis fills this gap, not only by isolating the effects of different moderators but also by providing quantitative estimates of the magnitude of effects. As such, the present meta-analysis focuses on those studies that have simultaneously compared performance for natural frequency and conditional probability formats. In the following sections, we review a broad range of potential moderators that have been put forth over the past two decades. In some cases, studies have examined the effect of moderators on the performance rates for one format only. We did not include these studies as they were more likely to differ in other ways that could not be controlled for in the analyses. While these studies cannot be included in the meta-analysis, we do mention these below as the moderators could still be coded for. We emphasise that not all moderators mentioned below were able to be included or coded for in our meta-analysis and we specify why this is the case in the respective sections. Nevertheless, we think it is important to review all potentially relevant moderators to provide context for those that were included. Each potential moderator is indicated with italics within the following sections. Moderators that were able to be included in the meta-analysis are listed in Table 2.1.

2.3.1 Problem Representation

The most widely debated aspects of the natural frequency facilitation effect relate to the representation of problems, and the different ways of presenting information in natural frequency, conditional probability, or both formats. Over the years, researchers have modified formats by adding or substituting information, or by altering the computational complexity of the problems to tease apart potential explanations for the effect. While there are many aspects of the problem representation that have been explored, we attempt to address those features that have received most empirical attention.

Computational or problem complexity. The argument that natural frequency formats are less computationally demanding than conditional probability problems is generally not debated and in fact, this was the only argument made in Gigerenzer and Hoffrage's

Table 2.1
Summary of Study Characteristics

		counts	
Name	Description	no	yes
short menu	short menu used in both formats	199	27
three hypotheses	problem complicated by third hypothesis	208	18
two or more cues	problem complicated by additional cues or cue values	212	14
probability question	probability question asked in natural frequency format	206	20
frequency question	frequency question asked in conditional probability format	216	10
enumerated population	enumerated population given for conditional probability	198	28
multiple events	conditional probability phrased in terms of multiple events	162	64
visual aid	visual aid used in both formats	200	26
base rate*	base rate in problem solved by participants	numerical	
hit rate*	hit rate in problem solved by participants	numerical	
false-alarm rate*	false-alarm rate in problem solved by participants	numerical	
show-up fee*	participants are paid a show-up fee	50	115
performance pay*	participants are paid a by performance	153	12
strict scoring	correct answer based on accuracy and/or protocol	193	33
within-subject*	within-subjects study design	212	14
both formats	both formats solved by each participant	174	52
additional problems	number of additional problems solved by each participant	numerical	
high numeracy*	participants have a high numeracy scores	26	26
experts	educated participants (students & professionals)	38	188

Notes: Study characteristics marked with * were coded but not able to be included in the full-sample meta-analysis because of the lack of a sufficient number of logits; where needed, we conducted separate subset-analyses that did not control for the larger set of study characteristics.

(1995) computational analysis. Specifically, to demonstrate that the computational advantage of natural frequencies lay in their more parsimonious segmentation of information, Gigerenzer and Hoffrage introduced the *short menu* of conditional probability and natural frequency formats. The short menu versions displayed only two pieces of information: $p(D)$ and $p(D \cap H)$ for the probability format and $n(D)$ and $n(D \cap H)$ for the natural frequency format. For example,

the probability problem stated: "The probability that a woman at age forty will get a positive mammography in routine screening is 10.3%. The probability of breast cancer and a positive mammography is 0.8% for a woman at age forty who participates in routine screening". For the conditional probability version, the joint probabilities simplified the computational algorithm, $p(H|D) = p(D \cap H)/p(D)$, similar to Equation (2.1), while for the natural frequency version the computational algorithm was the same as Equation (2.3) above, albeit with the sum in the denominator already calculated[3]. As reported in the results of their Study 1, Gigerenzer and Hoffrage found that when compared to the standard versions, the short menu versions resulted in improved performance for conditional probability formats, but not for natural frequency formats. Thus, the study showed that lowering computational complexity may explain the advantage of natural frequency formats over more computationally complex conditional probability formats. However, we note that in their study, short menu versions of natural frequency formats continued to outperform short menu versions of conditional probability formats despite similar computational algorithms: the natural frequency format improved from 46 percent for the standard format to 50 percent in the short format, and the conditional probability format improved from 16 percent to 28 percent correct solutions across problems.

While there is general agreement that short menu versions facilitate performance because there are fewer and easier calculation steps (Ferguson & Starmer, 2013; Mellers & McGraw, 1999), in particular for conditional probability formats, some authors have proposed alternative reasons for their advantage. For example, Mellers and McGraw (1999) suggest that presenting joint events makes the nested sets easier to visualise, thus facilitating computation. Fiedler et al. (2000) suggested that joint frequency and probability versions improved performance because these formats used a common reference scale (e.g., 8 *out of 1000* women; 95 *out of 1000* women), although we are unsure how the argument about the advantage of a common

[3]Mellers and McGraw (1999), Lesage et al. (2013), Fiedler, Brinkmann, Betsch, and Wild (2000) utilised a slightly different version of the short menu than Gigerenzer and Hoffrage (1995) which is known as the *joint menu*, where instead of providing $p(D)$, participants were provided with $p(H \cap D)$ and $p(\neg H \cap D)$ and they had to make the additional calculation: $p(D) = p(H \cap D) + p(\neg H \cap D)$ themselves. The natural frequency version is essentially the same, with information presented only in terms of joint frequencies with the complements provided.

reference scale provides any additional explanation for this effect. Unfortunately, Fiedler et al. (2000) misinterpreted the natural frequency format used in Gigerenzer and Hoffrage (1995), suggesting that they had compared the short menu version to the standard conditional probability version and thus, this was responsible for the facilitation effect. Not surprisingly, Fiedler and colleagues found that short menu versions resulted in better performance than standard versions. Admittedly, it is difficult to tease apart theoretical arguments for the enhanced performance on short menu versions: making the nested set relations transparent is argued to require simpler computational algorithms, a prediction that was originally made by Gigerenzer and Hoffrage (1995). We distinguish between short and standard menu versions in the coding of studies for the meta-analysis, to determine how much the facilitation effect is reduced as a result of the simpler computational structure.

To further address arguments regarding computational complexity, a number of studies have explored how manipulating the number and type of calculation steps affects performance. For example, Krauss, Martignon, and Hoffrage (1999) and Hoffrage, Krauss, et al. (2015) have argued that the information structure should still provide an advantage to natural frequencies even as Bayesian computations become more complex with an increasing numbers of cues and hypotheses. That is, although overall performance would be expected to decrease with additional computational complexity, the natural frequency facilitation effect should still hold given multiple cue values (e.g., a medical test can return a positive, negative, or unclear result; in this case *three cue values*), multiple hypotheses (e.g., three diagnoses are considered; in this case *three hypotheses*), more than one cue (e.g., multiple tests, such as a mammography and ultrasound; in this case, *two or more cues*). Girotto and Gonzalez (2001) reported that creating an additional mathematical *calculation step* by providing $n(\neg H \cap \neg D)$ rather than $n(\neg H \cap D)$ for the natural frequency format, reduced performance rates (in their study, from 53 percent to 35 percent).

In this connection, Ayal and Beyth-Marom (2014) manipulated the number of calculation steps required to solve natural frequency problems by providing participants with different pieces of relevant or irrelevant numerical information and found that performance decreased as the number of steps increased (e.g., from 55 percent with one step to 10 percent with four steps). Unfortunately, similar manipulations for conditional probability formats are often not tested,

and the relative detriment to performance cannot be compared. We intended to code for calculation steps in our meta-analysis, however, we found that it was not clear what would constitute a single mental step or if different mental steps should have equal weight (e.g., simple arithmetic versus recalculation or transformation). We do note that in one study minor manipulations to the numerical complexity of natural frequency versions by using large numbers or numbers not of multiples of 10 did not appear to reduce the facilitation effect (Misuraca, Carmeci, Pravettoni, & Cardaci, 2009). For our analysis, we code the number of cues and hypotheses within problems as an indicator of computational complexity.

A related manipulation has been to separate the problem format and question format such that the respondent is required to translate the information in the problem (e.g., natural frequencies) into a response for the alternative format (e.g., provide the answer as a probability). A number of studies have directly or unwittingly tested whether the facilitation effect is reduced if there is incongruence between problem and answer format (Ayal & Beyth-Marom, 2014; Evans et al., 2000; Fiedler et al., 2000; Girotto & Gonzalez, 2001). When the question asks for a numerical probability response, (i.e., *probability question*: What is the probability that a woman actually has breast cancer? _____%) this tends to reduce performance on natural frequency problems (Ayal & Beyth-Marom, 2014), yet when the question asks for a numerical response using frequencies or pairs of integers (i.e., *frequency question*: How many of these women do you expect to actually have breast cancer? _____ out of _____), the effect on performance for probability problems is mixed (Cosmides & Tooby, 1996; Evans et al., 2000). Few studies directly test incongruent problem and question formats for both types of problems, instead they focus on frequency or probability problems only (Ayal & Beyth-Marom, 2014), or compare the effect of congruence between problem and question formats on probability and normalised frequency versions (e.g., Cosmides & Tooby, 1996; Evans et al., 2000). Nevertheless, the results of these studies suggest that a numerical response that does not match the information format of the problem reduces the natural frequency facilitation effect (Johnson & Tubau, 2015). Accordingly, it is expected that incongruence would disadvantage whichever format is made incongruent and thus, this may explain some of the variability in performance rates across studies. Where a direct comparison between natural frequency and conditional probability formats includes a

manipulation of congruence for one or both formats, we include this moderator in the meta-analysis.

A number of other manipulations have been explored across studies, however in many cases the manipulations have been tested on a single format only or slight modifications to the formats have not been investigated systematically. Lesage et al. (2013) found that providing a total sample size or *enumerated population* on which to make calculations (e.g., "the study contains data from 8,500 children") did not enhance performance for probability versions, despite some indication that a reference class facilitates judgements on other Bayesian inference tasks (Neace, Michaud, Bolling, Deer, & Zecevic, 2008). Other studies state an enumerated population in probability versions but do not systematically test this format modification on performance rates (e.g., Fiedler et al., 2000; Konheim-Kalkstein, 2008; Macchi, 2000). However, meta-analyses can address this question by pooling estimates from different studies. It is unclear whether or not an enumerated population should enhance or decrease performance in the conditional probability format.

One final manipulation to problem complexity we think is worth mentioning involves textual modifications intended to increase or decrease problem comprehension. Increasing *verbal complexity* was shown to reduce performance (Johnson & Tubau, 2013), suggesting a role for basic text comprehension abilities in performance on word problems. Results of manipulations aimed at providing a *causal explanation* of a false alarm (e.g., that a benign but harmless cyst could cause a false-positive mammography result) to clarify the nested set structure or to facilitate the construction of a causal model (and therefore the determination of relevant model parameters) to use in solving conditional probability problems are mixed (Krynski & Tenenbaum, 2007; McNair & Feeney, 2014, 2015; Sloman et al., 2003). Further, when Moro, Bodanza, and Freidin (2011) tested one of these causal explanation versions against a natural frequency format, they found no effect of wording on performance rates for this manipulation. Unfortunately, as few studies manipulated and tested performance across formats, the relative advantages of such manipulations from one format to another were not able to be examined systematically.

Multiple events. In earlier work on the natural frequency effect, the distinction between a conditional probability and single-event

probability was often confused such that natural frequencies were thought to be an alternative to single-event probabilities (Barbey & Sloman, 2007; Cosmides & Tooby, 1996). However, conditional probability problems could be represented in terms of single events (e.g., probability between 0 and 1; chances of one event) or as *multiple events* (e.g., relative frequencies, see Figure 2.1A but with the phrasing "1% of women"), provided the information structure is normalised. To emphasise the different dimensions of statistical representations, Barton, Mousavi, and Stevens (2007) outlined a statistical taxonomy whereby representations can vary according to three orthogonal dimensions: number of events (single event or sets of events), numerical format (percentages, probabilities, fractions, or pairs of integers), and information structure (normalised or conjunctive). Thus, the statistical information in Bayesian inference tasks can be represented in multiple ways by crossing these dimensions.

To illustrate that the computational advantage of natural frequencies was related to the information structure rather than the number of events per se, Gigerenzer and Hoffrage (1995) demonstrated that the natural frequency facilitation effect did not extend to relative frequency versions of the conditional probability problems in their Study 2 (the relative frequency format tested by Gigerenzer & Hoffrage, 1995, was similar to Figure 2.1A but was presented in percentages and referred to women rather than a single woman). In further support of the relevance of information structure to the facilitation effect, Girotto and Gonzalez (2001) tested a representation that expressed chances of a single event, as pairs of integers (10 in 100 chances), in a format that mimicked the natural sampling structure of natural frequencies (compare C and D in Figure 2.1) and found that performance rates were similar for chance and frequency versions. However, simply using the term chance to represent problems does not seem to be sufficient to improve performance. Brase (2008) tested natural frequencies against chance representations with a natural sampling or normalised structure to distinguish the effect of multiple events and information structure. Brase found that chances with natural sampling improved performance compared to normalised chances but that natural frequency formats were still the superior representation. People do appear to understand the distinction between probability and frequency (Girotto & Gonzalez, 2002), however, participants who interpret chances as frequency representations tend to perform better on these problems (Brase, 2008).

From a computational perspective, proponents of the ecological rationality framework argue that chances with natural sampling mimic the computations for natural frequencies in Equation (2.3) and thus, performance should be improved relative to normalised versions (Gigerenzer & Hoffrage, 2007; Hoffrage, Gigerenzer, Krauss, & Martignon, 2002). However, other authors disagree as to whether the adoption of the chance formulation as an extension of the natural frequency format fits within the ecological rationality framework and its evolutionary claims about the acquisition of information in the form of frequencies experienced as a series of events (Barbey & Sloman, 2007; Girotto & Gonzalez, 2002). Rather, proponents of nested-sets theory argue that, like natural frequency formats, chance formulations are effective because they make the set structure more transparent (Girotto & Gonzalez, 2001). Given the different ways in which information structure, numerical format, and number of events can differ in the problem representations used across studies, the current meta-analysis will code each of these three dimensions to discern those features of the representation most influential for facilitation effects. As the chances with natural sampling format is not normalised, it is fundamentally different from regular probability formats and accordingly, we do not examine these studies together with standard probability formats but conduct a separate analysis comparing them to natural frequency formats. Owing to the similarity in the natural frequencies and chances with natural sampling formats in terms of information structure and numerical format, it is expected that the performance is similar when comparing these formats, but we anticipate that there may still be an added advantage of frequencies, consistent with the work reviewed (see also Gigerenzer & Hoffrage, 2007; Hoffrage et al., 2002).

Calculations. As noted in the preceding sections, the formulation of the Bayesian inference problems can differ in a variety of ways and some authors have suggested that the specific numerical values for the *base rate*, *hit rate*, and *false-alarm rate* have the potential to influence performance rates (see, e.g., Mellers and McGraw's, 1999, and Gigerenzer and Hoffrage's, 1999, discussion and reanalysis of the cab problem from Gigerenzer and Hoffrage, 1995). In a recent assessment of the influence of numeric task characteristics on solution rates, Hafenbrädl and Hoffrage (2015) examined how the

base rates, hit rates, and false-alarm rates of 19 different Bayesian inference problems influenced performance rates using data from the 15 problems in Gigerenzer and Hoffrage (1995) and 4 problems from Hoffrage, Hafenbrädl, and Bouquet (2015). In general, they found that while higher base rates and hit rates were associated with a higher number of Bayesian solutions, higher false-alarm rates were associated with a lower number of correct responses. Further, probability formats were more affected by these factors.

In this connection, Gigerenzer and Hoffrage (1995) explored how different cognitive shortcuts could approximate Bayesian solutions under certain task conditions and proposed reasons for why higher base rates and hit rates are associated with better performance. For example, when the base rate, $p(H)$, is small (and $p(\neg H)$ is close to one), then the rare-event shortcut can be applied to simplify computation by approximating $p(D|\neg H) \times p(\neg H)$ by $p(D|\neg H)$. Few studies have explored alternative solution strategies or the environments in which they work well (Gigerenzer & Hoffrage, 2007). A few studies have attempted to classify the most common errors on Bayesian inference problems from supportive protocols (that is, written or verbal protocols accompanying solutions to word problems; Hoffrage & Gigerenzer, 1998; Siegrist & Keller, 2011; Zhu & Gigerenzer, 2006), however, few explore the connection between solution strategies and the specific features of the problems. As many studies report outcomes aggregated across problems, we cannot examine the specific effects of the base rate, hit rate, and false-alarm rate on performance, systematically. However, we believe the recent analysis by Hafenbrädl and Hoffrage (2015) sufficiently addresses these points.

Visual aids. Of the manipulations intended to bolster performance on Bayesian inference problems, the addition of a *visual aid* has been the most broadly tested (see Figure 2.2 for examples of the different types of aids tested across studies). Visual aids have been utilised to examine whether performance can be improved further by helping participants to visualise relations between the different pieces of information[4]. Specifically, some authors have argued that if one makes the nested-set structure transparent through visual aids (e.g., one can see how different events are related to one another), then

[4]Note that we use the term visual *aid* to denote that the visual is presented alongside or in addition to the text, not as a replacement.

this modification can reduce the natural frequency facilitation effect by enhancing performance for conditional probability formats (e.g., Sloman et al., 2003; Yamagishi, 2003). Others have argued that visual aids, shown to be effective in communicating probability information in other domains (e.g., health risks), can further enhance the effect of natural frequencies (Garcia-Retamero & Hoffrage, 2013). The relevance of visual aids for helping reasoners solve Bayesian inference problems was not only evident in the use of pictoral analogs by participants in Gigerenzer and Hoffrage (1995), but also in early work by Cole (1988) who explored different possibilities for visualising laboratory results to aid Bayesian reasoning. Generally, visual aids improve performance across formats (Brase, 2009a; Garcia-Retamero & Hoffrage, 2013; Sloman et al., 2003; Yamagishi, 2003). However, the theoretical arguments for when and why visual aids should or should not facilitate are mixed. Arguments attributed to the ecological rationality framework, typically the account by Cosmides and Tooby (1996), include that visual aids should facilitate performance when they incorporate discrete, individualised events that can be counted or when iconicity is high (i.e., greater similarity between the icon and the object or event it represents) because these help to tap into the frequency-encoding mechanism (Brase, 2009b, 2014; Sirota, Kostovičová, & Juanchich, 2014). Nested-sets theory proponents argue that it is the nested-set component of visualisations that facilitate performance, or that pictures can boost performance by drawing on people's visual computation abilities (Yamagishi, 2003).

To address some of these predictions, a number of studies have sought to illustrate that not all visual representations are equally effective (Brase, 2009b, 2014; Moro et al., 2011; Sirota, Kostovičová, & Juanchich, 2014). Different visualisation techniques have been employed, some of which may indicate a relative effect size indicated by the size of elements in the display (see, e.g., Figure 2.2B icon arrays or frequency grids and Figure 2.2D roulette wheel) and others that simply reveal set structure (see, e.g., Figure 2.2A frequency or probability trees). Moro et al. (2011) criticised the use of visual aids that confound clarification of set structure with the relative size of an effect. Testing two visual aid designs that did not reveal the relative size of an effect, (e.g., a Venn diagram that represented sets without reference to the relative size of the elements), Moro et al. (2011) found that these visual aids did not significantly improve performance for natural frequency or conditional probability formats. Rather,

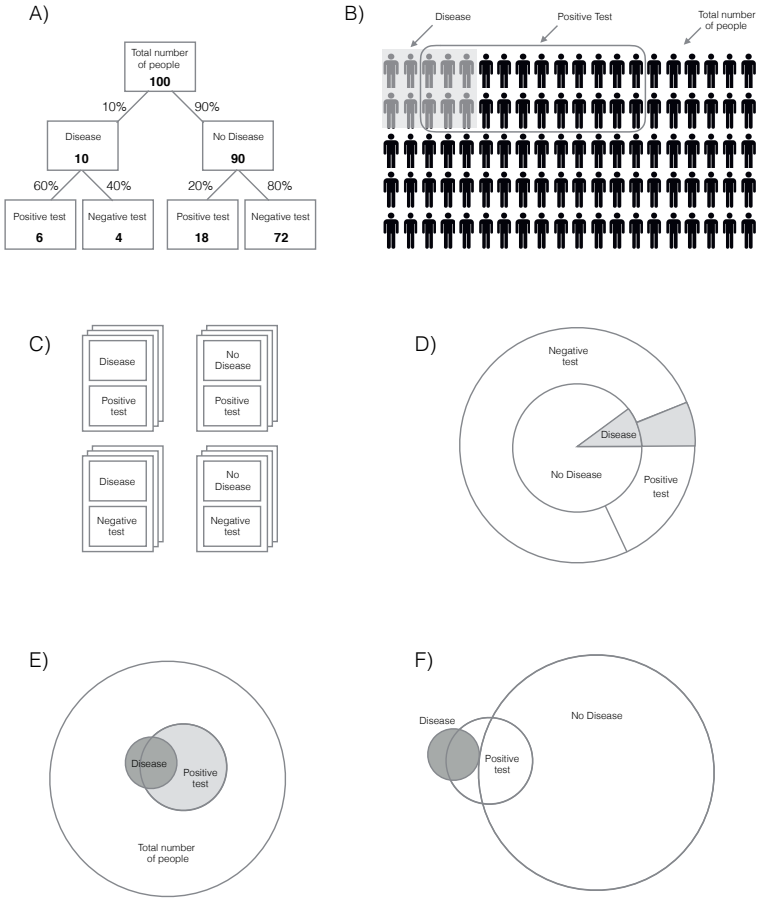

Figure 2.2: *Examples of visual aids for conditional probability and natural frequency problems. A) Natural frequency or conditional probability tree (see, e.g., Binder, Krauss, & Bruckmaier, 2015; Sedlmeier & Gigerenzer, 2001). Numbers in tree represent natural frequencies and numbers beside branches represent probability versions. B) Icon array/frequency grid (e.g., Brase, 2009a, 2014). C) Interactive cards (e.g., Vallée-Tourangeau, Abadie, & Vallée-Tourangeau, 2015). D) Roulette wheel (e.g., Brase, 2014; Yamagishi, 2003). E) Euler diagram and F) Area proportional Euler diagram (e.g., Micallef, Dragicevic, & Fekete, 2012; Sloman et al., 2003).*

performance on natural frequency formats was consistently better than for conditional probability formats regardless of the use of visual aids. Further, as most participants in both natural frequency and conditional probability conditions were able to accurately identify the set structure of problems illustrated with the visual aids, this suggests that the visual aids used in the study did not make the set structure of problems unclear. The present meta-analysis compares studies that use visual aids for both formats to those that do not use visual aids in order to establish whether visual aids reduce or enhance the natural frequency facilitation effect. We also code the specific type of visual aid employed.

2.3.2 Methodological Factors

Methodological or procedural factors may account for some of the variation across studies and a few studies have attempted to explore their effect on performance rates. For example, incentives to remunerate or motivate participants to complete a study (e.g., financial incentives, course credit) can enhance performance, particularly in cases where incentives are tied to performance (i.e., performance-based incentives; Cerasoli, Nicklin, & Ford, 2014). Brase (2009a) found performance-based incentives (*performance pay*) to be more effective than a *show-up fee* or course credit incentive structures at enhancing performance on Bayesian inference problems; there was no difference in performance for the latter two incentive structures. However, there are contradictory findings regarding the effect of incentives on performance according to presentation format. Sedlmeier and Gigerenzer (2001) found participants given rule-based training (Bayes rule) for solving conditional probability problems maintained high, stable performance at follow-up only when incentivised whereas participants receiving natural frequency training maintained performance regardless of incentives. Brase (2009a) found the effect of incentives was more evident for Bayesian inference problems of intermediate difficulty, notably variations of natural frequency formats, and not for the more difficult conditional probability formats. In contrast, Ferguson and Starmer (2013) found a main effect of incentives on performance but no interaction with presentation format. Given these mixed results and that theoretically it is not clear how incentives should influence performance across formats, we examine the incentive structure reported across all studies in the meta-analysis.

Multiple authors have suggested that the way in which Bayesian solutions are coded as correct may account for some of the variation across studies (Girotto & Gonzalez, 2001; McNair, 2015; McNair & Feeney, 2014). Studies vary as to whether a strict (point estimate) or more lenient estimate (point estimate plus or minus x percentage points) is used to classify correct Bayesian responses, and whether a supportive verbal or written protocol is required to determine Bayesian reasoning. Supportive protocols are used to classify marginally incorrect estimates that suffer from minor calculation errors as correct, or to reclassify as incorrect those responses that indicate a correct guess but that have not followed a Bayesian reasoning approach. In Gigerenzer and Hoffrage's (1995) initial study, supportive protocols were used as an additional requirement for determining the use of Bayesian reasoning and to provide insight into the types of errors that participants made, thus distinguishing between outcome and process. We are aware of only one study that compared performance rates given different scoring criteria. McNair and Feeney (2014) found that a *strict scoring* (an exact estimate) resulted in near-negligible rates of correct responses when compared with scoring that allowed for calculation errors (within 5 percentage points of the correct estimate). The present meta-analysis attempts to address whether scoring criteria can account for some of the variation found across studies.

Given the wide variety of study designs and procedures used across studies on natural frequencies, we also consider other methodological factors that could account for variation in effects. For example, we consider the potential for learning or practice effects to influence performance rates and account for whether a *within-subjects* design was employed, whether participants solved a single or *both formats* (e.g., natural frequency and conditional probability formats), and record the total number of problems (*additional problems*) participants were required to solve.

2.3.3 Individual Differences

Characteristics of the sample or the individual have been investigated in an attempt to determine for whom the natural frequency format is most effective. Some of these individual-level moderators have been investigated with the view that they reveal insights into the nature of the cognitive mechanisms involved (e.g., general purpose

or domain-specific; see Lesage et al., 2013; Sirota, Juanchich, & Hagmayer, 2014). Others have been explored in an attempt to elucidate boundary conditions or to identify characteristics of individuals that contribute to variations in performance (e.g., numeracy, education; see Brase, Fiddick, & Harries, 2006; Johnson & Tubau, 2013). For instance, exploring the natural frequency facilitation effect in different lay and expert populations can provide insights into how expertise, education, cognitive ability and development contribute to performance differences across formats and studies.

Individual differences have generally been explored in relation to differences in cognitive capacity or ability, for instance, in terms of educational achievement or cognitive development across the lifespan, to the importance of basic numerical skills or the role of expertise. There is little surprise that higher general intelligence or cognitive ability (i.e., cognitive reflection, thinking disposition) has been linked with improved performance across formats for Bayesian inference tasks (Lesage et al., 2013; McNair, 2015; Sirota & Juanchich, 2011; Sirota, Juanchich, & Hagmayer, 2014). Nevertheless, even amongst individuals with higher ability there are substantial variations in performance across studies and considerable room for improvement. In this connection, numerous studies have explored whether a basic level of numeracy is needed for the facilitation effect to emerge (e.g., the threshold hypothesis; see Hill & Brase, 2012) or whether natural frequencies can facilitate performance even for individuals with low numerical ability (Galesic, Gigerenzer, & Straubinger, 2009). While high numerates tend to perform better across formats, results are mixed as to whether the effect of numeracy is independent to that of information format (Chapman & Liu, 2009; Hill & Brase, 2012; Johnson & Tubau, 2013).

Similar results are found in relation to the effect of educational experience on performance (Brase et al., 2006; Siegrist & Keller, 2011). For instance, even medical professionals and students, who are exposed to conditional probabilities in medical textbooks and curricula (e.g., test statistics, such as positive predictive value), have been shown to benefit from natural frequency formats (Friederichs, Ligges, & Weissenstein, 2014; Hoffrage & Gigerenzer, 1998), with some suggestion that professionals may benefit more than lay audiences (Bramwell, West, & Salmon, 2006). A number of studies have also shown that natural frequency formats facilitate performance on Bayesian inference tasks for both children (Zhu & Gigerenzer, 2006) and adolescents

(Lesage et al., 2013), suggesting that the computational advantage is evident even at early stages of cognitive development. Galesic, Gigerenzer, and Straubinger (2009) also showed that natural frequencies facilitate performance in the elderly, who can face age-related cognitive declines in numerical reasoning abilities. Studies investigating training effects suggest performance on natural frequency formats is more robust over time (Kurzenhäuser & Hoffrage, 2002; Ruscio, 2003; Sedlmeier & Gigerenzer, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a), indicating the potential to further build on cognitive ability with training.

In summary, general cognitive ability is associated with improvements in performance on Bayesian inference problems across a range of individual difference factors. However, it is unclear whether cognitive abilities increase the facilitation effect of natural frequencies or simply improve performance across tasks irrespective of format. Accordingly, we seek to quantify to what degree these individual characteristics contribute to performance differences across studies. Specifically, we code for a variety of individual differences across studies, including sample characteristics (e.g., educational experience, or *expertise*) and whether the study made comparisons based on cognitive ability measures (e.g., *numeracy*). We anticipate that cognitive ability and educational expertise will increase performance rates for both formats and therefore it is unclear how these will influence the overall facilitation effect.

2.3.4 Overview of Meta-Analysis

As evident in our review, a variety of potential moderators have been proposed to account for the variation in the natural frequency facilitation effect found across studies. Some of these moderators have been introduced to explore or test specific theoretical accounts (e.g., cognitive ability, short menu format), while others have been proposed as factors to account for the variation in the size of effects across studies (e.g., scoring criteria). Predictions as to how each moderator should affect the size of the natural frequency facilitation effect are not always possible, because of mixed results and/or theoretical arguments in the literature (e.g., numeracy) or because studies have manipulated some moderators without testing them explicitly. Even in cases where moderators were not tested within any given study, meta-analyses are able to use the between-study variation to estimate

the effect of such moderators, provided sufficient variation. Accordingly, the meta-analysis examines whether each moderator further enhances performance for natural frequencies, whether performance is enhanced across formats, or whether performance improves on conditional probability formats such that the relative difference in performance between formats is reduced. Table 2.1 summarises the moderators that were coded for in the present meta-analysis.

2.4 Method

2.4.1 Literature Search and Inclusion Criteria

Multiple methods were utilised to identify relevant papers for the review. A literature search of relevant databases (PsycINFO; PubMed; and Web of Science) was conducted on the following search terms (with thesaurus to explode terms, if available): bayes* AND conditional probab*, natural frequenc*, nested set, OR (reason* OR inference). Cited reference searches were conducted on key theoretical or review papers in the field (Barbey & Sloman, 2007; Cosmides & Tooby, 1996) as well as on Gigerenzer and Hoffrage's (1995) initial paper. Finally, a message was posted to the Society for Judgment and Decision Making (JDM) list requesting any additional publications, published or unpublished. No language restrictions were applied explicitly. The search covered papers published until December 2015, inclusive.

Studies were considered eligible for inclusion if they directly compared natural frequency and conditional probability formats. Studies that examined performance rates for different moderators on one format only were excluded. For studies that included multiple variations of different formats, the formats that were most equivalent to one another were selected for comparison (e.g., when both formats included a visual aid). If the moderator applied to one format only (e.g., the inclusion of an enumerated population for probability conditions), the effect was included and coded accordingly (see Table 2.1). Where more than one comparison was possible, for example when a natural frequency format was compared with two different conditional probability versions (typically normalised frequencies), the effect for the two most equivalent formats were compared, or two separate effects were included and then multiple comparison effects were controlled for in the analysis. We included studies that

compared natural frequencies and chances with natural sampling, however these were analysed separately for the reasons specified in the review and again below (see section on *Coding of Moderators*; Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a). For the meta-analysis the sample sizes and performance rates for each format were required. When these were not reported in the publication, we attempted to obtain the relevant data from the corresponding author.

2.4.2 Screening for Eligible Studies

The database search yielded 213 relevant abstracts of which 35 reported on studies that met the inclusion criteria. The primary reasons for excluding studies were that they did not report an experiment (65 papers; 37% of excluded papers), studied a different type of probability problem (e.g., not conditional probabilities; 49 papers, 28%) or did not include one of the relevant formats for comparison (e.g., tested only natural frequency or probability problems; 44 papers, 25%). One additional paper was identified from the cited reference search, two additional papers were retrieved from the JDM mailing list (one currently unpublished), and one additional paper was sourced from the bibliography of another relevant paper during coding (the paper was published in a law journal and was not indexed for retrieval from the database search). For two of the relevant papers, an effect size could not be derived from the results or retrieved from the authors (Garcia-Retamero & Hoffrage, 2013; Kochetova-Kozloski, Messier, & Eilifsen, 2011) and these could not be included in the meta-analysis. Three papers compared natural frequencies and chances with natural sampling formats (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015b). As two of these papers did not also include a conditional probability format as a comparison (Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015b), these papers were coded separately and all chances with natural sampling conditions were analysed in a separate analysis (see subsection *"Subset analysis: Chances with natural sampling"*). Thus, a total of 35 papers that compared natural frequency and conditional probability formats were finally included in the meta-analysis.

2.4.3 Coding of Moderators

Each study was coded according to a coding manual that was developed on the basis of the literature reviewed. Binary variables were coded for the presence (1) versus absence (0) of the moderator. Basic study characteristics included year of publication (and publication status), sample size per group, sample characteristics (e.g., age and gender; these were not always available), and study design (within- or between-subject). Coding of papers was completed by research assistants and authors, with any disagreements resolved through discussion. One sixth of papers were double-coded and for each variable the agreement assessed using Cohen's κ, with a median value of 1 and an average of .89.

Problem representation. Problem representation moderators were recorded for each format separately (e.g., question format used in natural frequency format and in probability format), unless the moderator applied to only a single format (e.g., as in the case of an enumerated population). In studies where multiple problems were tested but the details for all problems were not provided, we made the assumption that all problems contained the same information, structure, and other representation factors (e.g., numerical format). Information structure or *short menu* was coded as either (0) for standard or (1) for short menu. A problem was coded as short menu if it presented only joint events, either with the denominator already calculated, as in Gigerenzer and Hoffrage's 1995 original study: $p(H \cap D)/p(D)$; or requiring a simple calculation as in Mellers and McGraw (1999): $p(H \cap D)/[p(H \cap D) + p(\neg H \cap D)]$. For the latter studies, the natural frequency version presented all joint frequencies (e.g., hit rate, false-alarm rate, and their complements) relative to a total sample rather than to subsets and the computation was the same as in standard versions. We coded studies that compared both formats in short menu, or both formats in standard menu. Task complexity was coded for the number of hypotheses, cues, and cue values presented in the problem. We set out to code the following variations: (a) two hypotheses, single dichotomous cue; (b) two hypotheses, two dichotomous cues; (c) three hypotheses, single dichotomous cue; (d) two hypotheses, single cue with 3 cue values; and (e) two hypotheses, three dichotomous cues. As only one study (Hoffrage, Krauss, et al., 2015) examined a single problem for

categories (d) and (e) (representing two effects each), and only two additional studies (Hill & Brase, 2015; Krauss et al., 1999) examined category (b) problems (representing 10 effects), these were combined under the more general *two or more cues* coding category, coded as present (1) or absent (0). This category represents more complex problem representations where two or more cues or cue values were examined (for example, multiple medical tests or a single test with multiple test results). *Three hypotheses* refers to case (c) where the problem included three hypotheses (e.g., one of three diagnoses) with a single dichotomous cue (e.g., a single medical test), and was also coded as present (1) or absent (0).

To account for any variation in effects attributable to the number of events referred to in probability formats, we coded moderator *multiple events* as (0) if the problem referred to the probability of a single event (e.g., the probability that a woman has breast cancer; chances that a student passes a test) and (1) if numbers referred to sets of events (e.g., the relative/absolute frequency of women having breast cancer). As all natural frequency problems relate to multiple events, this category only applied to the coding of the probability formats. We coded the numerical format used within each problem. However, there was minimal variation in the numerical formats used for the problems and questions. For conditional probability formats, only normalised frequency formats used pairs of integers (representing only 4 comparisons) whereas the remaining probability formats used percentages in the majority of cases, and in some cases fractions, or mixed numerical formats. Given the minimal variation in these numerical formats, we did not consider this moderator any further, except for cases in which an incongruent question format was used. In cases where a probability question was asked for the natural frequency format, we coded *probability question* as (1), and (0) otherwise. Similarly, when probability formats required a frequency response in pairs of integers, we coded *frequency question* as (1) and (0) otherwise.

An *enumerated population* (or reference class) was coded as present (1) if the probability problem included an enumerated population as part of the problem description, that is, a sample size was provided for the participant to use in their calculations (e.g., "the data refers to 1,000 women"), otherwise it was coded as (0). For each estimate, the *base rate*, *hit rate*, and *false-alarm rate* used within the problem was recorded, where possible. If an estimate represented the average

performance across multiple problems, or in cases where problems included multiple hypotheses or cues, no values were recorded. Studies that included a *visual aid* were coded for the type of visual aid used: (a) 2×2 contingency table, (b) frequency/probability tree, (c) Venn diagram, (d) picture, or (e) experienced-based interactive cards. Given that there were few instances of each type of visual, we compared effects that used any visual aid (a)–(e) coded as (1) versus no visual aid (0) in the analysis.

Three of the included studies (representing seven independent comparisons) made a direct comparison between natural frequency and chances with natural sampling problem formats. The only difference between the two conditions was related to the number of events (chances referred to the probability of a single event, and natural frequencies to multiple events). In all other respects, these chances formats differ almost entirely from other probability formats, specifically in terms of information structure (or lack of normalisation). We decided to include these effects when coding for studies because of recent interest in their similarities and the implications they are argued to have on theoretical perspectives (see, e.g., Brase, 2008; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a). However, we analyse these effects separately from the primary analysis.

Methodological factors. For each effect, we coded sample size and whether the study employed a *within-subjects* design, coded as (1), or not, coded as (0), for use in analyses as well as various study procedural factors. We coded whether participants received at least one problem of *both formats* as (1), and (0) otherwise, and the number of *additional problems* completed per format. We also recorded whether problems were presented in a fixed sequence or random order, however most studies used counterbalanced designs and it was not possible to test any order effects. Further, we recorded whether an incentive was offered and if so, the type of incentive: A *show-up fee* was coded as (1) when a it was given in the form of money or course credit, and (0) if participation was voluntary. Similarly, *performance-pay* was coded as (1) when payment depended on performance, and (0) when participation was voluntary. Where no information about incentives was stated explicitly, no category was recorded.

In relation to *strict scoring* criteria, we categorised studies as apply-

ing either a strict or more lenient criteria for scoring a respondent's solution as correct[5], and coded whether or not a supportive protocol was required. Specifically, we coded studies as requiring an: (a) exact estimate (include rounding), (b) exact estimate plus or minus a percentage margin, (c) exact estimate (including rounding) with supportive protocol, and (d) exact estimate (including rounding) or a supportive protocol[6]. Given that categories (a), (b), and (d) were determined to be more lenient criteria, and category (d) was not common, these were grouped together and *strict scoring* was coded (0) in these cases and (1) in case (c).

Individual differences. The primary sample of participants comprising each effect was coded to account for the education and expertise of the participants: (a) university undergraduate/postgraduate students, (b) medical students, (c) physicians, (d) general population, (e) older adults, (f) children, (g) secondary school students, (h) management executives. Owing to small samples across some of the categories, we considered categories (a), (b), (c), and (h) as instances of samples of (*experts*) in probabilistic reasoning or similar problems as described in the literature and coded these samples as (1) and (0) if they belonged to groups (d), (e), (f), or (g). Where *numeracy* was examined and reported as a moderator, samples were coded as having *high numeracy* (1) or low numeracy (0), typically determined by a median split of the sample in the study[7]. In all cases, the 11-item Lipkus numeracy scale (Lipkus, Samsa, & Rimer, 2001) was used to measure numeracy, although one study (Galesic, Gigerenzer, & Straubinger, 2009) added an additional item about a coin toss from Schwartz, Woloshin, Black, and Welch (1997). There were almost no studies that employed other measures of cognitive ability (see Lesage et al., 2013; Sirota, Juanchich, & Hagmayer, 2014, for exceptions).

[5]Only one study used average errors as a dependent variable (Fiedler et al., 2000), for which we obtained data classifying responses as correct or incorrect.

[6]A supportive protocol could be used to confirm a correct estimate, or to recode an incorrect estimate if the correct process was followed but a calculation error was made. In all but two studies (Vallée-Tourangeau, Abadie, & Vallée-Tourangeau, 2015; Zhu & Gigerenzer, 2006), the protocol was used to confirm a correct estimate.

[7]In one case, we obtained raw data from the authors in order to calculate separate effects for high and low numeracy (Vallée-Tourangeau, Abadie, & Vallée-Tourangeau, 2016).

2.4.4 Calculation of Outcome Measures

The focus of the meta-analysis is on the proportion of correct responses obtained for each of the two formats, that is, the natural frequency and the conditional probability formats. As mentioned above, we only considered studies that compared performance across formats eligible for inclusion. For this reason, each condition of each experiment yielded two proportions when two formats were compared, or three proportions when three formats were compared. We will refer to them as a tuple and assume, for the sake of exposition, that each tuple consists of two proportions, one for each format. Each study can yield either one tuple or several when it includes multiple conditions. Likewise, each of the 35 articles could include one study or several.

For each tuple, we calculated the proportion correct for both formats, using

$$p_F = \frac{c_F}{c_F + i_F} \qquad p_P = \frac{c_P}{c_P + i_P} \tag{2.4}$$

where subscripts F and P denote the natural frequency format and the conditional probability format, respectively, c denotes the number of correct responses, and i denotes the number of incorrect responses. Each observed proportion provides one estimate of the true proportion underlying the observed value. We therefore refer to observed proportions as estimates.

The meta-analytic model that aggregates the different estimates requires that they follow normal distributions. However, because p_F and p_P can only take values in the limited range of $[0, 1]$, they cannot be normally distributed and require transformation before being aggregated. One common transformation is the logit-transformation, defined as

$$l_F = ln\left[\frac{p_F}{1 - p_F}\right] = ln\left[\frac{c_F}{i_F}\right] \qquad l_P = ln\left[\frac{p_P}{1 - p_P}\right] = ln\left[\frac{c_P}{i_P}\right]. \tag{2.5}$$

For each format, the logit is the natural logarithm of the odds of a correct response, which is in turn defined as the ratio of correct to incorrect responses. When correct and incorrect responses are equally frequent, the odds are one and the logit takes a value of zero. When a correct response is more frequent than an incorrect response, the odds exceed one and the logit is positive. Conversely, more incorrect than correct responses imply a negative logit. From now on, we will refer to the logit when we would usually refer to a proportion.

Because each logit is estimated from a finite sample of participants, we record the corresponding sampling variance as a measure of its precision. Following Woolf (1955), these variances were calculated using

$$v_{l_F} = \frac{1}{c_F} + \frac{1}{i_F} \qquad v_{l_P} = \frac{1}{c_P} + \frac{1}{i_P}. \tag{2.6}$$

Most estimates were calculated from independent samples so that we assumed the covariances between them to be zero. However, some estimates were based on the same set of participants, which happened in either of three cases. First, in five studies, results were reported separately for each problem although all problems were solved by the same set of participants. In this case, we averaged proportions correct across problems, yielding only one estimate for each format. Second, in five experiments (yielding 18 logits), the same set of participants was used for both formats, yielding a within-subject estimate of the facilitation effect. Finally, five experiments (yielding 40 logits) shared participants across conditions. In both these cases, we estimated the covariance between the two logits based on the same set of participants using a method suggested by Stedman, Curtin, Elbourne, Kesselheim, and Brookhart (2011),

$$cov(l_a, l_b) = n \times \frac{n \times s - c_a \times c_b}{c_a \times c_b \times i_a \times i_b}, \tag{2.7}$$

where a and b denote two arbitrary formats, n denotes the sample size, and s denotes the number of participants giving a correct response under both formats[8].

We are not only interested in aggregating the proportion of correct responses for each format but also in the relative advantage of one format over the other. This effect size is captured by the odds ratio,

$$\text{OR} = \frac{c_F/i_F}{c_P/i_P}, \tag{2.8}$$

which gives the ratio of the odds for the natural frequency format to the odds for the conditional probability format. The logarithm of the

[8]There were eight logits for which the quantity s was not available. In three of these cases, either $c_F = 0$ or $c_P = 0$, so that $s = 0$. In the remaining five cases, we used the middlemost possible value and examined the effect of this choice on model outcomes (see section *Model Diagnostics*).

odds ratio is easily computed from the logits of both formats,

$$ln[\text{OR}] = ln[c_F/i_F] - ln[c_P/i_P] = l_F - l_P. \tag{2.9}$$

Intuitively, $ln[\text{OR}]$ is positive when the logit (and therefore the proportion correct) under the natural frequency format exceeds the logit under the conditional probability format, and is negative in the reverse case.

After averaging proportions and calculating covariances, we obtained a final set of $k = 226$ logit estimates from 115 conditions in 35 papers, as well as a $k \times k$ covariance matrix of these logits, $\mathbf{V}$. The estimates could then be aggregated to obtain summary estimates and examine the effects of different study characteristics. The following section will present the statistical strategy for conducting this meta-analysis.

2.4.5 Aggregation of Outcome Measures

To aggregate the $k = 226$ logits shown in Table 2.A in Appendix A, we use a bivariate mixed effects model (van Houwelingen, Zwinderman, & Stijnen, 1993). Because bivariate models are relatively new to the meta-analytical toolbox, this section gives a detailed exposition of the model.

According to the bivariate model, each experiment is characterized by two logits that represent the proportions correct for each format. Rather than combining these logits in an effect-size estimate, the model treats them separately and estimates how a given moderator affects responses under each format. Formally, each experiment produces a tuple Y_j,

$$Y_j = \begin{pmatrix} l_{F,j} \\ l_{P,j} \end{pmatrix} = \begin{pmatrix} \theta_{F,j} \\ \theta_{P,j} \end{pmatrix} + \begin{pmatrix} e_{F,j} \\ e_{P,j} \end{pmatrix}, \tag{2.10}$$

which consists of two observed logits, l_F and l_P that serve as estimates of the true logits, θ_F and θ_P. Because estimates are based on random samples of participants, the true logits differ from their observed values by residuals e_F and e_P. The residuals are assumed to reflect only sampling variation and follow a multivariate normal distribution

with covariance matrix

$$\mathbf{V} = \begin{pmatrix} v_{l_1} & cov(l_1, l_2) & \cdots & cov(l_1, l_k) \\ cov(l_2, l_1) & v_{l_2} & \cdots & cov(l_2, l_k) \\ \vdots & \vdots & \ddots & \vdots \\ cov(l_k, l_1) & cov(l_k, l_2) & \cdots & v_{l_k} \end{pmatrix} \tag{2.11}$$

where the (estimated) sampling variances of each logit are given along the main diagonal and the (estimated) covariances between the logits are either assumed to be zero or estimated using Equation (2.7), as discussed above.

The true logits of each estimate j are further assumed to consist of two components, reflecting the effects of moderators and residual heterogeneity. Specifically, we decompose the true logits into the following linear combination,

$$\begin{pmatrix} \theta_{F,j} \\ \theta_{P,j} \end{pmatrix} = \begin{pmatrix} \beta_{F,0} \\ \beta_{P,0} \end{pmatrix} + \begin{pmatrix} \beta_{F,1} & \beta_{F,2} & \cdots & \beta_{F,12} \\ \beta_{P,1} & \beta_{P,2} & \cdots & \beta_{P,12} \end{pmatrix} \times \begin{pmatrix} x_{1,j} \\ \vdots \\ x_{12,j} \end{pmatrix} + \begin{pmatrix} u_{F,j} \\ u_{P,j} \end{pmatrix}, \tag{2.12}$$

which describes each format's true logit as the sum of a base logit, the effects of moderators, and the effect of residual heterogeneity. Consider first the base logits, $\beta_{F,0}$ and $\beta_{P,0}$. For each format, they give the average true logits underlying estimates from a standard study that has none of the 12 characteristics examined in the full-sample meta-analysis.

Consider next the effects of moderators. For each of the twelve study characteristics, $x_{m,j}$ is a binary variable that takes value 1 when estimate j is based on a study with characteristic m, and 0 otherwise. A value of 1 on $x_{j,m}$ adds the products $\beta_{F,m} \times 1$ and $\beta_{P,m} \times 1$ to the base logits, implying that $\beta_{F,m}$ and $\beta_{P,m}$ capture the effects of study characteristic m on the true logits. In contrast, a value of 0 on $x_{j,m}$ leaves the true logit unaffected by study characteristic m.

Finally, consider the effects of residual heterogeneity, $u_{F,j}$ and $u_{P,j}$. They capture, for example, differences in study design that are left unexplained by the set of moderators. Because studies differ in the exact problems they use and because these problems vary in difficulty, such residual heterogeneity is realistic. The mixed-effects model assumes that the effect of residual heterogeneity follows a bivariate normal distribution with covariance matrix

$$\mathbf{T} = \begin{pmatrix} \tau_F^2 & \tau_{FP} \\ \tau_{FP} & \tau_P^2 \end{pmatrix}, \tag{2.13}$$

where τ_F^2 denotes the variance of $u_{F,j}$, τ_P^2 denotes the variance of $u_{P,j}$, and τ_{FP} denotes their covariance. In addition, the model assumes that the sample of studies is a random sample from the population of possible studies. Provided that these assumptions are met, the mixed-effects model can be used to make inferences about the population of hypothetical studies. An alternative approach does not include random effects from residual heterogeneity. Such a fixed-effects approach does not allow inferences beyond the sample of studies included (Hedges & Vevea, 1998) and is therefore not considered for the present analysis.

The goal of the present meta-analysis is to estimate all β and τ coefficients. That is, we want to estimate both base logits, the average effects of all twelve study characteristics, and the (co-)variances of the random effects of residual heterogeneity. To assess the precision of these estimates, we compute cluster-robust standard errors that account for potential associations between results reported in the same paper (Hedges, Tipton, & Johnson, 2010). To estimate the parameters, individual logits are weighted by their precision using the covariance matrices $\mathbf{V}$ and $\mathbf{T}$. This implies that logits based on large, independent samples receive, ceteris paribus, higher weight than those based on small, correlated samples.

2.5 Results

In this section, we examine the pool of collected studies and report results of our meta-analytical model that aggregates the individual study results and delineates the effects of different study characteristics. This model was estimated using the statistical package R, version 3.3.3 (R Core Team, 2014), and the development version of the `metafor` package (Viechtbauer, 2010).

2.5.1 Distribution of Observed Values

Of the $k = 226$ logits collected for analysis, $k_F = 111$ concern the natural frequency format and $k_P = 115$ concern the conditional probability format. The disparity of k_F and k_P is due to the fact that in four cases an alternative probability format (normalised frequencies) was included along with the usual conditional probability format.

Figures 2.4 and 2.5 show forest plots of the proportions correct for the natural frequency and conditional probability formats, respec-

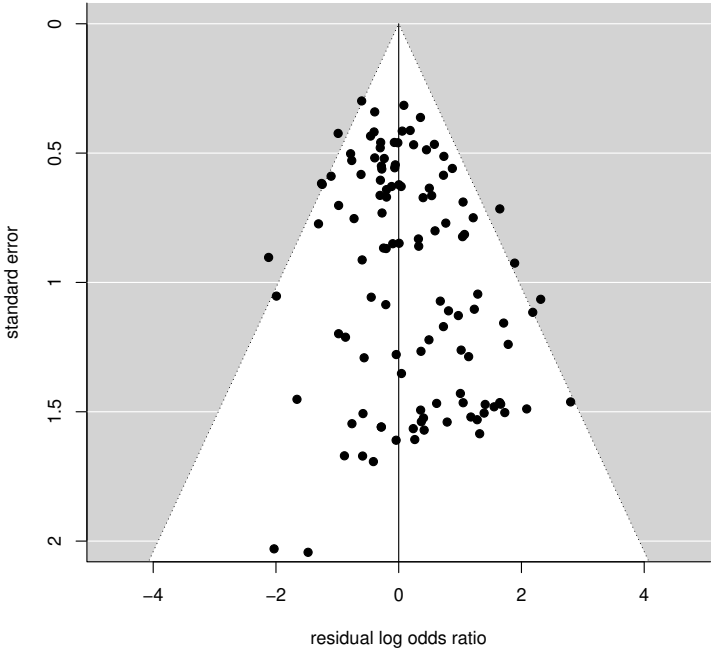

Figure 2.3: *Funnel plot of model residuals: For each observed log odds ratio, the x-axis gives the residuals from a univariate mixed-effects model that includes the same study characteristics as the bivariate model. The y-axis gives the standard errors of the observed log odds ratios, and the triangle defines the 95 percent confidence interval. In the absence of publication bias, the funnel plot is symmetric; asymmetry indicates publication bias. We can observe that the plot appears largely symmetric with slightly more studies having positive residuals than negative ones.*

tively. Here, each estimate is represented by a square, enclosed by its 95% confidence intervals. The distributions of observed proportions differ markedly between formats, with the majority falling above .2 in Figure 2.4 and below .2 in Figure 2.5. Proportions from the conditional probability format therefore appear to be smaller, on average, than those from the natural frequency format.

To examine evidence of selective reporting of studies, Figure 2.3 shows a funnel plot of the observed effects. Selective reporting refers to a biased publication system that favors results of a particular direction of the effect and/or size of the p-value. Because the effect

size is given by the odds ratio, we estimated a univariate mixed-effects model (Hedges & Vevea, 1998) of the estimated ln[OR]. This model contained the same set of moderators as our original model but takes ln[OR] as the outcome variable. Figure 2.3 plots the residuals from this model against each estimate's standard error. One would expect estimates with small standard errors to have residuals closer to zero and those with large standard errors to fluctuate more strongly around the estimated value. The white area reflects this intuition and gives the 95%-confidence intervals.

Funnel plots are commonly used to detect asymmetries that indicate a relative lack of studies with residuals in a particular direction. Although the plot appears largely symmetrical, we can observe that the presence of two very imprecise studies with negative residuals in the lower left portion of the plot creates a relative lack of studies with similar imprecision and positive residuals. Indeed, a formal test of funnel plot asymmetry in meta-analyses of odds ratios (Peters, Sutton, Jones, Abrams, & Rushton, 2006) rejects the null hypothesis of funnel plot symmetry at the one percent significance level ($t = 2.7$, $p = .0077$). Although such funnel-plot asymmetry cannot unambiguously establish selective reporting (see, e.g., Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006), we conclude that there is some evidence of publication bias. Because over-reporting of studies with negative residuals has a negative effect on most of the estimated coefficients, we note that the estimates reported here may underestimate the underlying true effects. However, because the degree of asymmetry is small, we suspect that the under-estimation is limited[9].

2.5.2 Model Diagnostics

The meta-analytical model introduced in the previous section uses the observed proportions to estimate the underlying true proportions for studies with different combinations of study characteristics. Before we discuss the summary estimates, let us consider a few model diagnostics.

[9] Ideally we would re-estimate the model after correcting for publication bias (e.g., Duval & Tweedie, 2000) or employ methods that are unaffected by publication bias (e.g., van Assen, van Aert, & Wicherts, 2015). However, these methods are either not suitable or not available for the bivariate mixed-effects model employed here.

For each observed logit, the polygons in Figures 2.4 and 2.5 give the estimated average proportion for studies with the same characteristics, along with their 95% confidence intervals. However, the estimates do not include the effects of residual heterogeneity, that is heterogeneity in the true effect that is unaccounted for by the study characteristics included in the model. The model estimates $\hat{\tau}^2_F = .54$, $\hat{\tau}^2_P = .50$, and $\hat{\rho}_{FP} = .90$, where ρ_{FP} denotes the correlation between τ^2_F and τ^2_P. The effect of residual heterogeneity is estimated to be, on average, larger in the natural frequency format than in the conditional probability format. At the same time, the high correlation between the random effects implies that studies with large residual heterogeneity in one format have, on average, large residual heterogeneity on both formats. A formal test of residual homogeneity ($Q_{188} = 773.19, p < .0001$) suggests that the presence of residual heterogeneity is not merely a sampling artifact. To put the estimated amount of residual heterogeneity into perspective, consider the ratio of residual heterogeneity to the total variation in observed logits (Higgins & Thompson, 2002; Jackson, White, & Riley, 2012). This ratio is estimated at $I^2_F = 76\%$ for the natural frequency and $I^2_P = 75\%$ for the conditional probability format. Thus, in both formats, more than 70 percent of the variance in effects is due to residual heterogeneity, which would be considered a considerable amount of heterogeneity in medical research (Higgins, Thompson, Deeks, & Altman, 2003). Although psychological studies tend to be less stringently controlled, the amount of heterogeneity can nevertheless be considered large.

Like all model parameters, the amount of residual heterogeneity is estimated with some level of imprecision. We use the upper and lower bounds of the confidence intervals of all three heterogeneity parameters to examine the effect of this imprecision. In total, these parameter estimates can be combined in $2\times2\times2 = 8$ different ways. For each of these eight combinations, we re-estimate the model with τ^2_F, τ^2_P, and τ^2_{FP} fixed a priori. Because the amount of heterogeneity affects the weights assigned to each observed logit, each re-estimation of the model yields a different set of estimated summary effects. However, none of the estimated changes in proportions varies by more than two percentage points, leading us to conclude that the imprecision in the estimated amount of heterogeneity has no considerable effect on the conclusions of our analysis.

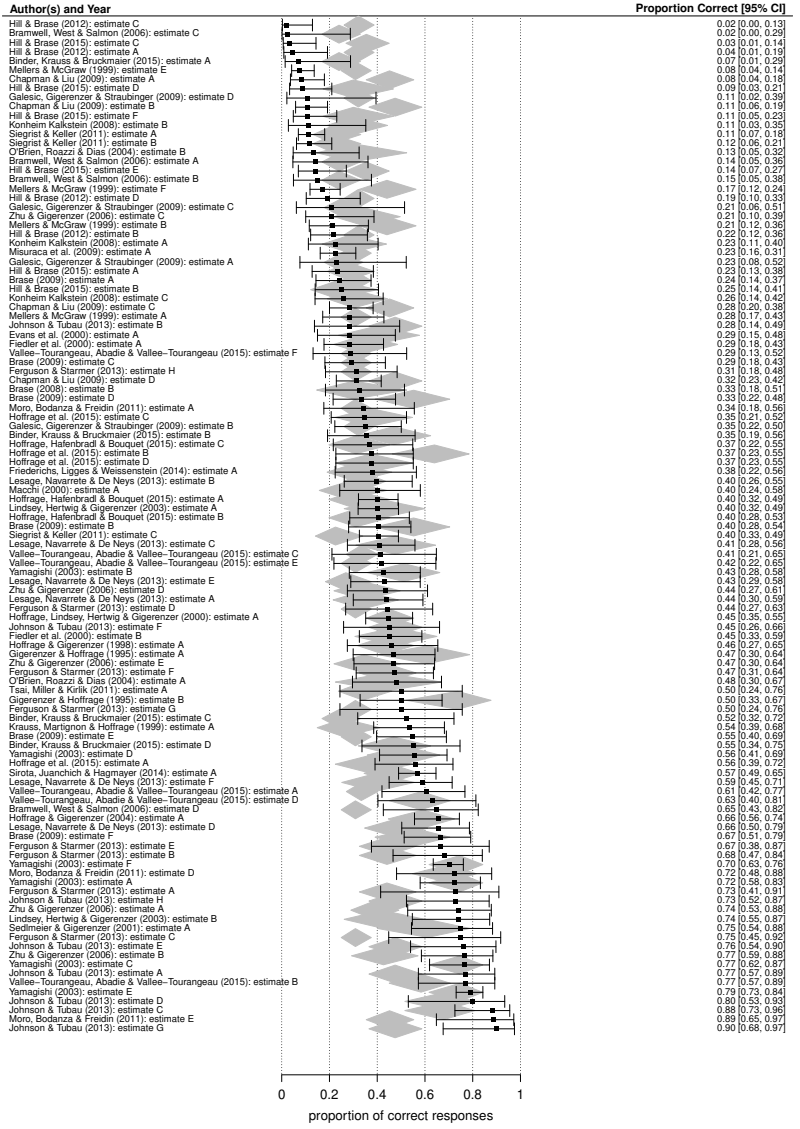

Figure 2.4: *Forest plot of proportions correct in natural frequency format: The black squares represent the observed values and the whiskers their corresponding confidence intervals. The polygons represent the estimated average proportions (and their confidence intervals) for studies with the same characteristics. Estimates A, B, C,... denote different comparisons from the same paper, referenced in Table 3.*

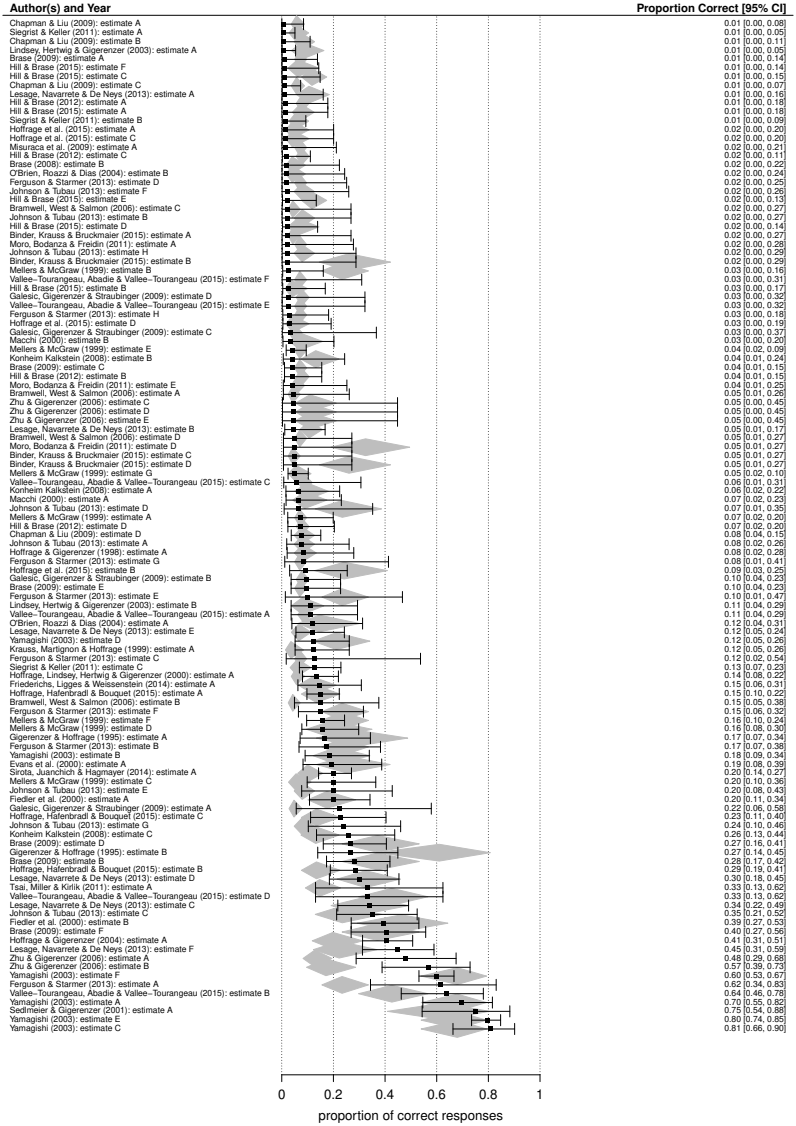

Figure 2.5: *Forest plot of proportions correct in conditional probability format: The black squares represent the observed values and the whiskers their corresponding confidence intervals. The polygons represent the estimated average proportions (and their confidence intervals) for studies with the same characteristics. Estimates A, B, C,... denote different comparisons from the same paper, referenced in Table 3.*

Figures 2.4 and 2.5 show that accuracy appears to be lowest at the very top and bottom of each forest plot, where extreme observed logits are found. To identify such outliers, we computed the standardized residuals, dividing each residual by its standard deviation (Viechtbauer & Cheung, 2010). This metric follows a standard normal distribution where all logits with residuals $|z| > 1.96$ may be defined as outliers, 18 in the present case. Outliers can have an undue influence on the model parameters when their leverage is high, that is when their study characteristics differ strongly from the average. Leverage is expressed by an estimate's hat value and there are 4 logits with hat values indicating high leverage, however none of them is an outlier. Therefore, we would expect that none of the outliers has undue influence on model parameters. Indeed, we computed the Mahalanobis distances of the observed logits, which indicate their individual effects on the estimated summary values. Again, no single observed logit was found to have an undue effect on the model estimates, which may not be surprising given the total number of logits. Therefore, no observed logit was excluded from the meta-analysis.

Another sensitivity check concerns the estimation of the sample covariances in the section *Calculation of Outcome Measures*. As mentioned before, the estimation required knowledge of quantity s, the number of participants with a correct response in both formats. This number was not available for all studies but the range of possible numbers was restricted by zero on the lower end, and c_F and c_P (whichever is lower) on the upper end. From this range, we used the middlemost value for estimating the covariance. To examine the effect of this choice on the outcomes of the meta-analysis, we re-estimated the model for all possible combinations of s values. Given the ranges of the five estimates, there are $2 \times 4 \times 10 \times 2 \times 6 = 960$ different possibilities and model re-estimations. The model parameters changed very little with no estimate of β_F or β_P varying by more than 0.1 and no value of τ_F^2 or τ_P^2 ranging by more than 0.02. We therefore suspect that this factor is unlikely to affect our conclusions.

Last, the model assumes normality in the errors and Figure 2.6 shows a quantile plot of the residuals, where the straight line indicates full normality. Indeed, the observed residuals follow the line closely and do not seem to deviate systematically, indicating an approximate normal distribution.

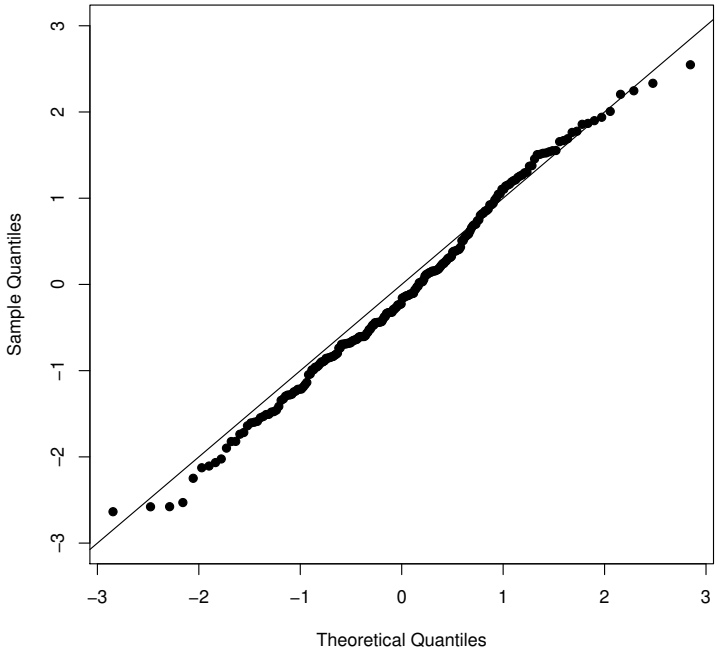

Figure 2.6: *Quantile plot of residuals: When the residuals follow a normal distribution, they should line up along the diagonal line where the theoretical quantiles of a normal distribution equal the observed quantiles. Indeed, the observed quantiles follow approximately a normal distribution and deviations are small and non-systematic.*

2.5.3 Summary Estimates and Study Characteristics

The upper panel of Table 2.2 shows the estimated summary effects. The first row gives the base proportions of both formats. Recall that the base proportions give the proportions correct for a standard study with none of the characteristics examined below. To obtain these proportions from the logits, we re-converted both estimates of β_0, using

$$\Delta_{F,0} = \frac{e^{\beta_{F,0}}}{1 + e^{\beta_{F,0}}} \qquad \Delta_{P,0} = \frac{e^{\beta_{P,0}}}{1 + e^{\beta_{P,0}}} \tag{2.14}$$

which is the inverse of Equation (2.5). For the natural frequency format, the model estimates that on average 24 percent of participants

Table 2.2
Estimated Average Proportions and Implied Odds Ratios

		natural frequency		conditional probability		implied effect size	
No	variable	Δ_m	$CI_{.95}$	Δ_m	$CI_{.95}$	OR	$CI_{.95}$
0	baseline	.24	[.13, .40]	.04	[.01, .14]	7.11	[4.37, 11.56]
1	short menu	+.12	[−.13, +.47]	+.11	[−.00, +.38]	3.08	[1.75, 5.42]
2	three hypotheses	+.19	[−.05, +.45]	+.09	[+.01, +.25]	4.66	[2.32, 9.37]
3	two or more cues	−.04	[−.20, +.37]	−.02	[−.04, +.07]	11.37	[4.59, 28.20]
4	probability question	−.02	[−.17, +.27]	+.02	[−.02, +.13]	4.42	[2.08, 9.37]
5	frequency question	−.01	[−.18, +.38]	−.02	[−.04, +.06]	12.40	[3.90, 39.44]
6	enumerated population	−.00	[−.15, +.24]	−.00	[−.02, +.06]	7.06	[3.40, 14.69]
7	multiple events	+.02	[−.17, +.37]	+.02	[−.02, +.12]	5.53	[3.31, 9.24]
8	visual aid	+.23	[+.02, +.45]	+.22	[+.04, +.53]	2.52	[1.36, 4.67]
9	strict scoring	−.02	[−.16, +.23]	+.01	[−.03, +.15]	4.72	[2.67, 8.33]
10	both formats	+.13	[−.15, +.53]	−.01	[−.03, +.09]	16.19	[8.75, 29.96]
11	additional problems	+.01	[−.03, +.05]	+.00	[−.01, +.02]	6.72	[4.14, 10.91]
12	experts	+.07	[−.11, +.34]	+.03	[−.03, +.24]	5.96	[4.39, 8.09]
0	chances	.40	[.16, .71]	.19	[.03, .67]	2.77	[1.49, 5.16]
0	no incentive	.41	[.27, .57]	.10	[.05, .21]	6.10	[3.46, 10.77]
1	show-up fee	−.03	[−.22, +.20]	+.01	[−.07, +.21]	4.90	[3.60, 6.68]
2	performance pay	+.23	[+.07, +.36]	+.11	[−.00, +.28]	6.62	[2.57, 17.07]
0	low numeracy	.26	[.00, .97]	.04	[.00, .56]	9.37	[4.10, 21.41]
1	high numeracy	+.25	[+.04, +.46]	+.11	[−.01, +.50]	5.92	[3.32, 10.57]

Notes: Top panel gives results of full-sample meta-analysis and bottom panels give results of subset analyses; numbers without sign give baseline proportions and numbers with sign give changes in proportions; confidence intervals based on cluster-robust standard errors (Hedges et al., 2010); numbers are rounded.

give a correct response. In contrast, on average only 4 percent of participants are estimated to give a correct response with the conditional probability format. These proportions imply an odds ratio of 7.1, which can be considered a strong effect. Even at the lower bound of the 95% confidence interval, this corresponds to an effect that is exceptionally large. On average, the share of participants who correctly solved Bayesian inference problems is 20 percentage points higher when they are presented in natural frequencies rather than conditional probabilities.

The remaining rows of the top panel in Table 2.2 give the changes in proportions that can be attributed to different moderators. For each format, we obtained Δ_m associated with study characteristic m by adding β_0 and β_m, converting the resulting logit into a proportion,

and subtracting the base proportion,

$$\begin{aligned}\Delta_{F,m} &= \frac{e^{\beta_{F,0}+\beta_{F,m}}}{1+e^{\beta_{F,0}+\beta_{F,m}}} - \frac{e^{\beta_{F,0}}}{1+e^{\beta_{F,0}}} \\ \Delta_{P,m} &= \frac{e^{\beta_{P,0}+\beta_{P,m}}}{1+e^{\beta_{P,0}+\beta_{P,m}}} - \frac{e^{\beta_{P,0}}}{1+e^{\beta_{P,0}}}.\end{aligned} \tag{2.15}$$

Using the estimates of $\Delta_{F,m}$, $\Delta_{P,m}$ we are now able to examine the effects of each of the twelve study characteristics on the proportions correct and the resulting odds ratios. Note that a study characteristic that affects performance of both formats equally can have strong effects on the odds ratio, because it alters the relative advantage of one format over the other. For example, increasing performance from 50% to 60% in one format and from 10% to 20% in the other reduces the odds ratio from $\frac{.5}{.5}/\frac{.1}{.9} = 9$ to $\frac{.6}{.4}/\frac{.2}{.8} = 6$. Note also that we discuss the different study characteristics individually, although most studies include one or more of them simultaneously. For example, the original study by Gigerenzer and Hoffrage (1995) used *strict scoring*, tested *experts*, and had participants complete 14 *additional problems* in *both formats*. To obtain the proportions correct estimated by our model for this scenario, one needs to add the effects of all these moderators to the base proportions.

As shown in Table 2.2, *visual aids*, *three hypotheses*, and *short menu* were the strongest problem representation characteristics across formats. Methodological factors were also influential, in particular *both formats*. In many cases, moderators that improved performance for probability formats also improved performance for natural frequency formats. However, performance rates varied markedly across studies, and it is sobering to acknowledge that even with the inclusion of influential moderators, many participants were still unable to solve Bayesian inference problems in natural frequency formats. Below, we discuss the results and their practical and theoretical implications following our moderator analysis.

Problem representation moderators. Problem representation moderators affected natural frequency and conditional probability formats in a similar direction although to varying degrees and often with greater percentage increases for natural frequency formats. However, owing to the relative improvement or decline found for both formats, in some cases the size of the facilitation effect was reduced. A subset

analysis on studies that compared natural frequencies to chances with natural sampling suggests that, despite a similar information structure, natural frequencies retained a facilitative effect.

Short menu As predicted by Gigerenzer and Hoffrage (1995), short menu formats have a positive effect on probability formats but perhaps unexpected was the improvement in performance for the short menu versions of natural frequencies. A short menu is estimated to increase the share of correct responses by 12 percentage points in the natural frequency condition and by 11 percentage points in the conditional probability format. Although both proportions increase by similar amounts, the parallel increase would lower the average odds ratio to 3.1 because the absolute difference in proportions now constitutes a smaller relative advantage. Thus, presenting short menu formats improves performance as predicted, but there is some added facilitative effect for short menu formats presented in joint frequencies.

As mentioned in the introduction, there were differences in the conjunctive structure of information presented in the short versions used across studies: three studies provided $p(D)$ and $p(D \cap H)$ (Ferguson & Starmer, 2013; Gigerenzer, 1996b; O'Brien, Roazzi, & da Graca B. B. Dias, 2004) and three studies provided $p(D \cap H)$ and $p(D \cap \neg H)$ (Fiedler et al., 2000; Lesage et al., 2013; Mellers & McGraw, 1999). Consistent with Gigerenzer and Hoffrage's (1995) argument, in both cases the equation for short and standard natural frequency versions remains the same (albeit in some cases with the denominator already calculated) whereas for conditional probability versions the equation is substantially simplified owing to the fact that the base rate is no longer needed in the calculation. Rather, participants must determine which of the joint events are relevant to the solution; this is most obvious in the case where only $p(D)$ and $p(D \cap H)$ are provided but may also be facilitated when all joint events are provided (Ottley et al., 2016; Wu, Meder, Filimon, & Nelson, 2017).

From a nested-sets theory perspective, Lesage et al. (2013) and Mellers and McGraw (1999) argue that formats presenting joint events facilitate performance because they help participants to visualise the nested or subset structure, although Mellers and McGraw (1999) argued that this would be the case for common but not rare events. It is not entirely clear how the results of the meta-analysis support or

refute these claims, as the mechanism by which the subset structure is revealed has not been specified. Nor is it clear how the joint event formats help participants to visualise the nested structure. Similarly, the theory does not extend this explanation to further describe the improved performance of short natural frequency formats. Although the ecological rationality framework makes specific predictions for the facilitation of performance for short probability formats based on a computational analysis, the additional facilitation effect for natural frequency formats is also not immediately clear from a computational perspective. One potential explanation is that the base rate is not explicitly mentioned in the joint event or short menu versions which may reduce errors associated with selecting an incorrect denominator. To test this explanation would require primary data from these studies.

Three hypotheses The improvement in performance across both conditional probability and natural frequency formats on problems involving three hypotheses was unexpected. While the natural frequency facilitation effect was anticipated to remain, the added complexity of the additional hypothesis was expected to decrease performance across formats. Introducing a third hypothesis increases performance in both formats, albeit to different extents. Performance for the natural frequency format is increased by 19 percentage points, and more than the 9 percentage point increase for the conditional probability format. Again, increased performance in both formats lowers the estimated odds ratio to 4.7, implying a reduced advantage of the natural frequency format over the probability format, but the effect remains considerable nonetheless. Looking at the three-hypotheses problems included in the meta-analysis gives some indication as to why this effect may have been found. Three papers examined three-hypotheses problems: Yamagishi (2003, contributing 12 logits) tested numerical variations of the gemstone problem, Johnson and Tubau (2013, contributing four logits) included a gemstone problem and an apple distribution problem based on the probabilities used in the gemstone problems in Yamagishi (2003), and Hoffrage, Krauss, et al. (2015, contributing two logits) used a medical test that could identity the presence of one of two diseases (or neither disease). In all cases except for the Hoffrage, Krauss, et al. (2015) medical diagnosis problem, the problems segment information in a way that requires

that only two hypotheses need to be considered for the solution. To illustrate, the gemstone problem segments information into three hypotheses (blurred, cracked, and clear gemstones) and the goal is to infer how many of the stones that pass inspection are clear. The hit rates and true negative rates are 100% for two of the hypotheses (a machine retains all clear stones and rejects all cracked stones), reducing it to a two hypotheses format with a perfect hit rate. For this reason Johnson and Tubau (2013) described the problems as simple because the format presents a simpler structure. This may explain the absence of a negative effect based on computational complexity predictions.

We can also further speculate as to why a positive effect was found for three-hypotheses problems. The numerical formats used in these problems, particularly for the conditional probability formats, differed from many of the other formats. In six of the three-hypotheses problems (representing 12 logits), common fractions were used to represent proportions (these were the only problems that ever used proper fraction representations), namely $1/2$, $1/3$, or $1/4$. Calculations involving fractions are typically difficult mathematical operations to learn (Siegler, Thompson, & Schneider, 2011), however, dividing quantities into common proportions or parts of a whole (e.g., half, third, quarter) is taught relatively early in formal schooling and fractions are frequently used to divide resources (e.g., sharing food). It is possible that the use of the part-to-whole representations afforded by the fractions and wording used within these problems facilitated the segmentation of information into proportions, similar to the effect of presenting joint frequencies (e.g., imagine cutting a cake in thirds, keeping a third, and cutting one of the thirds in half and comparing the two proportions). This interpretation is supported when considering the high performance on the conditional probability formats ($70 - 81$ percent correct responses) when the gemstone problems were also accompanied by the roulette wheel visual aid in the studies by Yamagishi (2003). The roulette wheel segments the information in such a way that its segments line up to create a clear visual segmentation, which may have facilitated performance in these conditions (Brase, 2014). In this connection, Yamagishi (2003) argued that such a visual aid could tap into people's visual computational abilities.

Two or more cues Adding additional dichotomous cues or cue values (two or more cues) to the standard Bayesian inference task reduces performance for both conditional probability and natural frequency formats (by 4 and 2 percentage points, respectively), although the general facilitation effect of natural frequencies remains. Using two or more cues instead of one was estimated to have a small, negative effect on responses in the natural frequency format and a negligible effect on responses in the conditional probability format. Jointly, these two effects increased the implied odds ratio to 11.4. However, both effects are imprecisely estimated from the small set of studies that are currently available, and the results may change, in both sign and magnitude, with more diverse examinations.

Multiple events In the present meta-analysis, studies that employed conditional probability formats with multiple-event phrasing had a small improvement on performance rates by 2 percentage points compared to problems with single-event phrasing[10]. Curiously, the framing of the conditional probability format appears to have affected performance in the natural frequency format, although these effects are imprecisely indicated and may reflect sampling variation. Taken together, the two effects decrease the estimated average odds ratio to 5.5. One explanation for these effects is that the types of problems that were used in these studies were generally easier problems. However, as we cannot examine the problems in greater detail, this is only speculative.

Incongruent question formats and enumerated population Incongruent question formats (*probability question, frequency question*) appear to offer only small disadvantages to the respective formats (each decreases 2 percentage points). As these estimates are based on a small number of studies that manipulated these characteristics, the effects are estimated imprecisely and may require further examination. However, based on the evidence currently available, the effects appear to be small and negligible. Similarly, augmenting the probability formats with an *enumerated population* does not appear to facilitate performance (no percentage change). This effect is imprecise

[10]To ensure the effect was not driven by the few studies that examined normalised frequency formats presented using numerical frequencies (e.g., 10 in 1000) we repeated the analysis with these studies excluded. The effect did not change.

for natural frequency formats but is estimated with greater precision for the probability formats. Thus, small changes to the complexity of the problem or introducing a requirement to convert problem and question formats appear to have a negligible effect on performance, particularly in contrast to the strong effects of information structure.

Visual aids The strongest moderator of performance for both natural frequency and probability formats involved the inclusion of a visual aid. Supplementing both formats with a visual aid increased performance by 23 and 22 percentage points for the natural frequency and probability formats, respectively. The strong improvements in performance for both formats decreased the odds ratio to 2.5, although this still indicates a strong natural frequency facilitation effect. When visual aids are used for conditional probability formats they enhance performance to a similar level as natural frequency formats without visual aids. However, given that visual aids improve performance on natural frequency formats to the same degree, it appears that they may have an independent effect to that of format. Visual tools have been used throughout history to convey meaning and represent relations between concrete and abstract concepts, are beneficial for understanding concepts in mathematics and problem-solving more broadly, and tend to be spontaneously produced when individuals attempt to solve probability problems (B. Tversky, 2001; Zahner & Corter, 2010). In this connection, visual aids have been shown to improve comprehension of health risks in risk communications (Galesic, Garcia-Retamero, & Gigerenzer, 2009; Garcia-Retamero & Cokely, 2013).

There have been several recent studies that have explored the features of visual designs related to performance on natural frequency formats. Micallef et al. (2012) explored different Euler diagrams and frequency grids that varied whether or not the area of the different regions was proportional to the quantities represented in the problem. They tested six different variations and found that none of the variations were superior to one another, and only slightly improved performance rates compared to a natural frequency text. However, performance rates across formats in this study were generally poor. Khan, Breslav, Glueck, and Hornbaek (2015) investigated the different qualities of visualisations that could facilitate performance on Bayesian inference problems, again only with respect to

natural frequency formats. Khan et al. (2015) distinguished between visualisations that emphasised branching or information structure (e.g., trees), nested-set relations (e.g., Euler diagrams), and frequencies or the quantities involved (e.g., icon arrays), and developed two hybrid diagrams that combined branching with frequency (a Sankey diagram) or branching that illustrated all sets (a double-tree). The frequency grid and double-tree diagrams resulted in the best performance (20 percent of participants answered correctly in both conditions), however performance was similar across conditions. Unfortunately, owing to the small samples of studies exploring visual aids, we were not able to explore the specific features of the visual aids that were more or less likely to facilitate performance in the present meta-analysis.

Subset analysis: Chances with natural sampling Studies investigating the facilitative effect of chances with natural sampling also suggest that the interpretation of chances as frequencies rather than as a single event improves solution rates (Brase, 2008, 2014). In the present study, natural frequency representations demonstrate superior performance to chances with natural sampling. The second panel of Table 2.2 shows the results of a subset analysis of $k_{chs} = 18$ logits from studies examining chances. For this subset analysis, we used the same model as for the full-sample meta-analysis but excluded any study characteristics. The lack of study characteristics as covariates means that the model does not account for large parts of the systematic variation in observed logits. As a consequence, the estimated amounts of residual heterogeneity are higher than in the original model ($\tau^2_F = .6, \tau^2_P = .7$) and the null hypothesis of residual homogeneity is rejected ($Q = 78.1, p < .0001$). The average performances are estimated at 40 percent and 19 percent, for the natural frequency and conditional probability format, respectively. The implied odds ratio is estimated at 2.8. This analysis is based on few observed logits and does not control for other study characteristics, so that the estimated proportions should not be compared with the results of the full-sample meta-analysis. Nonetheless, the results suggests that although the information structure is the same, single-event representations may still be a more difficult representation for participants to solve.

One additional aspect to consider when interpreting this analysis

is the difference in question format used within these studies. Girotto and Gonzalez (2001) argued that the question format for natural frequency versions prompts participants to compute two terms of the ratio (e.g., those who test positive, and those that test positive and have breast cancer) whereas probability versions do not, thus making the solution easier for natural frequency versions (in addition to information structure). For natural frequency and chances with natural sampling formats, two step questions that first require participants to provide the denominator of the ratio (e.g., people who test positive) followed by the numerator (e.g., people have breast cancer and test positive) were better than the standard question format (i.e., _____ out of _____) for both natural frequency and chances with natural sampling formats. In all of the studies comparing chances with natural sampling and natural frequency formats the two part question was used and may explain the higher performance for both formats in these studies (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015b). Only Brase (2008) used a two part question form when comparing normalised chances and natural frequency formats, limiting our ability to test the moderator in the present meta-analysis.

Methodological factors. Methodological factors were found to account for differences in performance for the different formats, suggesting that some of the variation in the natural frequency facilitation effect reported across studies may result, in part, from differences in study design. As we will see, some of the methodological moderators do not affect both formats in the same way, thus distorting the facilitation effect relative to a typical experimental setup. The results can therefore guide methodological decisions in future empirical work.

Both formats Study designs that involved participants receiving one or more conditional probability problems *and* one or more natural frequency problems appeared to help performance for natural frequency but hinder performance for conditional probability formats. Exposing participants to both formats (e.g., in a within-subject design) is estimated to increase performance for natural frequency formats by 13 percentage points on average, and slightly decrease performance for conditional probability formats; this combination increases the odds ratio to 16.2, which is substantial. One potential explanation for

these effects is that solving both types of formats may interfere with the strategies applied to subsequent problems. In some cases, this may have an advantage whereas in other cases it may not. As the majority of studies that used multiple format designs also counterbalanced the order of problems, we are unable to test any order effects, and no order effects were reported in these studies.

Additional problems Performance on natural frequency and conditional probability formats improved with the number of problems that participants were required to solve. The experimental set-up expressed in the baseline proportions assumed that each participant was given only one problem. Assuming linearity, the model estimates that each additional problem increases performance, on average, by around 1 percentage point for natural frequency format and .5 percentage points for the conditional probability format. Because some studies have their participants solve ten Bayesian problems, implying an increase in performance by 10 and 5 percentage points, respectively, the estimated effect due to practice can be considerable. In such cases, the odds ratio would decrease to 5.3. It may be that further opportunities to solve problems may increase the potential for one to try out different solution strategies that turn out to be successful for at least one of the problems or that the underlying structure of problems is more likely to be recognised. However, our hypotheses about the reason for this effect is purely speculative; rather we stress that methodological differences can account for some of the variation across studies.

Strict scoring criteria Contrary to the results of McNair and Feeney (2014), there was no effect of a stricter scoring criteria on performance rates, as defined in the current study. McNair and Feeney (2014) compared scoring that involved either an exact estimate or an exact estimate $+/-5$ percentage points for conditional probability problems. In the reviewed studies, the $+/-\%$ range varied across studies from $1-5$ percentage points (Bramwell et al., 2006; Chapman & Liu, 2009) making it difficult to determine an appropriate cut-off value to indicate greater or lesser leniency. Nevertheless, in the present study we sought to compare lenient against the strictest scoring criteria which we determined to be one that required not only a correct estimate but also a correct protocol to support the

solution. The justification for these scoring criteria, as stated by the authors of the reviewed studies, was to ensure that a correct estimate was *not* a result of a guess or an alternative, non-Bayesian strategy. However, the strictness of the scoring criteria had a negligible effect on performance in both formats, with a decrease of 2 and a slight increase of 1 percentage points for natural frequency and probability formats, respectively. These estimates, particularly for the natural frequency format, are imprecise and may change when additional studies that employ strict coding criteria are included.

Subset analysis: Incentives Of the $k = 226$ logits in the full sample, data on incentives is available for only $k_{inc} = 165$ logits, of which 115 use show-up fees, 12 use performance pay, and 38 use neither. For this reason, we could not include incentives in the full-sample meta-analysis but examined them in a separate analysis that does not include other study characteristics as covariates. The model used for aggregating the logits is the same as before with incentive included as the only covariate. Again, the amounts of residual heterogeneity are higher than in the original model ($\tau_F^2 = 1.0, \tau_P^2 = 1.9$) and the null hypothesis of residual homogeneity is rejected ($Q = 1373.5, p < .0001$). The third panel of Table 2.2 shows the effects of incentives. The proportions for experiments without incentives are estimated at 41 percent for the natural frequency format and 10 percent for the conditional probability format. Unlike a show-up fee, which appears to have a negligible effect, performance pay is estimated to increase performance by 23 and 11 percentage points, respectively, again without controlling for differences in study designs, which complicates comparisons with the full-sample meta-analysis.

There were only two studies that systematically examined the effect of performance-based incentives (Brase, 2009a; Ferguson & Starmer, 2013, representing a total of 12 logits), with many studies incentivising participants with the award of course-credit or payment of a show-up fee. As stated previously, the two studies that investigated the effect of performance-based incentives on performance across formats found contradictory effects: Brase (2009a) found incentives facilitated performance on natural frequency problems whereas Ferguson and Starmer (2013) found a general effect of incentives irrespective of format. Unfortunately, owing to the limited studies on performance-based incentives we cannot resolve this conflict.

Rather, we can conclude that, similar to studies in other domains, performance-based incentives may generally improve performance across formats (Cerasoli et al., 2014).

Individual characteristics. Owing to only a few studies examining numeracy, we could only include experts as a moderator in the full-sample meta-analysis. However, we conducted a subset analysis on studies that included effects for low and high numerate participants.

Experts Natural frequencies benefit experts and non-experts alike, and the results of the meta-analysis do not suggest that participants with greater educational or professional experience are better able to solve problems presented in either format, as compared to lay samples. The slight increases in performance in both formats are imprecisely estimated and may change direction as further studies comparing expert and novice samples accumulate. However, we acknowledge that a limitation to our analysis is the reduction of a broad range of samples to a dichotomy of samples that were presumed to be experts and those that were presumed to be non-experts, although, the distinctions were consistent with arguments made within the literatures reviewed. Medical professionals, management executives and university or postgraduate students are often described as comprising more expert, educated samples. While we found that a broad range of samples were used across studies, the majority were based on university students, consistent with the general criticism of research in the behavioural sciences that studies tend to be based almost entirely on university student samples (Henrich, Heine, & Norenzayan, 2010). Expanding research to include more diverse samples could provide valuable insights into how Bayesian reasoning is learned and evolves over time. In particular, we think that a greater focus on Bayesian reasoning in children is warranted in order to explore developmental trajectories, for example, to elucidate which mathematical concepts are influential in revealing solution strategies (see, e.g., Zhu & Gigerenzer, 2006) or to learn where difficulties first emerge and how they can be targeted. In this connection, further work on older adults, who may suffer from cognitive declines with age, would be important from a more applied perspective, particularly given that older adults are increasingly faced with the implications of medical test results.

Subset Analysis: Numeracy Of the $k = 226$ logits in the full sample, data on numeracy is available for only $k_{num} = 52$ logits. For this reason, we could not include numeracy in the full-sample meta-analysis but examined it in a separate analysis that does not include other study characteristics as covariates. The model used for aggregating the logits is the same as described earlier with only numeracy included as a covariate. The lack of other study characteristics as covariates means that the model does not account for large parts of heterogeneity. As a consequence, the amounts of residual heterogeneity are higher than in the original model ($\tau_F^2 = 1.5, \tau_p^2 = 1.0$) and the null hypothesis of no residual heterogeneity is rejected ($Q = 267.5, p < .0001$). The bottom panel of Table 2.2 shows the effects of numeracy. Ignoring differences due to other study characteristics, for participants with low numeracy scores it is estimated that 26 percent achieve the correct solution for the natural frequency format and 4 percent for the conditional probability format, resembling the baseline proportions in the full-sample meta-analysis. The performance of participants with high numeracy scores exceeds those of participants with low scores by 25 percentage points in the natural frequency format and 11 percentage points in the conditional probability format. These effects appear large but ignore differences in study design and cannot be directly compared to effect estimates from the full-sample meta-analysis[11]. Additional studies on the effect of numeracy are required for other covariates be to included in such an analysis, and for reducing the current imprecision of the estimated effect of numeracy.

Nevertheless, our analysis shows that natural frequency formats continue to offer an advantage over conditional probability formats for low numerates and this difference is maintained for high numerates. These results support prior work showing a general format effect for natural frequencies (Hill & Brase, 2012; Johnson & Tubau, 2013) and do not suggest that the effect is stronger for high numerates

[11] There is, however, reason to assume that any effect estimated in this meta-analysis underestimates the effect of numeracy because all studies included here are based on median-splits in which group assignment to numeracy groups is based on the sample median score. When the distribution of test scores in the sample does not reflect the population distribution, this procedure may be biased because (some of) those below the sample median may fall above the population median. More accurate results can be obtained using "hard" cutoffs as are common in clinical research where assignment is based on pre-defined values. Such studies often require larger samples of participants.

(Chapman & Liu, 2009). Johnson and Tubau (2013) suggested that numeracy and information format can interact with the complexity of problems, for example in terms of information segmentation (e.g., in the gemstone problem described above) and verbal complexity, to impede performance for low numerates who may be more affected by such manipulations. Unfortunately, we were unable to examine other moderators used in these studies to draw any conclusions about interactions with verbal or computational complexity manipulations (Johnson & Tubau, 2013). There were also few studies that examined claims involving other measures of cognitive ability. In this view, Johnson and Tubau (2015) introduce a broader mathematical framework that takes into account individual differences in other areas, such as text comprehension and problem-solving abilities, to explore insights into solution strategies for Bayesian inference problems. We discuss this framework further below.

2.6 General Discussion

The results of the meta-analysis demonstrate that the natural frequency facilitation effect is fairly robust and is largely retained with respect to a range of individual, methodological, and problem representation moderators. Visual aids and short menu formats were the most influential moderators of the effect, enhancing performance for both natural frequency and conditional probability formats. These results suggest that not only is information structure an important component of the effect but that helping participants to *visualise* the information structure of the problem can improve performance. The computational complexity of the problems themselves also affects performance: adding additional cues or cue values does reduce performance rates across problems as anticipated, but adding an additional hypothesis can increase performance if the features of the problem allow one to ignore one or more of the hypotheses. The influence of methodological factors on performance rates, such as small improvements when individuals complete multiple problems or both formats, suggests that further insights could be gained from examining training or transfer effects across problems.

2.6.1 Limitations of the Meta-Analysis

Like all meta-analyses, the present meta-analysis is only as good as is the data available and most limitations of primary studies apply to their meta-analysis. For example, we were unable to analyse every moderator that has been tested across studies, as there were too few performance estimates for many potential moderators. For this reason, the meta-analysis does not do justice to the more subtle differences in study designs, although we attempted to control for the most important differences. Inevitably, there are more differences than the ones accounted for here and even among those, gaps in the available evidence lead to imprecise estimates.

Furthermore, some of the moderators included in the meta-analysis likely remain understudied. Given the variety of samples, problem representation manipulations, and methodologies utilised across studies, we had to collapse some of the coding categories as there was too little data to allow us to explore all potential coding categories in our analyses (e.g., adults versus children). In each case, we have attempted to justify collapsing coding categories given theoretical or methodological arguments found within the literature. We have also made sure to note caveats to any results involving the affected coding categories. Nonetheless, it remains possible that as a result subtle differences between these categories were overlooked.

2.6.2 Theory Building: Bridges or Fences?

The literature on the natural frequency facilitation effect has been enriched but also hindered by theoretical debates that foster the current dichotomy: ecological rationality framework versus nested-sets theory. In a recent theoretical review of Bayesian reasoning with natural frequencies, Brase and Hill (2015) criticised the persistence of the two "camps" into which many researchers place themselves, arguing that progress depends on researchers engaging in integration rather than competition. Similarly, Johnson and Tubau (2015) have pushed for theory integration and proposed a framework for understanding Bayesian word problems that connects problem solving with mathematical cognition. In the following section, we discuss the implications of the results for the respective theories and make recommendations as to where further work can help to strengthen or clarify the premises of the theories or promote theory integration.

***Ecological* not evolutionary rationality framework.** The computational analysis of Bayesian inference problems originally put forward by Gigerenzer and Hoffrage (1995) is supported by the results of the meta-analysis in that short menu probability formats offered an advantage over standard conditional probability formats. Further, the improvement on short menu natural frequency formats suggests that simplifying even elementary arithmetic operations or specifying only the joint frequencies can facilitate the selection of relevant information. The results also suggest a role for visual displays in facilitating cognitive operations, a finding that connects to a vast literature on the development and use of visualisations for supporting thought (B. Tversky, 2001, 2011; Zahner & Corter, 2010). Although some proponents of the ecological rationality framework argue that specific types of visualisations may be most beneficial for Bayesian reasoning problems from an evolutionary perspective (see Brase, 2009b), others have not made specific predictions. We anticipate that proponents would argue that the more relevant question would be to ask which visual tools are most effective for different cognitive problems (see Zahner & Corter, 2010, for examples of different types of visual aids that are generated for different types of probability problems). Literature on the emergence of visual aids from a cultural or educational perspective may provide insights to guide research here.

Some of the criticism aimed at the ecological rationality framework as it has been applied to Bayesian reasoning is that the framework has not adequately addressed the question of how information about the occurrence of joint events is acquired or accumulated, nor has it offered a clear explanation for why some participants continue to have difficulty with natural frequency formats. This criticism could also be aimed at nested-sets theory. We turn to work on the *description-experience gap* (Hertwig, Barron, Weber, & Erev, 2004; Hertwig & Erev, 2009) to examine this criticism. Research on the description-experience gap emerged in consideration of contradictory findings related to how participants made decisions on the basis of probability distributions (Hertwig & Erev, 2009). When participants were given the opportunity to sample event occurrences, from which probabilities and payoffs can be inferred, common errors found in studies that simply provide probabilities diminish or disappear. For example, base rates are more likely to be used when experienced rather than simply described (Koehler, 1996). Like others (Hoffrage, Krauss, et al., 2015;

Schulze & Hertwig, 2016), we wonder whether natural frequencies offer an intermediate solution by improving on conditional probability formats such that the information is presented in a format that more closely resembles how the information is naturally acquired, but that stops short of providing people with the understanding of information structure that coincides with experience. For example, in line with related work on experience-based probability learning tasks, experience may facilitate the construction of a causal model (see, e.g., Sobel et al., 2004). We return to this aspect below when discussing future directions.

One of the most fundamental premises of the ecological rationality framework is the question of how cognitive processes map onto structures, or the *ecological* aspect of the framework (Gigerenzer & Hoffrage, 2007), yet this has received the least attention in research on Bayesian reasoning (see Gigerenzer & Hoffrage, 1995; Hafenbrädl & Hoffrage, 2015). Rather, much of the criticism of the theory resides in objections to an evolutionary argument regarding the potential for the human mind to have evolved a frequency-processing mechanism (Navarrete & Santamaria, 2011). This argument is not wholly supported by proponents of the theory, but nevertheless has become much more central to the debate than the ecological argument on which it is was originally based. At this point, we think it is pertinent to emphasise the response of Gigerenzer and Hoffrage (2007) to the many different interpretations of the ecological rationality framework put forward by Barbey and Sloman (2007):

> The evolutionary perspective ... provides a general framework for finding the right questions... An ecological framework postulates that thought does not simply emerge inside the mind. Every theory of reasoning needs to specify both cognitive strategies and the environmental structures under which these strategies work well (p.266).

We wonder whether some of the debate between the ecological rationality framework and nested-sets theory would dissipate should the strict modularity view lose its emphasis.

Clarification of nested-sets theory. Given that nested-sets theory and the ecological rationality framework make similar arguments on the importance of the nested information structure to the facilitation

effect of natural frequencies, the two theories can also draw on similar results from the meta-analysis to support their premises. For example, the facilitative effect of short menu formats supports the premise of nested-sets theory that clarifying the nested-set structure of the problem can improve performance. Although the theory justifies the benefits of short menu formats on the basis of their ability to help participants visualise the subset structure, the mechanism by which this is revealed is not entirely clear. Mandel (2007) proposes that representations that reveal nested-sets and minimise computational complexity will enhance performance (holding transparency constant), which he refers to as the *complexity principle* of nested-sets theory. It is not clear how this principle differs from the computational argument made by Gigerenzer and Hoffrage (1995). Further, it is not clear how the theory accounts for the reduction in the natural frequency facilitation effect given short menu formats; the set structure is clarified in the same way in short menu versions of both formats.

Similarly, the prediction that visual aids can improve performance on conditional probability problems is supported in the present meta-analysis. Visual aids could enhance performance to a level similar to natural frequencies without visual aids. Given that visual aids also improved performance for natural frequency formats, we wonder whether the theory would argue that visual aids and natural frequency formats have an additive effect or whether the different methods for revealing subset structures build on one another. Mandel (2007) suggests that nested-sets relations can be clarified through different modalities (which he called the *multi-modal principle*), however, to our knowledge relations between these modalities has not been clarified. For instance, in relation to visual aids, the theory falls short of discussing whether the relative size or structure of the visual aid is important for clarifying the subsets (e.g, see Moro et al., 2011), or whether any visual design that illustrates subsets is sufficient.

One argument mounted in favour of nested-sets theory is that it has broader applications, such that transparent nested-set manipulations can facilitate *deductive* reasoning as well (Amitani, 2015; Barbey & Sloman, 2007). For example, Euler circles showing the subset structure of syllogisms can facilitate solutions (Sloman et al., 2003). Mandel (2007) alludes to the fact that there can be multiple ways in which the clarity of a representation can be improved, however, details of the range of strategies have not been elaborated. In fact, some proponents

of the theory emphasise that *any* manipulation that draws one's attention to the nesting of events will facilitate reasoning (Lesage et al., 2013). Unfortunately, this claim allows the theory to accommodate a broad range of results in support of its premises without providing explanations for the mechanisms (whether these are the same or different across representations). For example, in what ways can nested-set relations be made transparent? In which modalities? Is there a hierarchy of manipulations and how do different modalities operate in connection with one another? The theory needs to imply conditions for testing these mechanisms. On the other hand, the ecological rationality framework has also been criticised for not being able to explain how variations in the transparency of nested-sets, for example, textual manipulations, affect performance (Mandel, 2007).

At this point, we are inclined to agree with Mandel (2007) in his description of nested-sets theory as an assemblage of empirical findings collected in rebuttal to the frequentist mind perspective. Barbey and Sloman (2007) attempted to align the theory with dual-process models of reasoning to suggest that the deliberate rule-based system induces the use of rules about elementary set operations when set relations are transparent, although this dual process view is not wholly supported (Evans & Elqayam, 2007; Lagnado & Shanks, 2007; Mandel, 2007; Samuels, 2007). Sloman et al. (2003) allude to the fact that the ability to reason in relation to sets and subsets, including their relations and their relative sizes, is necessary for many problems (including those under primitive conditions, such as sharing resources) yet argues that claims of adaptiveness do not provide any further explanatory power. Yet, how did we come to operate so well on set relations? Nested-sets theory would benefit from further explication of its key propositions and principles, and from addressing the ecological nature of its predictions.

In summary, one of the major points of difference between the ecological rationality framework and nested-sets theory is the emphasis that evolutionary theory has in the foundations of the respective theories. Yet, it tends to be the proponents of nested-sets theory who continue to point to strong evolutionary claims within the ecological rationality framework, more so than its proponents. Rather, we think that a focus on the *ecological* aspect of the ecological rationality framework generates more interesting research questions. In any case, the two theoretical perspectives are broader than their application to Bayesian reasoning problems, which offer an environment for

testing the predictions of the respective theories. However, if these predictions are ill-specified, this limits the insights that can be gained from this research.

2.6.3 Future Directions

The research reviewed in the present meta-analysis represents just a small part of research on Bayesian reasoning. While elementary Bayesian textbook problems have offered an experimental paradigm for testing theories, they represent a small world and this no doubt limits the breadth of questions that can be addressed about Bayesian inference. Indeed, research within this paradigm led to the important insight that information representation matters to Bayesian reasoning and allowed researchers to explore the importance of information structure relative to other factors (e.g., numeracy, visual aids). However, we can still not offer a coherent explanation for why, in some cases, the majority of participants have difficulties with Bayesian reasoning problems as they have been studied here. Schulze and Hertwig (2016) suggest that differences in research methodology can help to explain some of these findings, for example, the finding that children are good intuitive statisticians but adults are not (e.g., employing experienced-based versus descriptive methods, respectively). We discuss ways in which research within this paradigm can be improved to generate further insights into the information representations that boost probabilistic inference.

Where and when do difficulties in Bayesian reasoning arise? In the reviewed studies, Bayesian reasoning has typically been defined as the ability to provide a correct probability estimate given the information provided. Focusing on this endpoint has limited our ability to understand how or determine why many of the interventions reviewed in our meta-analysis do or do not work (McNair, 2015). We discuss two opportunities to build on insights from research with Bayesian textbook problems to explore how representation can affect how people process the information and to explore differences across a range of performance criteria.

First, the use of a narrow criterion to evaluate performance on these descriptive tasks restricts arguments about Bayesian inference to a rather narrow mathematical definition and detracts from other potential questions (Domurat, Kowalczuk, Idzikowska, Borzymowska,

& Nowak-Przygodzka, 2015; McNair, 2015; Wu et al., 2017). For example, given the different information formats, are people able to make the correct choice or inference given the information provided, irrespective of whether or not they can calculate a correct estimate? The results of Wu et al. (2017) and Domurat et al. (2015) suggest that making a correct choice does not necessarily imply that one calculates an exact estimate. Wu et al. (2017) employed an information selection task that required participants to select which of two tests had a better chance of answering a specific query (i.e., which of two genetic tests was most effective for identifying a species) presented either in natural frequency, conditional probability, or visual formats. The authors found no relation between probability judgement errors and choices, such that lower probability judgement errors were not necessarily associated with better choices. Similarly, Domurat et al. (2015), employing a natural sampling approach for participants to learn probabilities, found that the majority of participants made choices that satisfied Bayes' theorem despite participants' verbal reports suggesting that non-Bayesian solution strategies were often applied. These results highlight how different performance criteria can lead to different conclusions about people's capabilities. It is an open question as to whether the information formats and the moderators examined in the present meta-analysis would have similar effects given different performance criteria.

Second, focusing on the difficulties that arise during the solution process can help identify the features of information representations that underlie the facilitation effect. Questions related to process have not played a central role in many studies despite the fact that Gigerenzer and Hoffrage (1995) included both performance and process measures in their original study. Their analysis of the visual analogs and solution strategies participants wrote down while solving Bayesian reasoning tasks helped them to identify cognitive shortcuts that could lead to solution rates similar to Bayes' theorem given specific features of the problem. Process measures represented an important part of their ecological analysis, to identify cognitive mechanisms and to identify when cognitive shortcuts could lead to correct solutions, for both conditional probability and natural frequency formats. However, only recently has there been renewed interest in the processes leading up to a correct solution rather than on endpoints alone (Brase & Hill, 2015; Johnson & Tubau, 2015; McNair, 2015).

For instance, Sirota, Vallée-Tourangeau, Vallée-Tourangeau, and Juanchich (2015), McNair (2015), and Johnson and Tubau (2015) have suggested that we draw on theories in the field of problem-solving and mathematical cognition for insights into how people approach and solve Bayesian reasoning problems. For example, Sirota, Vallée-Tourangeau, et al. (2015) suggest decomposing the question *"What facilitates Bayesian reasoning?"* into *"What facilitates the insight?"* or understanding of information structure, and *"What facilitates the computation?"*. Similarly, McNair (2015) suggests that focusing on process can help to identify cognitive abilities that are influential in early stages of the problem-solving process, and may indicate a lack of formal knowledge, and those that occur later and indicate a lack of ability to apply knowledge. Johnson and Tubau (2015) propose that text comprehension and problem solving are interrelated processes, and that understanding the relation between these processes is central to improve our understanding why and at which point of the process different intervention strategies, such as natural frequencies, work. In the following section we suggest how different research methodologies offer opportunities to gain these insights.

Broadening methodological scope. A resounding criticism of work on Bayesian reasoning with elementary word problems is that these tasks are limited in what they can tell us about how people acquire, learn, or represent information about the occurrence of joint events (Brase & Hill, 2015; Girotto & Pighin, 2015; Mandel, 2014; Vallée-Tourangeau, Sirota, Juanchich, & Vallée-Tourangeau, 2015). For instance, in textbook tasks estimates are typically provided, and multiple estimates are not collected across time. Our understanding of probabilistic inference from descriptive tasks can be improved by incorporating different research methodologies, potentially leading to insights into theories and the assumptions on which they rest (as an example, consider research on the description-experience gap, Hertwig & Erev, 2009).

To test an assumption of the ecological rationality framework about how participants learn joint frequencies of events, Leach (2002) provided participants with either summary estimates of the relation between one of two symptoms and one of two diseases in a descriptive task, or with patient record cards that contained information about

symptoms and diseases in an experiential natural sampling task. In both conditions, participants made highly accurate posterior probability estimates about the relation between symptoms and diseases. Similarly, Vallée-Tourangeau, Abadie, and Vallée-Tourangeau (2015) found that participants who received natural frequency or conditional probability text formats improved when these were accompanied by interactive cards displaying the joint occurrences of events, although performance was superior when the text represented natural frequencies. For both formats, the interactive cards allowed participants to see the subset structure or the relation between hypotheses and data. As far as we are aware, this is the only study that combines interactive sampling of joint events with a conditional probability descriptive information format. Experience-based methodologies could also be employed to explore updating or revision of posterior probabilities given new information.

Further, process tracing methods such as eye-tracking and verbal or written protocols can be employed to examine where attention is focused, to gauge the weight a reasoner is giving to different pieces of information, or to identify when in the process deviations occur (Johnson & Tubau, 2015; McNair, 2015). Another promising line of research is to assess the prior distributions people have for different types of real events (Mandel, 2014). In this connection, Obrecht, Anderson, Schulkin, and Chapman (2012) asked obstetricians and gynaecologists to estimate the probability of Down syndrome given a positive test result and subsequently, to report estimates of the base rate (babies with Down syndrome), hit rate (positive test result given Down syndrome), and false-alarm rate (positive test result given no Down syndrome) in either natural frequencies or conditional probabilities. Cognisant of differences in the physician's personal experiences, Obrecht et al. (2012) examined accuracy in terms of whether the participants' base rate, hit rate, and false-alarm rate estimates were consistent with their posterior probability estimate. Posterior probability estimates were more consistent when information was requested in natural frequencies as opposed to conditional probabilities, a finding that generates questions about how people store information about the occurrence of joint events. There are many opportunities for research on Bayesian reasoning to broaden its methodological scope and explore a variety of research questions on information structure and reasoning through descriptive tasks.

2.6.4 From Textbook Problems to Real World Applications

At this point, it is important to remember that the study of Bayesian reasoning using word problems emerged in consideration of real world contexts where information is communicated in similar formats (e.g., Eddy, 1982). As Navarrete, Correia, Sirota, Juanchich, and Huepe (2015) emphasise, we should not lose sight of the ultimate goal to foster understanding in contexts where probabilistic inference from description problems is required. A few studies have sought to examine the effect of information representation formats in groups who are required to make these types of probabilistic inferences in a given domain, for example: medical professionals (Ben-Shlomo, Collin, Quekett, Sterne, & Whiting, 2015; Bramwell et al., 2006; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Hoffrage & Gigerenzer, 1998), managers (Hoffrage, Hafenbrädl, & Bouquet, 2015), and advanced law students and professional jurists (Lindsey, Hertwig, & Gigerenzer, 2003). Attempts to improve probabilistic inferences through training participants to translate probabilities into natural frequency representations (as opposed to rule-based training using Bayes' rule) have generally been shown to be effective over the longer term (Kurzenhäuser & Hoffrage, 2002; Sedlmeier & Gigerenzer, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a). However, we are unaware of more formalised efforts to implement training in curricula. Further, a better understanding of the errors participants make with each format and the implications of these errors on inference is needed. For example, future work could focus on the different cognitive shortcuts participants make and identify the conditions under which these shortcuts can approximate correct solutions (e.g., see Gigerenzer & Hoffrage, 1995; Hafenbrädl & Hoffrage, 2015).

2.6.5 Conclusions

The facilitative effect of natural frequencies is robust and our meta-analysis identified conditions under which performance on conditional probability problems can be improved further. Thus, although there remains room for improvement on natural frequency formats, the results of the meta-analysis suggest that natural frequencies are favourable to conditional probability formats. Even though short menu formats and visual aids can improve performance in both natural frequency and conditional probability formats, these moderators

are still better used with natural frequency formats, with visual aids offering the strongest advantage to performance. We had hoped to examine the relative benefits of different visual designs but were limited by the number of studies that have explored different design features and thus, further work is needed to establish which visual aids are most effective and why (although see Böcherer-Linder & Eichler, 2016 Wu et al., 2017, and Khan et al., 2015, for exceptions).

There is also preliminary evidence to suggest that higher numeracy and performance-based incentives can improve performance on Bayesian inference tasks. However, there is a lack of comparative studies examining these moderators, as well as non-university samples (e.g., children or experts), and textual manipulations aimed at improving problem comprehension (as opposed to texts aimed to emphasise set relations). The meta-analysis also identified how variations in performance could be explained by differences in study designs, such as when participants complete both problem formats. We suggest that current research methodologies employed in the study of elementary textbook tasks can be extended to incorporate more experience-based and process-tracing approaches. However, what we have learned from the many studies included in our review is that not only can natural frequencies improve Bayesian inference, but that there is still ample room for improvement. We hope that future work will focus not only on different performance criteria but on the processes leading up to the correct solution as well, with the aim to understand why many participants continue to have difficulties solving Bayesian inference tasks. Ultimately, future work will move beyond the current theoretical dichotomy of the ecological rationality framework and nested-sets theory to focus on integration, not only between but also beyond these two perspectives.

Appendix A Data of Meta-Analysis

Table 2.A
Relevant Data Used in Meta-Analysis

Estimate	format	*c*	*i*	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
Gigerenzer and Hoffrage (1995)																		
A) exp 1: standard	F	14	16	N	N	N	N	N	N	Y	N	Y	N	Y	Y	—	Y	14
A) exp 1: standard	P	5	25	N	N	N	N	N	N	Y	N	Y	N	Y	Y	—	Y	14
B) exp 1: short	F	15	15	Y	N	N	N	N	N	Y	N	Y	N	Y	Y	—	Y	14
B) exp 1: short	P	8	22	Y	N	N	N	N	N	Y	N	Y	N	Y	Y	—	Y	14
Hoffrage and Gigerenzer (1998)																		
A) ∅ all four problems	F	11	13	N	N	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
A) ∅ all four problems	P	2	22	N	N	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
Krauss et al. (1999)																		
A) version 1 vs 3	F	22	19	N	Y	N	N	N	N	Y	N	—	—	N	N	—	Y	0
A) version 1 vs 3	P	5	36	N	Y	N	N	N	N	Y	N	—	—	N	N	—	Y	0
Mellers and McGraw (1999)																		
A) exp 1: Nat, standard	F	13	33	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
A) exp 1: CP, standard	P	3	39	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
B) exp 1: Nat, joint	F	9	33	Y	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
B) exp 1: CP, joint	P	1	38	Y	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
C) exp 1: Sys, standard	P	7	28	N	N	N	N	Y	Y	N	N	N	N	N	N	—	Y	0
D) exp 1: Sys, joint	P	7	37	Y	N	N	N	Y	Y	N	N	N	N	N	N	—	Y	0
E) exp 2: Nat, standard	F	10	121	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
E) exp 2: CP, standard	P	5	117	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
F) exp 2: Nat, joint	F	24	115	Y	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
F) exp 2: CP, joint	P	15	81	Y	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
G) exp 2: Sys, standard	P	7	131	N	N	N	N	Y	Y	N	N	Y	N	N	N	—	Y	0
Evans et al. (2000)																		
A) exp 3: NF easy	F	8	20	N	N	N	Y	N	N	N	N	—	—	N	N	—	Y	7
A) exp 3: CP hard	P	5	21	N	N	N	Y	N	N	N	N	—	—	N	N	—	Y	7
Fiedler et al. (2000)																		
A) exp 1: incompatible	F	14	35	N	N	N	Y	N	Y	N	N	Y	N	N	N	—	Y	3
A) exp 1: incompatible	P	9	36	N	N	N	Y	N	Y	N	N	Y	N	N	N	—	Y	3
B) exp 1: common	F	24	29	Y	N	N	Y	N	Y	N	N	Y	N	N	N	—	Y	3
B) exp 1: common	P	20	31	Y	N	N	Y	N	Y	N	N	Y	N	N	N	—	Y	3
Hoffrage, Lindsey, Hertwig, and Gigerenzer (2000)																		
A) example 1, apps in medicine	F	43	53	N	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
A) example 1, apps in medicine	P	13	83	N	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
Macchi (2000)																		
A) PF	F	12	18	N	N	N	N	N	Y	Y	N	—	—	Y	N	—	Y	0
A) NPP	P	2	28	N	N	N	N	N	Y	Y	N	—	—	Y	N	—	Y	0
B) NPF	P	1	29	N	N	N	N	Y	Y	N	N	—	—	Y	N	—	Y	0
Sedlmeier and Gigerenzer (2001)																		
A) exp 2: strict, trees, test2	F	18	6	N	N	N	Y	N	N	Y	Y	Y	N	N	N	—	Y	6
A) exp 2: strict, trees, test2	P	18	6	N	N	N	Y	N	N	Y	Y	Y	N	N	N	—	Y	6
Lindsey et al. (2003)																		
A) law students, p(profile)	F	51	76	N	N	N	Y	N	Y	Y	N	Y	N	N	Y	—	Y	0
A) law students, p(profile)	P	1	126	N	N	N	Y	N	Y	Y	N	Y	N	N	Y	—	Y	0
B) jurists, p(profile)	F	20	7	N	N	N	Y	N	Y	Y	N	N	N	N	Y	—	Y	0
B) jurists, p(profile)	P	3	24	N	N	N	Y	N	Y	Y	N	N	N	N	Y	—	Y	0
Yamagishi (2003)																		
A) exp 1: viz	F	34	13	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
A) exp 1: viz	P	30	13	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
B) exp 1: no viz	F	17	23	N	N	Y	N	N	N	Y	N	Y	N	N	N	—	Y	0
B) exp 1: no viz	P	7	31	N	N	Y	N	N	N	Y	N	Y	N	N	N	—	Y	0
C) exp 2: viz	F	33	10	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
C) exp 2: viz	P	34	8	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
D) exp 2: no viz	F	25	20	N	N	Y	N	N	N	Y	N	Y	N	N	N	—	Y	0
D) exp 2: no viz	P	5	36	N	N	Y	N	N	N	Y	N	Y	N	N	N	—	Y	0
E) exp 3: viz a	F	160	42	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
E) exp 3: viz a	P	157	40	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
F) exp 3: viz b	F	141	60	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
F) exp 3: viz b	P	119	79	N	N	Y	N	N	N	Y	Y	Y	N	N	N	—	Y	0
Hoffrage and Gigerenzer (2004)																		
A) med students, short menu	F	63	33	Y	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
A) med students, short menu	P	39	57	Y	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
O'Brien et al. (2004)																		
A) exp 2: curta	F	12	13	Y	N	N	N	Y	N	Y	N	—	—	N	N	—	Y	0
A) exp 2: curta	P	3	22	Y	N	N	N	Y	N	Y	N	—	—	N	N	—	Y	0
B) exp 2: padrao	F	3.5	22.5	N	N	N	N	Y	N	Y	N	—	—	N	N	—	Y	0
B) exp 2: padrao	P	.5	25.5	N	N	N	N	Y	N	Y	N	—	—	N	N	—	Y	0
Bramwell et al. (2006)																		
A) pregnant women	F	3	18	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
A) pregnant women	P	1	21	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
B) companions	F	3	17	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
B) companions	P	3	17	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
C) midwives	F	.5	20.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0

Relevant Data Used in Meta-Analysis (continued)

Estimate	format	*c*	*i*	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
C) midwives	P	.5	22.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
D) obstetricians	F	13	7	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
D) obstetricians	P	1	20	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
Zhu and Gigerenzer (2006)																		
A) exp 1: adults	F	17	6	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	6
A) exp 1: adults	P	11	12	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	6
B) exp 2: adults	F	23	7	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	9
B) exp 2: adults	P	17	13	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	9
C) exp 2: 4th grade	F	6.5	24.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
C) exp 2: 4th grade	P	.5	10.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
D) exp 2: 5th grade	F	13.5	17.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
D) exp 2: 5th grade	P	.5	10.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
E) exp 2: 6th grade	F	14.5	16.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
E) exp 2: 6th grade	P	.5	10.5	N	N	N	N	N	N	Y	N	—	—	N	N	—	N	9
Brase (2008)																		
B) exp 1: NF	F	9.5	19.5	N	N	N	N	Y	N	Y	N	Y	N	N	N	—	Y	0
B) exp 1: Normalized	P	.5	28.5	N	N	N	N	Y	N	Y	N	Y	N	N	N	—	Y	0
Konheim-Kalkstein (2008)																		
A) exp 1: correct solutions	F	7	24	N	N	N	Y	N	N	Y	N	—	—	N	N	—	Y	0
A) exp 1: correct solutions	P	2	29	N	N	N	Y	N	N	Y	N	—	—	N	N	—	Y	0
B) exp 4: lo num	F	2	16	N	N	N	Y	N	Y	N	N	—	—	N	N	N	Y	0
B) exp 4: lo num	P	1	23	N	N	N	Y	N	Y	N	N	—	—	N	N	N	Y	0
C) exp 4: hi num	F	9	26	N	N	N	Y	N	Y	N	N	—	—	N	N	Y	Y	0
C) exp 4: hi num	P	8	23	N	N	N	Y	N	Y	N	N	—	—	N	N	Y	Y	0
Brase (2009a)																		
A) course req., no viz	F	12.5	39.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
A) course req., no viz	P	.5	50.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
B) course req., viz	F	21	31	N	N	N	N	N	Y	Y	Y	Y	N	N	N	—	Y	0
B) course req., viz	P	14	36	N	N	N	N	N	Y	Y	Y	Y	N	N	N	—	Y	0
C) flat pay, no viz	F	14	34	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
C) flat pay, no viz	P	2	45	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
D) flat pay, viz	F	16	32	N	N	N	N	N	Y	Y	Y	Y	N	N	N	—	Y	0
D) flat pay, viz	P	13	36	N	N	N	N	N	Y	Y	Y	Y	N	N	N	—	Y	0
E) var. pay, no viz	F	23	19	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
E) var. pay, no viz	P	4	38	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
F) var. pay, viz	F	28	14	N	N	N	N	N	Y	Y	Y	N	Y	N	N	—	Y	0
F) var. pay, viz	P	17	25	N	N	N	N	N	Y	Y	Y	N	Y	N	N	—	Y	0
Chapman and Liu (2009)																		
A) medical, lo num	F	5.5	60.5	N	N	N	N	N	N	Y	N	Y	N	N	Y	N	Y	0
A) medical, lo num	P	.5	87.5	N	N	N	N	N	N	Y	N	Y	N	N	Y	N	Y	0
B) car, lo num	F	9.5	78.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
B) car, lo num	P	.5	65.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
C) medical, hi num	F	26	66	N	N	N	N	N	N	Y	N	Y	N	N	Y	Y	Y	0
C) medical, hi num	P	1	91	N	N	N	N	N	N	Y	N	Y	N	N	Y	Y	Y	0
D) car, hi num	F	29	63	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
D) car, hi num	P	7	85	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
Galesic, Gigerenzer, and Straubinger (2009)																		
A) older adults, hi num	F	3	10	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	N	1
A) older adults, hi num	P	2	7	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	N	1
B) younger adults, hi num	F	15	28	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
B) younger adults, hi num	P	4	38	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
C) older adults, lo num	F	2.5	9.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	N	1
C) older adults, lo num	P	.5	14.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	N	1
D) younger adults, lo num	F	1.5	12.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
D) younger adults, lo num	P	.5	17.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
Misuraca et al. (2009)																		
A) ø cond. 1-4 vs 5	F	27.5	93.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
A) ø cond. 1-4 vs 5	P	.5	30.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	Y	0
Moro et al. (2011)																		
A) exp 1: NF	F	7.5	14.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
A) exp 1: Nested-sets CP	P	.5	21.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
D) exp 2: gemstone, viz	F	13	5	N	N	N	N	N	N	Y	Y	Y	N	N	Y	—	Y	0
D) exp 2: gemstone, viz	P	1	20	N	N	N	N	N	N	Y	Y	Y	N	N	Y	—	Y	0
E) exp 2: gemstone, no viz	F	16	2	N	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
E) exp 2: gemstone, no viz	P	1	22	N	N	N	N	N	N	Y	N	Y	N	N	Y	—	Y	0
Siegrist and Keller (2011)																		
A) exp 1: mammography	F	15	117	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	0
A) exp 1: mammography	P	1	133	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	0
B) exp 2: mammography	F	9	68	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	0
B) exp 2: mammography	P	1	70	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	0
C) exp 3: ø social & cookie	F	55	81	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	1
C) exp 3: ø social & cookie	P	9	61	N	N	N	N	N	N	Y	N	—	—	Y	N	—	N	1
Tsai, Miller, and Kirlik (2011)																		
A) NF	F	6	6	N	N	N	N	N	N	Y	N	Y	N	Y	N	—	Y	5
A) CP	P	4	8	N	N	N	N	N	N	Y	N	Y	N	Y	N	—	Y	5
Hill and Brase (2012)																		
A) exp 2: ø med & NB, lo num	F	1.5	32.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
A) exp 2: ø med & NB, lo num	P	.5	37.5	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
B) exp 2: ø med & NB, hi num	F	10	36	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
B) exp 2: ø med & NB, hi num	P	2	45	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
C) exp 3: ø med & NB, lo num	F	1	49	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1
C) exp 3: ø med & NB, lo num	P	1	58	N	N	N	N	N	N	Y	N	Y	N	N	N	N	Y	1

Relevant Data Used in Meta-Analysis (continued)

Estimate	format	c	i	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
D) exp 3: ø med & NB, hi num	F	9	38	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
D) exp 3: ø med & NB, hi num	P	3	38	N	N	N	N	N	N	Y	N	Y	N	N	N	Y	Y	1
Ferguson and Starmer (2013)																		
A) experts, incentive, short	F	8	3	Y	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
A) experts, incentive, short	P	8	5	Y	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
B) novices, incentive, short	F	15	7	Y	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
B) novices, incentive, short	P	4	19	Y	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
C) experts, incent., standard	F	9	3	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
C) experts, incent., standard	P	1	7	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
D) novices, incent., standard	F	11.5	14.5	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
D) novices, incent., standard	P	.5	24.5	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	0
E) experts, no incent., short	F	8	4	Y	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
E) experts, no incent., short	P	1	9	Y	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
F) novices, no incent., short	F	16	18	Y	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
F) novices, no incent., short	P	5	28	Y	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
G) experts, no incent., stand.	F	6	6	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
G) experts, no incent., stand.	P	1	11	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
H) novices, no incent., stand.	F	11	24	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
H) novices, no incent., stand.	P	1	33	N	N	N	N	N	N	Y	N	—	—	N	N	—	Y	0
Johnson and Tubau (2013)																		
A) exp 1: complicated, hi num	F	20	6	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
A) exp 1: complicated, hi num	P	2	24	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
B) exp 1: complicated, lo num	F	6.5	16.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
B) exp 1: complicated, lo num	P	.5	22.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
C) exp 1: simple, hi num	F	30	4	N	N	Y	N	N	N	N	N	Y	N	N	Y	Y	Y	0
C) exp 1: simple, hi num	P	12	22	N	N	Y	N	N	N	N	N	Y	N	N	Y	Y	Y	0
D) exp 1: simple, lo num	F	12	3	N	N	Y	N	N	N	N	N	Y	N	N	Y	N	Y	0
D) exp 1: simple, lo num	P	1	14	N	N	Y	N	N	N	N	N	Y	N	N	Y	N	Y	0
E) exp 2: long, hi num	F	16	5	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
E) exp 2: long, hi num	P	4	16	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
F) exp 2: long, lo num	F	9.5	11.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
F) exp 2: long, lo num	P	.5	23.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
G) exp 2: short, hi num	F	18	2	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
G) exp 2: short, hi num	P	5	16	N	N	N	N	N	N	N	N	Y	N	N	Y	Y	Y	0
H) exp 2: short, lo num	F	17.5	6.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
H) exp 2: short, lo num	P	.5	20.5	N	N	N	N	N	N	N	N	Y	N	N	Y	N	Y	0
Lesage et al. (2013)																		
A) exp 1: total sample, relative	F	18.5	23.5	N	N	N	N	N	N	Y	N	N	Y	N	N	—	Y	1
A) exp 1: total sample, relative	P	.5	42.5	N	N	N	N	N	Y	N	N	Y	N	N	N	—	Y	1
B) exp 1: no tot.samp., relative	F	17	26	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	1
B) exp 1: no tot.samp., relative	P	2	41	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	1
C) exp 1: total sample, absolute	F	18	26	Y	N	N	N	N	Y	N	N	Y	N	N	N	—	Y	1
C) exp 1: total sample, absolute	P	15	29	Y	N	N	N	N	Y	N	N	Y	N	N	N	—	Y	1
D) exp 1: no tot.samp., absolute	F	27	14	Y	N	N	N	N	N	N	N	Y	N	N	N	—	Y	1
D) exp 1: no tot.samp., absolute	P	13	30	Y	N	N	N	N	N	N	N	Y	N	N	N	—	Y	1
E) exp 2: relative	F	18	24	N	N	N	N	N	N	N	N	N	N	N	N	—	N	1
E) exp 2: relative	P	6	44	N	N	N	N	N	N	N	N	N	N	N	N	—	N	1
F) exp 2: absolute	F	30	21	Y	N	N	N	N	N	N	N	N	N	N	N	—	N	1
F) exp 2: absolute	P	21	26	Y	N	N	N	N	N	N	N	N	N	N	N	—	N	1
Friederichs et al. (2014)																		
A) NF, no viz	F	11	18	N	N	N	Y	N	N	Y	N	—	—	N	N	—	Y	2
A) CP, no viz	P	5	29	N	N	N	Y	N	N	Y	N	—	—	N	N	—	Y	2
Sirota, Juanchich, and Hagmayer (2014)																		
A) exp 2: med & children's	F	86	65	N	N	N	N	N	N	Y	N	—	—	Y	Y	—	Y	3
A) exp 2: med & children's	P	30	121	N	N	N	N	N	N	Y	N	—	—	Y	Y	—	Y	3
Binder et al. (2015)																		
A) mammography, no viz, task 1	F	1.5	19.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
A) mammography, no viz, task 1	P	.5	22.5	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
B) mammography, viz tree, task 1	F	8.5	15.5	N	N	N	N	N	N	Y	Y	N	N	N	N	—	N	0
B) mammography, viz tree, task 1	P	.5	20.5	N	N	N	N	N	N	Y	Y	N	N	N	N	—	N	0
C) economics, no viz, task 1	F	11	10	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
C) economics, no viz, task 1	P	1	20	N	N	N	N	N	N	Y	N	N	N	N	N	—	N	0
D) economics, viz tree, task 1	F	11	9	N	N	N	N	N	N	Y	Y	N	N	N	N	—	N	0
D) economics, viz tree, task 1	P	1	20	N	N	N	N	N	N	Y	Y	N	N	N	N	—	N	0
Hill and Brase (2015)																		
A) MTurk, ST	F	9.5	31.5	N	N	N	N	N	N	N	N	Y	N	N	N	—	N	2
A) MTurk, ST	P	.5	37.5	N	N	N	N	N	N	N	N	Y	N	N	N	—	N	2
B) MTurk, ø BC & CF	F	10	30	N	Y	N	N	N	N	N	N	Y	N	N	N	—	N	2
B) MTurk, ø BC & CF	P	1	36	N	Y	N	N	N	N	N	N	Y	N	N	N	—	N	2
C) online, ST	F	1.5	45.5	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	2
C) online, ST	P	.5	46.5	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	2
D) online, ø BC & CF	F	4	42	N	Y	N	N	N	N	N	N	Y	N	N	N	—	Y	2
D) online, ø BC & CF	P	1	45	N	Y	N	N	N	N	N	N	Y	N	N	N	—	Y	2
E) paper, ST	F	7	42	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	2
E) paper, ST	P	1	47	N	N	N	N	N	N	N	N	Y	N	N	N	—	Y	2
F) paper, ø BC & CF	F	5.5	44.5	N	Y	N	N	N	N	N	N	Y	N	N	N	—	Y	2
F) paper, ø BC & CF	P	.5	48.5	N	Y	N	N	N	N	N	N	Y	N	N	N	—	Y	2
Hoffrage, Krauss, et al. (2015)																		
A) exp 1: 2 hyp, 3 cue values	F	18.5	14.5	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
A) exp 1: 2 hyp, 3 cue values	P	.5	32.5	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
B) exp 1: 3 hyp, 1 dichot cue	F	12	20	N	N	Y	N	N	N	Y	N	—	—	Y	Y	—	Y	1
B) exp 1: 3 hyp, 1 dichot cue	P	3	29	N	N	Y	N	N	N	Y	N	—	—	Y	Y	—	Y	1
C) exp 1: 2 hyp, 2 dichot cues	F	11.5	21.5	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1

Relevant Data Used in Meta-Analysis (continued)

Estimate	format	c	i	short menu	2+ cues	3 hypot.	prob. quest.	freq. quest.	enum. pop.	mult. events	vis. aid	show fee	perf. pay	strict scor.	both formats	high num.	exp.	add. probs.
C) exp 1: 2 hyp, 2 dichot cues	P	.5	32.5	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
D) exp 1: 2 hyp, 3 dichot cues	F	12	20	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
D) exp 1: 2 hyp, 3 dichot cues	P	1	31	N	Y	N	N	N	N	Y	N	—	—	Y	Y	—	Y	1
Hoffrage, Hafenbrädl, and Bouquet (2015)																		
A) undergraduates, all tasks	F	53	79	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
A) undergraduates, all tasks	P	19	108	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
B) junior executives, all tasks	F	23	34	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
B) junior executives, all tasks	P	18	45	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
C) senior executives, all tasks	F	11	19	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
C) senior executives, all tasks	P	7	24	N	N	N	N	N	N	N	N	—	—	Y	N	—	Y	1
Vallée-Tourangeau, Abadie, and Vallée-Tourangeau (2015)																		
A) exp 1&2: lo interact., hi num	F	17	11	N	N	N	N	N	N	Y	N	N	N	N	N	Y	Y	2
A) exp 1&2: lo interact., hi num	P	3	24	N	N	N	N	N	N	Y	N	N	N	N	N	Y	Y	2
B) exp 1&2: hi interact., hi num	F	20	6	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	2
B) exp 1&2: hi interact., hi num	P	21	12	N	N	N	N	N	N	Y	Y	N	N	N	N	Y	Y	2
C) exp 1&2: lo interact., lo num	F	7	10	N	N	N	N	N	N	Y	N	N	N	N	N	N	Y	2
C) exp 1&2: lo interact., lo num	P	1	17	N	N	N	N	N	N	Y	N	N	N	N	N	N	Y	2
D) exp 1&2: hi interact., lo num	F	12	7	N	N	N	N	N	N	Y	Y	N	N	N	N	N	Y	2
D) exp 1&2: hi interact., lo num	P	4	8	N	N	N	N	N	N	Y	Y	N	N	N	N	N	Y	2
E) exp 3: standard	F	7.5	10.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
E) exp 3: standard	P	.5	17.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
F) exp 3: fleshed out	F	5.5	13.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0
F) exp 3: fleshed out	P	.5	18.5	N	N	N	N	N	N	Y	N	Y	N	N	N	—	Y	0

Notes: c and i denote the number of correct and incorrect responses, respectively; when $c = 0$ or $i = 0$, 0.5 was added to all counts of the same experiment; Y denotes the presence and N denotes the absence of a study characteristic; effects from studies including chances formats (Brase, 2008; Girotto & Gonzalez, 2001; Sirota, Kostovičová, & Vallée-Tourangeau, 2015a) are excluded; table includes only the moderators used in the analyses.

Chapter 3

Satisficing: Integrating Two Traditions

This chapter is forthcoming in *Journal of Economic Literature* as: Artinger, F., Gigerenzer, G., & Jacobs, P., Satisficing: Integrating Two Traditions.

Ever since Simon (1955) initiated the behavioral revolution in economics, its poster child has been satisficing. Satisficing refers to the observation that agents make choices with the help of aspiration levels that do not necessarily coincide with utility maximization. The normative appeal of utility maximization has led many to dismiss satisficing uniformly as an undesirable quirk of human behavior. In this article, we distinguish two separate research traditions that can be traced back to Simon's (1955) original visions of satisficing but are largely disconnected today. Reviewing both traditions, we integrate them within a framework for understanding decision making beyond utility maximization.

Specifically, we argue that the rationality of satisficing strategies depends on the class of decision environment. Broadly, environments can be divided into two classes. Under risk, optimization sets the rational benchmark and satisficing can yield suboptimal decisions. Under uncertainty or intractability, where the optimal action cannot be determined, satisficing can outperform complex strategies, including rational choice models. Satisficing has been examined in both types of decision environments, resulting in two distinct and largely unconnected literatures.

In section 3.1 we trace the historical trajectory of the notion of satisficing and provide a conceptual overview of the meaning of the term. In section 3.2, we provide a review of the two traditions in research on satisficing that have evolved. In section 3.3, we propose a unifying framework to study when and why satisficing can be rational

and what that means. Section 3.4 closes with four methodological conclusions, advocating competitive out-of-sample tests to evaluate decision strategies under uncertainty and intractability.

3.1 Satisficing

3.1.1 Historical Context

In his seminal contribution to economics, Herbert Simon advocated and developed a model of bounded rationality, positing that "the task is to replace the global rationality of economic man with a kind of rational behavior that is compatible with the access to information and computational capacities that are actually possessed by organisms, including man, in the kinds of environments in which such organisms exist." (Simon, 1955, p.99). The period in which this paper was published was characterized by the popularization of neoclassical rational choice theory, or global rationality, as Simon referred to it. This body of theories includes von Neumann and Morgenstern's (1944) and Savage's (1954) work on expected utility theory and the work by (Nash, 1950) on equilibrium in non-cooperative games. Common to these theories is the assumption of an agent who has complete information about the available alternatives, including perfect foresight about all possible consequences and sufficient knowledge of the probabilities with which they occur. Agents are then predicted to act as if they were solving an optimization problem to maximize expected utility. M. Friedman (1953, p.15) maintained that such an assumption is justified, irrespective of whether it is deemed realistic, because "the relevant question to ask about the 'assumptions' of a theory is not whether they are descriptively 'realistic', for they never are, but whether they are sufficiently good approximations for the purpose in hand. And this question can be answered only by seeing whether the theory works, which means whether it yields sufficiently accurate predictions". Compared with this approach, Simon's differed in two respects: First, his interest was in models of agents' actual decision processes, not only of their outcomes. Second, he was interested in situations "where the conditions for rationality postulated by the model of neoclassical economics are not met" (Simon, 1989, p.377).

The early responses in economics to Simon's writings were twofold. On the one hand, the proposition of bounded rationality evoked

vigorous defenses of rational choice theory[1]. At the same time, many economists were somewhat open to the notion of bounded rationality. This group included Robert Solow, who — reviewing Simon's (1957) book "Models of Man: Social and Rational", which expands on his 1955 paper — found himself "torn between an impulse to display the interdisciplinary scope of my ignorance by commenting on every essay and a more rational disposition to fight it out along the main line" (Solow, 1958, p.81). Although the book received very positive reviews by some economists (e.g., Shubik, 1958), Simon's departure from the neo-classical economic canon presumably made it difficult for many economists to incorporate his ideas into their theorizing.

For the years up to 1969, we found only 35 citations of Simon's 1955 article on Web of Science. Only during the 1970s did his early contributions start to gain recognition in the economic literature. Inspired by Simon's being awarded the 1978 Nobel Memorial Prize in Economic Sciences, a community of economists and psychologists dedicated their work to studying the behavioral foundations of economic theory, which developed into behavioral economics. Today, Simon's work is often cited as the predecessor of Tversky and Kahneman's (1974) work on heuristics and biases. However, not until 1981 did Tversky and Kahneman begin to relate their work to Simon's study of bounded rationality (Gigerenzer, 2004). Yet bounded rationality does not mean the same in both programs. To Simon, it meant the study of behavior in situations where the conditions assumed in neoclassical economics are not met, whereas Kahneman and Tversky assumed that these conditions are met and that deviating behavior implies a lack of rationality. Simon (1985, p.297) made this difference between them clear: "Bounded rationality is not irrationality."

At the same time, both approaches to behavioral economics can be characterized as empirically falsifying the assumptions underlying neoclassical economic theory (for complementary reviews on the topic, see Crawford, 2013; Harstad & Selten, 2013; Rabin, 2013). Unlike Simon, however, the heuristics-and-biases program attributed behavioral deviations from neoclassical theory to flaws in people's minds rather than to potential flaws in the application of the theory. This allowed contemporary behavioral economics to retain the underlying norm of an agent who integrates all information and maximizes utility.

[1] For an overview of the arguments put forth in favor of rational choice theory over the years as well as the counter-arguments, see (Conlisk, 1996).

Simon's writings, in contrast, were followed by research that studied decisions beyond the domain of rational choice theory, including the work by Cyert and March (1963) on the behavioral theory of the firm, Winter's (1971) work on evolutionary economics, and the work by Gigerenzer and colleagues on fast-and-frugal heuristics (Gigerenzer, Hertwig, & Pachur, 2011; Gigerenzer & Selten, 2001). These analyses are based on a satisficing agent and study the decision processes, routines, and rules of thumbs that agents and organizations actually use when facing complex and dynamic environments that provide only limited information.

Winter (1971) provides a quote of Simon as a central source for his own inspiration:

> "The equilibrium behavior of a perfectly adapting organism depends only on its goals and its environment; it is otherwise completely independent of the internal properties of the organism (...) [T]o predict the short-run behavior of an adaptive organism, or its behavior in a complex and rapidly changing environment, it is not enough to know its goals. We must also know a great deal about its internal structure and particularly its mechanisms of adaptation." (Simon, 1959, p.255)

That is, equilibrium strategies derived from a stylized representation of the world, specifically its incentive structure, can substantially differ from the strategies that agents actually use to navigate an uncertain and complex world. Going back to V. L. Smith (1962), there is a substantial literature in economics demonstrating that equilibrium also obtains with naïve, merely privately informed agents (for a review, see V. L. Smith, 2008). Gode and Sunder (1993) even find that zero-intelligence traders, who randomize within their budget constraints, produce allocative efficiency. Much of the work in behavioral economics does not make a clear distinction between the individual and the aggregate levels of analysis but, in contrast to Friedman, interprets the methodological tool of *homo economicus* literally and sets out to refute it.

In order to account for differences between an equilibrium perspective and the actual behavior of an agent, V. L. Smith (2008) proposes a distinction between two types of analyses[2]. The first, *con-*

[2]The principal distinction between two such rational orders can already be found

structivist rationality, applies deductive reasoning from first principles: It identifies the incentive structure and deduces the equilibrium by sufficiently abstracting and simplifying. In contrast, an analysis of *ecological rationality* proceeds empirically by determining the decision strategies used by agents and then evaluating the performance of that strategy competitively against other relevant strategies in the given context. The term ecological rationality thereby refers to the degree to which a strategy is adapted to the environment, evaluated in terms of a fitness measure such as profit or accuracy of predictions (Gigerenzer et al., 1999).

In the present article, we examine satisficing through the lens of ecological rationality. This perspective, yet uncommon in economics, offers a framework for thinking about decision strategies in a broader way. As we will argue, it explains how Simon's early writings inspired two largely distinct research traditions. Because this perspective examines strategies relative to the environment, we include a short primer on different degrees of uncertainty that lead to fundamentally different classes of decision environments.

3.1.2 Risk, Ambiguity, Intractability, and Uncertainty

Keynes (1921) and Knight (1921) use a dichotomy of two broad categories of environments, both of which are characterized by the absence of certainty: risk and uncertainty. Whereas risk is commonly understood, uncertainty has been assigned different meanings. In order to define those relevant for this article, we begin with the terminology of Savage (1954) developed in "Foundations of Statistics," in which he axiomatized subjective expected utility theory. Building on this terminology allows us to offer a more detailed definition of different kinds of uncertainty (see also Table 3.1).

Savage defines a decision problem as a pair $\{S, C\}$, where S is the exhaustive and mutually exclusive set of all future states of the world and C the exhaustive set of their consequences associated with each alternative. The alternatives or actions are defined on $\{S, C\}$, and each state s in S has an assigned probability . Choice under certainty means that for each alternative, there is only one state with probability 1; all others have probabilities 0. Choice under

in the writings of A. Smith (1776/1976, 1779/1981), Hume (1739/2000), and later von Hayek (1937, 1945), as well as Savage (1954) and Simon (1955, 1956).

Table 3.1
Decision Environments

Decision Environment Under Risk					
Alternative	State 1	State 2	State 3	$\cdots$	State M
	$p = p_1$	$p = p_2$	$p = p_3$	$\cdots$	$p = p_M$
a_1	$c_{1,1}$	$c_{1,2}$	$c_{1,3}$	$\cdots$	$c_{1,M}$
a_2	$c_{2,1}$	$c_{2,2}$	$c_{2,3}$	$\cdots$	$c_{2,M}$
a_3	$c_{3,1}$	$c_{3,2}$	$c_{3,3}$	$\cdots$	$c_{3,M}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_N	$c_{N,1}$	$c_{N,2}$	$c_{N,3}$	$\cdots$	$c_{N,M}$
Decision Environment Under Ambiguity					
Alternative	State 1	State 2	State 3	$\cdots$	State M
	$p =$**??**	$p =$**??**	$p =$**??**	$\cdots$	$p =$**??**
a_1	$c_{1,1}$	$c_{1,2}$	$c_{1,3}$	$\cdots$	$c_{1,M}$
a_2	$c_{2,1}$	$c_{2,2}$	$c_{2,3}$	$\cdots$	$c_{2,M}$
a_3	$c_{3,1}$	$c_{3,2}$	$c_{3,3}$	$\cdots$	$c_{3,M}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
a_N	$c_{N,1}$	$c_{N,2}$	$c_{N,3}$	$\cdots$	$c_{N,M}$
Decision Environment Under Uncertainty					
Alternative	State 1	State 2	State 3	$\cdots$	**??**
	$p =$**??**	$p =$**??**	$p =$**??**	$\cdots$	$p =$**??**
a_1	$c_{1,1}$	$c_{1,2}$	$c_{1,3}$	$\cdots$	**??**
a_2	$c_{2,1}$	$c_{2,2}$	$c_{2,3}$	$\cdots$	**??**
a_3	$c_{3,1}$	$c_{3,2}$	$c_{3,3}$	$\cdots$	**??**
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
??	**??**	**??**	**??**	$\cdots$	**??**

risk means that more than one state has non-zero probability and that the probabilities attached to each state are known; the expected utility of an alternative is the sum of the consequences multiplied by their respective probabilities over all possible states. A situation of ambiguity is identical to this apart from the probability distribution not being known, as in the gambles underlying the Ellsberg paradox (Ellsberg, 1961; see also Anscombe and Aumann, 1963). What these three situations — certainty, risk, and ambiguity — have in common is that the complete set of alternatives, future states of the world, and consequences is known. Such problems are said to be *well-defined*.

A well-defined problem can be tractable or not. Any decision problem under certainty, risk, and ambiguity is considered computationally intractable (with subdivisions into NP-hard, NP-complete, etc.) if the set of alternatives or states is so large that the best one cannot be identified by mind or machine. This means that no efficient (i.e., polynomial-time) algorithm exists to solve it (e.g., Garey & Johnson, 1979). Examples include games such as chess and Go. To understand the order of magnitude of this limitation, note that chess has approximately 10^{120} unique sequences of moves or games, a number greater than the estimated number of atoms in the universe (Shannon, 1950). Many important tasks are intractable, including scheduling, capital budgeting, and itinerary problems, among others (Markose, 2005; Papadimitriou & Steiglitz, 1998). Savage (1954, p.16) is explicit that intractable problems are outside of the domain of expected utility theory. Solving intractable problems requires a different kind of decision theory, in both descriptive and normative terms, that includes heuristic strategies.

The final class of problems involves degrees of what we call uncertainty, which has elsewhere been termed radical or fundamental uncertainty (King & Kay, 2020). In economics, the terms ambiguity and uncertainty are commonly used interchangeably (Etner, Jeleva, & Tallon, 2012). However, the distinction between them is fundamental. Ambiguity means that a problem is well-defined, that is, the exhaustive set of alternatives, possible states, and their consequences is known. Uncertainty, in contrast, means that the problem is ill-defined, that the exhaustive set of states of the world and their consequences is not knowable or foreseeable at the point of decision making. Savage (1954, p.16) lists as an example planning a picnic, where events can occur that one cannot know ahead of time. He points out that expected utility theory cannot and should not be applied under uncertainty.

One of the contributions of this article is to relate these classes of situations to satisficing. In section 2, we will show that there are two different traditions of satisficing, one assuming well-defined situations such as risk and the other addressing situations of uncertainty and intractability. First, however, we define the basic concepts of satisficing.

3.1.3 Satisficing: Definition

To illustrate his vision of bounded rationality, Simon spends a good portion of his landmark 1955 article describing a satisficing decision strategy. Such a strategy, he posits, is more descriptive of human decision processes than the traditional model of rational choice, for which he sees "a complete lack evidence that, in actual human choice situations of any complexity, these computations can be, or are in fact, performed" (Simon, 1955, p.104). His alternative model consists of two elements that are characteristic for choice processes: (i) the aspiration level, which is a simplified value function that can be adapted over time, and (ii) search.

The first element, a direct simplification of neoclassical theory, is perhaps the most controversial element of his proposal. Consistent with our earlier notation, Simon (1955, p.104) suggests that "[o]ne route to simplification is to assume that [the value function] $V(s)$ necessarily assumes one of two values, $(1; 0)$, or one of three values, $(1; 0; -1)$, for all s in S. Depending on the circumstance, we might want to interpret these values, as (a) (satisfactory or unsatisfactory), or (b) (win, draw or lose)." A binary value function implies the use of an aspiration level, which refers to the minimum level of a given scale of interest that is acceptable to the agent in order to choose an alternative. The term aspiration level has a long tradition in psychological theory (Gardner, 1940), appearing first in German (as "Anspruchsniveau") in work by Dembo (1931) and then in English (as "level of aspiration") in work by (Lewin, 1935). Although Simon first defines the aspiration level in terms of the value or utility of an alternative, he later maintains that it is more realistically set separately for each attribute under consideration (cf. Simon, 1955, p.109–110). When choosing among different houses, for example, agents may have multiple aspiration levels, one for each attribute such as price, floor size, location, or number of rooms, etc. instead of one aspiration level for the overall utility of each alternative. In this view, agents are not assumed to integrate different attributes on a cardinal scale but instead to evaluate each attribute separately. This explicitly allows for incommensurability, the proverbial comparison of apples and oranges where one cannot readily compare two attributes on the same scale.

Aspiration levels do not necessarily remain constant over time. After one or more alternatives are encountered that do not meet the

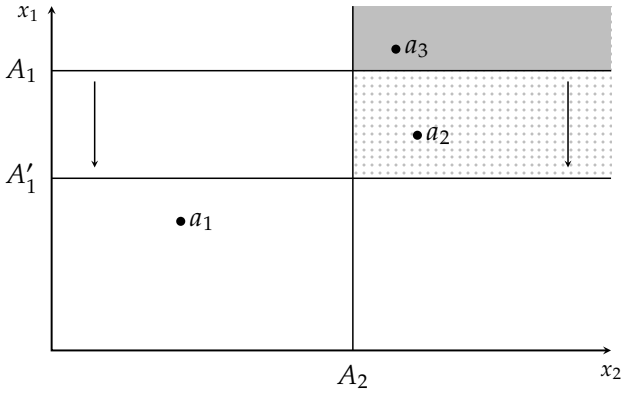

Figure 3.1: *Aspiration-level adaptation. Aspiration levels for alternatives a_1, a_2, a_3 with two attributes, x_1 and x_2, initially set at A_1 and A_2, which defines the set of acceptable alternatives (shaded area). At $t = t_1$, A_1 is lowered to A_1' while A_2 remains. Alternative a_1 lies below both aspiration levels and is never chosen, whereas a_3 is always chosen and a_2 is chosen only after t_1.*

aspiration level, the agent may adjust the level and then proceed to examine the next alternative until one is encountered that fulfills the most recent aspiration level. For an illustration, consider Figure 3.1, which extends Figure II of Simon (1955). Aspiration levels for alternatives a_1, a_2, a_3 with two attributes, x_1 and x_2, are initially set at A_1 and A_2, which defines the set of acceptable alternatives (shaded area). At $t = t_1$, A_1 is lowered to A_1', while A_2 remains. This adjustment redefines the set of acceptable options from the shaded area to the shaded and dotted areas. Alternative a_1 lies below both aspiration levels and is never chosen, whereas a_3 is chosen at all times and a_2 is chosen only after t_1.

The division into acceptable and unacceptable alternatives could be interpreted as a choice of a less-than-optimal alternative. Simon (1955) addresses such concerns by pointing out that the dichotomy of the value function can reflect preferences (e.g., as an approximation to a value function with sufficiently strong decreasing marginal returns) but can also be considered an element to navigate the environment, for example, in situations where alternatives are decided upon sequentially. Note that this implies that the alternative chosen is the "best so far", a phrase that Simon later said is a better summary of

the satisficing idea than the common notion of an alternative that is "good enough" (Gigerenzer, 2004).

The second element of Simon's satisficing assumes that the agent has to search for information. Simon points to two possibilities for exploration. First, the agent is unaware of the complete set of alternatives and discovers alternatives sequentially. Second, the agent is aware of all alternatives but is agnostic about the complete set of states of the world that each alternative entails. The agent can search either externally by acquiring additional pieces of information or internally by picturing the consequences of different alternatives. In both cases, information gathering leads agents to sequentially explore alternative-state combinations.

Both elements of satisficing, aspiration level and search, jointly define the decision process in situations where the agent cannot fully explore the space spanned by alternatives and states[3]. The aspiration level governs both the quality of the alternative chosen and the duration of the search process — even in situations where the cost of gathering information is not known to the agent. Setting the aspiration level therefore is at the heart of a satisficing model.

Generally, models of aspiration level setting and adaptation can take many forms, ranging from computationally intensive methods, including Bayesian approaches, to simple ones. Although Simon himself derived a computationally intensive method in the appendix of his 1955 paper (see also Gilboa & Schmeidler, 2001; Wall, 1993), he notes that such a method is psychologically implausible given agents' typical lack of necessary information and the complexity of the models. Instead, he posits that an agent "will set his acceptance [level] quite high, watch the distribution of offers he receives, and gradually and approximately adjust his acceptance [level] downward or upward until he receives an offer he accepts — without ever making probability calculations." (Simon, 1955, p.117). Simon's basic model of aspiration-level adaptation can be summarized as follows:

[3]Simon provides the example of a chess player pondering the next move by simulating internally how the game would continue under different alternatives until a move is found that clearly leads to a winning position. He points out that this particular task is only manageable by using a satisficing strategy, as it reduces the space of alternatives from an estimated 10^{24} to fewer than 100 moves to be considered.

Step 1: Set an aspiration level.
Step 2: Continue search until finding the first alternative that meets or exceeds the aspiration level.
Step 3: If no alternative meets the aspiration level within a fixed period, adapt it by a particular value and return to Step 2.

In his article of 1955, Simon leaves his class of models unnamed and only later introduces the term satisficing, in a follow-up article in Psychological Review (Simon, 1956). In that article, he shows how an organism with very limited cognitive capabilities can successfully apply a well-adapted satisficing strategy to maintain its subsistence. Although the organism in the article of 1956 relies solely on a satisficing strategy, Simon highlights that the satisficing model is but one strategy for many realistic decision situations; other situations may call for different strategies (Simon, 1955, p.104).

3.2 Two Traditions of Satisficing

Ever since Simon's inception of the term satisficing, it has been used widely and diversely in the literature, which has produced a fragmented understanding of it. In its most general sense, satisficing is defined simply as the antithesis of optimization, without imposing additional constraints on the decision model. By this definition, any decision model that does not rely on optimization techniques "satisfices". We could not find this interpretation of satisficing in Simon's early writings, but it emerged later (e.g., Simon, 1979), after the term had become emblematic of his more general critique of rational choice theory.

We argue that satisficing has been understood in two different ways, each of which originates in the writings of Simon. Both research traditions make use of aspiration levels in their decision models, albeit in different ways. The key difference between these two traditions lies in the different decision environments they assume: Whereas one tradition is primarily concerned with decisions under risk, with and without search, the other tradition examines search-based decisions under uncertainty and intractability. In this section we give an overview of both these traditions and their relevant literature selectively highlight some of the central contributions.

3.2.1 Satisficing Under Risk

The first of these research traditions examines satisficing under risk, where all information is available. On the basis of Simon's conception of satisficing above, we make a distinction between models that focus on aspiration levels only and those that model both aspiration levels and search. Although models in the former group lack one of our defining elements, they represent a widespread interpretation of the term.

Models Without Search

We begin with the group of static models, which is rooted in Simon's proposition of a simplified binary value function. However, over time more diverse implementations have emerged; Table 3.2 classifies this diverse set of models into broad categories.

One straightforward means of introducing aspiration levels into the neoclassical framework has been to modify the utility function. In place of a standard Bernoulli function (Bernoulli, 1738/1954) of concave curvature, utility functions are modified to accommodate aspiration levels. Borch (1968) was one of the first to put forth a model of a decision maker seeking to minimize the probability of bankruptcy. A decision maker in this model seeks to have positive wealth and chooses courses of action that jointly maximize the probability of achieving this outcome. Such a maximization is computationally equivalent to maximizing Simon's (1955) step-utility function, which assumes a value of zero for all outcomes below the aspiration level and a value of one for all outcomes at or above the aspiration level.

The problem with models of successful probability maximization lies in their coarseness. Assuming that each alternative exceeding the aspiration level yields the same amount of utility appears somewhat counter-intuitive. One attempt to overcome this issue is presented by Diecidue and van de Ven (2008). Rather than introducing a step utility function, their model maintains and augments a concave utility function: The value of an alternative is described by the sum of the expected utility of the alternative, its probability of success, and its probability of failure. Formally, the value of alternative a is given by

$$V(a) = \sum_{e=1}^{E} p_e \, u(c_e) + \mu P(c^+) - \lambda P(c^-) \tag{3.1}$$

Table 3.2
Models of Satisficing without Search

Approach	Reference	Keyword
Utility Function	Charnes and Cooper (1963)	binary utility
	Borch (1968)	binary utility
	Fishburn (1977)	risk preferences
	Payne et al. (1980, 1981)	risk preferences
	Brown and Sim (2009)	risk preferences
	Brown et al. (2012)	risk preferences
	Diecidue and van de Ven (2008)	augmented concave utility
	Kahneman and Tversky (1979)	utility with reference point
	Köszegi and Rabin (2006)	utility with reference point
Preference Orders	Jamison and Lau (1973)	semi-orders
	Krishnan (1977)	semi-orders
	Lioukas (1984)	semi-orders
	Rubinstein and Salant (2006)	semi-orders
	van Rooij (2010)	semi-orders
	Aleskerov et al. (2007)	interval orders
Choice Rules	Manzini et al. (2013)	lexicographic semi-orders
Strategic Interaction	Wierzbicki (1982)	principal-agent problem
	Haller (1985)	principal-agent problem
	Pazgal (1997)	cooperation
	Oechssler (2002)	cooperation
	Güth et al (Güth, 2010; Güth et al., 2010)	games
	Papi (2012, 2013, 2018)	consumer choice in monopoly

where c_e denotes the payoff conditional on event e that occurs with probability p_e, c^+ and c^- denote the set of payoffs above and below an aspiration level, respectively, and μ and λ are constant behavioral quantities. Here, c^+ and c^- are defined by an aspiration level.

Whereas the models by Diecidue and van der Ven use expected utility theory as a starting point, prospect theory (Kahneman & Tversky, 1979; A. Tversky & Kahneman, 1992) represents a more radical departure from traditional economic theory. Importantly, prospect theory uses a reference point that separates the domain of gains from that of losses. Gains are valued along a conventional concave function, and losses are valued along a convex function that is steeper in slope than the gain function. The reference point has been interpreted as an aspiration level (e.g., Heath, Larrick, & Wu,

1999); exceeding it yields returns in a conventional concave fashion, whereas falling short of it yields disproportionately negative returns.

An alternative approach to modeling satisficing modifies the preference relation underlying the utility function. In this literature, satisficing is most commonly understood in terms of semi-orders (Luce 1956), a class of preference orderings that allow for intransitive indifference relations of the following type: $a_1 \sim a_2; a_2 \sim a_3; a_1 \succ a_3$. That is, an agent is indifferent between a_1 and a_2 and between a_2 and a_3 in paired comparisons, albeit preferring a_1 over a_3. This intransitive relationship is implied by the differential threshold of vision or touch formulated in Weber's law $jnd = \frac{\delta v}{v}$, where jnd is the just-noticeable difference, v the value of the stimuli, and δv the change in the stimuli. Similarly, Luce (1956) contends that such indifference relations occur when agents are only able to distinguish alternatives that are sufficiently distinct and shows that semi-orders are consistent with maximizing a generalized form of utility function.

van Rooij (2010) argues that satisficing gives rise to a preference semi-order, resulting from agents' inability to distinguish alternatives above the aspiration level. Like a binary utility function, this interpretation of satisficing emphasizes the perceived equivalence of different alternatives. However, as A. Tversky (1969) points out, a semi-order can result from a lexicographic choice rule. By this choice rule, attributes are examined in a fixed order: Initially, alternatives are ranked according to the first attribute; only if they are too similar is the second attribute considered. Similarity is usually assessed using a threshold that can be interpreted as an aspiration level. For instance, Manzini and Mariotti (2007, 2012) develop axiomatic characterizations of choice data that are consistent with the use of lexicographic choice rules. Using this framework, Manzini et al. (2013) characterize a specific lexicographic procedure in which a satisficing strategy is applied at the first stage, followed by a maximization procedure on the selected subset if no unique solution is found beforehand.

Overall, static models of satisficing under risk use expected utility theory as a starting point and modify it to incorporate an aspiration level and make the theory more consistent with observed behavior. Their similarity to expected utility theory enables satisficing to be contrasted with utility maximization, where satisficing is often understood as perceived equivalence of two alternatives that objectively differ in quality. In this modeling approach, satisficing is considered a deviation from rational choice.

Models of Search

The second branch of satisficing under risk is characterized by the use of aspiration levels in the context of search. Here, an agent is initially unaware of the full set of available alternatives and needs to explore them sequentially. Depending on the search problem, the agent can either recall alternatives discovered earlier or only select the alternative that was last examined. Even if earlier alternatives can be chosen, search may be costly, resulting in a trade-off between investing in further exploration and exploiting current knowledge. Given limited resources such as time, it can be advantageous to terminate search before exploring all available alternatives (Caplin & Dean, 2015; Gabaix, 2014; Reis, 2006; Sims, 2003).

Risky search implies that the agent may be unaware of the available alternatives but has meta-information, e.g., regarding their distribution or the cost of search. The first model of search under risk was developed by Simon. In the appendix of his 1955 paper, firmly within the rational choice tradition, he develops an optimal search model that relies on an aspiration level for selling a house. Each day the agent receives a price offer from a known distribution. The agent sets the reservation price, or aspiration level, such that it maximizes the expected value. Following this example, Stigler (1961) popularized the topic of search in economics, emphasizing optimal search. At the same time, the topic of search rose to prominence in statistics, with the theory of optimal stopping (DeGroot, 1970). A common finding in this literature is that optimally behaving agents should continue search until finding an alternative that meets a fixed utility threshold, or aspiration level, similar to Wald's (1948) approach to statistical inference. During the 1970s, this sequential paradigm was also adopted by economists who then used reservation prices as aspiration levels to characterize optimal search (e.g., Rothschild, 1974; Telser, 1973). Since then, models of optimal search under a range of assumptions made aspiration-based stopping rules a tradition in economic and statistical theory staying within the tradition of expected utility theory (e.g., Gilboa & Schmeidler, 1995, 2001; Rubinstein & Salant, 2006).

In search problems without recall, agents can observe the value of an alternative directly but alternatives are only available sequentially. A classic problem in this literature is the secretary problem, where the goal is to choose the best alternative from a random sequence, of

which only the most recently seen alternative is available for choice. As Gilbert and Mosteller (1966) demonstrate analytically, the optimal strategy is a satisficing strategy. When each alternative is described solely by its rank within the observed sample, the chances of choosing the best alternative are maximized when the agent examines the first percent of the sequence, uses the best alternative encountered so far as an (implicit) aspiration level, and then selects the first alternative exceeding this aspiration level. Gilbert and Mosteller (1966) describe a computationally more intensive satisficing strategy for maximizing the probability of choosing the best alternative. Abstracting from the classical secretary problem, Dudey and Todd (2001) introduce additional goals beyond the probability of finding the best alternative, such as maximizing the expected value of the chosen secretary. According to their results, achieving these goals requires shorter search than would be necessary to maximize the probability of finding the best alternative. This divergence in goals, they argue, may explain the finding that experimental participants search less than necessary to find the best alternative. The search models presented here are derived deductively. They are obtained from the properties of the decision problem by determining the optimal decision strategy that achieves a given goal. Notably, these optimal responses often take the form of a satisficing model.

When studying whether people use an aspiration level, the empirical challenge is that it is not sufficient to rely merely on observed outcomes, as M. Friedman (1953) postulates. This challenge provides the motivation for studying the decision process in terms of (i) the search process identifying the information sequentially inspected by the agent, (ii) the stopping rule specifying the aspiration level that terminates search, (iii) and the decision strategy specifying how the agent derives the decision from the information inspected (Gigerenzer et al., 2011; Handel & Schwartzstein, 2018).

Caplin, Dean, and Martin (2011) are among the first to empirically demonstrate the use of aspiration levels akin to Simon (1955) in an incentivized experiment. Participants need to infer the values of the alternatives based on attributes represented by positive or negative numerical values, facilitating commensurability, and to indicate their preferred alternative at any given moment. The authors show that a satisficing model best describes behavior: Participants switch from lower- to higher-value alternatives, indicating that information is being absorbed on an item-by-item basis. Search stops when

participants encounter an alternative that exceeds their aspiration level.

The empirical evidence consistently shows that participants do not choose the best alternative that would be possible in the event of omniscience. However, if people have the opportunity to learn through experience, they are able to approximate an optimal stopping rule (Goldstein, McAfee, Suri, & Wright, 2020; Hey, Permana, & Rochanahastin, 2017; Manski, 2017) which includes the response time when searching internally (Navarro-Martinez, Loomes, Isoni, Butler, & Alaoui, 2018). Taking the sequential nature of search into account, participants generally choose the best alternative among those observed (Bearden & Connolly, 2007; Caplin & Dean, 2011; Caplin et al., 2011; Reutskaja, Nagel, Camerer, & Rangel, 2011), thereby meeting the requirement of rational choice theory for sequential search (Caplin & Dean, 2011). That is, people do indeed choose the "best so far".

3.2.2 Satisficing under Uncertainty and Intractability

As noted in the previous section, Simon (1955) was the first to develop an optimal search model that uses an aspiration level. Yet he suggests that this is inadequate in many settings:

> "It is interesting to observe what additional information the seller needs in order to determine the rational acceptance price, over and above the information he needs once the acceptance price is set. He needs, in fact, virtually complete information as to the probability distribution of offers for all relevant subsequent time periods. Now the seller who does not have this information, and who will be satisfied with a more humbling kind of rationality, will make approximations to avoid using the information he doesn't have." (Simon, 1955, p.117)

One way to address such a situation is by applying heuristics. Since the 1970s, the term heuristics has acquired a negative connotation in economics, psychology, and management, referring to the shortcomings of human reasoning (A. Tversky & Kahneman, 1974). In computer science, however, it is used in in line with its original Greek meaning "to find out or discover" to describe comparatively simple algorithms for making intelligent inferences with incomplete

information in situations of uncertainty or intractability[4]. We follow this tradition and use the term to describe simple decision processes that use limited information.

Heuristics are typically derived from observation of expert decision making in natural environments that are often fraught with uncertainty. To assess the performance of strategies under uncertainty, one cannot rely on the axioms of rational choice theory, which apply solely to situations of risk. Instead, one can assess quality by comparing performance among a set of strategies, for example, that of rational choice strategies with satisficing heuristics. In these comparisons, heuristics often perform surprisingly well or even outperform highly complex strategies, vindicating their use by experts. Determining the conditions under which strategies work well under uncertainty forms the subject of the study of ecological rationality. In other words, heuristics provide the answer to the question of how decisions are made, while ecological rationality is the answer to the question of why a given strategy works well.

Heuristics are often specified in terms of the three elements highlighted before: a search rule, a stopping rule, and a decision rule (Gigerenzer, 1996b). Using this taxonomy, we can characterize satisficing models more precisely as decision models that (i) search through alternatives or attributes and (ii) use an aspiration level in their stopping rules. Learning and adaptation can lead agents to rely on specific classes of strategies tailored to classes of decision problems. The resulting assemblage of strategies represents an "adaptive toolbox" (Gigerenzer & Selten, 2001). We will review several classes of decision problems along with classes of satisficing heuristics (see Table 3.3 for an overview; heuristics are ordered by their appearance in the text).

Aspiration-level adaptation

V. L. Smith (1962) observed a paradox: Markets quickly converge to equilibrium even though agents operate under information conditions that are much weaker than specified by the theory that characterizes the aggregate market. But what are the decision strategies that agents actually use in such a context to solve the problem? Addressing

[4]The first textbook on heuristics in computer science was written by Pearl (1984), who, like Simon, received the Turing Award, the highest honor in the field of computer science and often compared to the Nobel Prize; see Lucci and Kopec (2015) for an up-to-date treatment.

Table 3.3
Classes of Satisficing Heuristics

Aspiration Level	Search Rule	Stopping Rule	Decision Rule
	Aspiration-Level Adaptation		
A_1: minimum value on single attribute (e.g., profit or price)	by alternatives a_i	$a_{i1} > A_1$; if after time t search could not be stopped, lower A_1 by Δ	Choose a_i
	Multi-attribute aspiration-level adaptation		
A_j: minimum value on multiple attributes	by alternatives a_i	$a_{ij} > A_j$ for all x_j; if after time t search could not be stopped, lower A_j by Δ_j	Choose a_i
	Hiatus (recency)		
A_t: time passed since last event	by alternatives a_i	$a_{it} > A_t$	Classify a_i into one of two categories
	Tallying		
A_j: minimum values on multiple attributes	by attributes x_j	$a_{ij} > A_j$ for k out of n	Classify a_i into one of two categories
	Fast-and-frugal Trees		
A_j: minimum values on ordered attributes	by attributes x_j	$a_{ij} > A_j$ for first j that permits exiting the tree	Classify a_i into one of two categories
	Take-the best / Δ-inference		
A_j: minimum value of difference in x_j between a_1 and a_2	by attributes x_j in order of validity	first attribute for which $\lvert a_{1j} - a_{2j} \rvert > A_j$	Choose alternative with higher value on x_j
	Elimination-by-aspects		
A_j: minimum level on attributes	by attributes x_j in random order	for each x_j, eliminate all a_i with $a_{ij} < A_j$, until only one a_i remains	Choose remaining a_i

Notes: a_i denotes alternative $1, 2, ..., i, ..., N$, x_j denotes attribute $1, 2, ..., j, ..., n$, a_{ij} denotes the value of alternative a_i on attribute j, A_j denotes the aspiration level for x_j; for ease of presentation, we assume that the aspiration level is a minimum rather than a maximum value and is lowered when adapted.

this puzzle, Artinger and Gigerenzer (2016) conduct an analysis investigating both constructivist and ecological rationality by studying pricing in the online used car market, which is characterized by a large degree of uncertainty. They find that the market is well fitted to the aggregate data by an equilibrium model by Varian (1980) that captures both observed price dispersion and average price in the tradition of a constructivist analysis. Unlike the equilibrium model, the aspiration-level adaptation heuristic, originally proposed by Simon (1955), correctly predicts the actual dynamic setting of the price. This heuristic is virtually used by all dealers, who initially start with a high aspired price A_1 and lower it at fixed time intervals t, usually by a predetermined margin δ, until a car sells. The aspiration-level adaptation heuristic systematically captures observed phenomena such as high initial price, price stickiness, and the "cheap twin paradox" whereby two highly similar cars at the same dealership have a different price tag due to the simple fact that the price of the car that has been on offer for longer has been reduced after a fixed time interval.

Artinger and Gigerenzer (2016) show that the parameters the dealers use — the initial price A_1, the duration t that the price is held constant, and price reduction δ — vary systematically with the local market conditions, an indication of the ecological rationality of aspiration-level adaptation pricing. Specifically, the higher the population density in the local market and number of dealerships, the shorter the duration t that the price is held constant. In a more densely populated area with more competition, a dealer can more quickly infer that a car is unlikely to sell for a given price, whereas in less densely populated areas with less competition the price needs to be held constant for a longer time.

Pricing offers an illustration of a simple heuristic that formulates an aspiration level on one attribute, the price, and adapts it if necessary. Camerer et al. (1997) hypothesize that taxi drivers terminate their shifts after earning a daily income target. However, if there is an increase of demand on a given day, and taxi drivers could predict it, this would imply that drivers stop their shift too early. Subsequent research has tested this hypothesis by comparing the descriptive powers of neoclassical and reference-dependent versions of utility theory (e.g., Crawford & Meng, 2011; Farber, 2008, 2015). However, Artinger, Gigerenzer, and Jacobs (2020a) find that hourly earnings per driver are barely predictable and therefore hypothesize that drivers

use some form of heuristic rather than strategies that rely on rational expectations, such as utility theory. To test their hypothesis, the authors compare two utility models and four satisficing heuristics in predicting taxi drivers' shift ends. They find that the behavior of the vast majority of drivers is best predicted by satisficing heuristics that terminate shifts after working for a fixed number of hours or after earning a fixed amount of income. The authors speculate that the choice of the aspiration variable reflects which attribute drivers prioritize, whereas the aspiration level is chosen based on experience such that a reasonable balance of income and leisure is reached. Both heuristic models use fixed aspiration levels that are not adapted over time. Similarly, Berg (2014) reports that developers of high-rise office building and malls decide in favor of an investment if they can get at least "X% return" in two or three years, that is, a return that exceeds an aspiration level A for a time t.

Multi-attribute aspiration-level adaptation

In other situations such as academic job search, several attributes are relevant, for instance, prestige of the institution, salary, location, quality of local schools, and spouses' and family preferences. In addition, some of these attributes cannot be traded for others but may be perceived as incommensurable. For these situations, Selten (1998; Sauermann and Selten, 1962) provides a solution that closely follows Simon's (1955) aspiration-level adaptation heuristic (Table 3.3). Agents have an aspiration grid for a set of n incommensurable attributes. After each alternative a_i is examined, aspiration levels A_j are adapted for some attributes j according to a ranking that is affected by preferences and may change as search progresses. When the aspiration for a specific attribute is not met by the examined alternative, it is lowered; otherwise it is increased. When no further increases are feasible, the alternative that meets all aspiration levels is chosen.

In an experimental set-up, Stüttgen, Boatwright, and Monroe (2012) use an eye-tracking device that enables monitoring the search, stop, and decision process for choices among brands of instant noodles with which participants are familiar. Such a naturalistic task makes it possible to investigate a situation with multiple, potentially incommensurable attributes. They test the predictive accuracy of two different models: one based on expected utility and the other on

Simon's (1955) aspiration-level heuristic, albeit without adaptation, where an agent forms aspiration levels for each attribute separately. The satisficing model predicts the observed data much more closely than the utility model does, suggesting that even in such a mundane task, incommensurability is at work. In particular, when an alternative is found that meets all aspiration levels, search is concluded after a verification phase in which the agents acquire additional information.

Hiatus heuristic

Aspiration levels can also be used for predicting whether an event observed in the past will occur again in the future. A case in point is that marketing practitioners often rely on the hiatus heuristic, predicting that customers will make further purchases if they have made a purchase within the past t days, otherwise not (Table 3.3). Here, the aspiration level At refers to time passed. This practice contrasts sharply with the Pareto-NBD model (Schmittlein, Morrison, & Colombo, 1987), a stochastic model that also predicts future purchases. Using purchasing data from three industries, Wübben and von Wangenheim (2008) set out to demonstrate the superiority of the stochastic model but find that the hiatus heuristic yielded the same or better out-of-sample predictions than did the Pareto-NBD model (see also Persson & Ryals, 2014).

Theirs is not an isolated finding. An exclusive reliance on recency, that is, the time since the last event occurred, which ignores all other variables, has long been considered irrational yet has been observed in many different domains (e.g., Gallagher, 2014; Kunreuther, 1976; Malmendier & Nagel, 2011; Slovic, Kunreuther, & White, 1974). Using 60 different data sets across many different domains, Artinger, Kozodi, Wangenheim, and Gigerenzer (2018) show that the hiatus heuristic is the single best strategy for predicting future events such as purchases, outperforming logistic regressions and highly sophisticated machine learning methods such as random forests that make use of recency, frequency, and any other available information.

Tallying

Predicting the next president of the US is a problem entailing high uncertainty about voters' behavior. In November 2016, when prediction markets, polls, and big data analytics predicted Hillary Clinton's

victory by a large margin, Lichtman (2020) instead predicted that Donald Trump would win. Using a tallying heuristic, Lichtman's model has correctly predicted all presidential elections since 1984. The heuristic considers 13 attributes that Lichtman calls "keys", which comprise yes-no questions such as whether the incumbent party holds more seats in the U.S. House of Representatives after midterm elections than it did after the previous midterm election, whether the incumbent-party candidate is the sitting president, and whether real annual per capita economic growth during the term equals or exceeds mean growth during the two previous terms. The tallying rule is:

> Keys to the White House: If six or more attributes are negative, then the challenger will win.

Note that there is no attempt to estimate the weights of the attributes or their covariances; all are given equal weight and simply counted. Unlike in aspiration-level adaptation, where the attributes are given and search takes place by evaluating alternatives, here the alternatives are given and search takes place by evaluating attributes. Search is then stopped when 6 out of 13 attributes are negative. As an aside, 12 of the 13 attributes concern the party holding the White House and its candidate, and only one is about the challenger. This logic implies that voters did not vote for Trump; they instead voted against the previous governing party and its candidate.

Åstebro and Elhedhli (2006) report evidence that a tallying heuristic for classifying early stage ventures performs at least as well as computationally intensive models, while being faster and requiring less information. The heuristic first tallies the positive and negative attributes of a specific venture. If the tally of the positive attributes exceeds the aspiration level and the tally of the negative attributes falls short of the aspiration level, a venture is classified as promising. Testing the strategy competitively against a linear model, the authors found that the satisficing strategy reached a predictive accuracy of 83%, compared with 79% for the linear model. Similarly, Jung, Concannon, Shroff, Goel, and Goldstein (2020) show that a system of bail decisions based on tallying leads to recommendations that are as accurate as those of complex machine learning systems, but less costly and more transparent.

In general, consider a choice between alternatives a_1 and a_2, and n binary or continuous attributes x_j with aspirations levels A_j. The prediction or choice is made by a simple rule:

> Tallying: Choose a_1 if $a_{1,j} \geq A_j$, for at least k out of n attributes. Otherwise choose a_2.

Note that aspiration-level-adaptation requires that all A_j are met, whereas tallying requires that only some k with $(k \leq n)$ of A_j are met. That makes it possible to consider a larger number of attributes and determine whether a critical subset is met. Tallying is compensatory, treating the attributes equally and exchangeable. In contrast, the next class of heuristics treats these in a non-compensatory way: if the first aspiration level is met, search is stopped and a decision is made.

Fast-and-frugal trees

In classification problems, an agent needs to assign an alternative to one of several classes according to its attributes. One class of heuristics that addresses these problems are fast-and-frugal trees (Martignon, Katsikopoulos, & Woike, 2008). Fast-and-frugal trees order attributes sequentially and define aspiration levels for each of them. Unlike other classification trees, however, fast-and-frugal trees can reach a decision after each attribute is examined. That is, for attributes that are split at their aspiration levels, they have $n + 1$ exits for n attributes, of which the first $n - 1$ attributes have one exit each and the final one has two. Thus, a fast-and-frugal tree has $n + 1$ exits, while a full tree has $2n$ exits. These two features, order and aspiration levels, prevent estimation error or even intractability with large numbers of attributes.

For illustration, consider Figure 3.2, which displays a decision tree by Aikman et al. (2020) developed for the Bank of England to identify banks at default risk. This tree uses $n = 3$ attributes, leading to $n+1 = 4$ exits overall. The first attribute that is examined is the bank's leverage ratio. If this ratio falls short of the aspiration level of 4.1%, the bank is immediately classified as vulnerable, without inspecting subsequent attributes. Only when the leverage ratio is at least 4.1% is the next attribute examined, the market-based capital ratio, which can also lead to an immediate decision. The sequential nature of the fast-and-frugal tree models a form of incommensurability: If a bank has an extremely poor leverage ratio, an excellent loan-to-deposit

ratio cannot compensate for that. This is analogous to many other real-life systems: A healthy liver cannot compensate for a failing heart. The exit structure of the fast-and-frugal tree reflects the cost structure of misclassifications (Luan, Schooler, & Gigerenzer, 2011). The bank classification tree has a "red flag" exit after each attribute and thereby minimizes misses at the cost of false alarms compared to the three other possible trees with the same order of attributes but different exit structures. In principle, there are two means of parameterizing the heuristic: by estimating the parameters statistically or by a combined method in which an expert determines the attributes, their order, and the exits and then determines the thresholds for each variable using statistical methods. Aikman et al. (2020) show that the statistically estimated fast-and-frugal tree is better than a logit model at predicting vulnerable banks and that the combined method outperformed both. The construction of fast-and-frugal trees, as well as of tallying models, and their predictive accuracy relative to regression and machine learning models is explained by Katsikopoulos, Şimşek, Buckmann, and Gigerenzer (2020).

Take-the-best, δ-inference, and elimination-by-aspects

Consider a choice problem where the agent needs to select one alternative from a set based on the attributes of each alternative but does not know the overall value or utility of each alternative. For example, an agent wants to identify the house with the highest quality based on attributes such as price, floor size, location, or number of rooms. Here, the agent knows the available options but does not know the outcome (quality) and needs to infer it from the given attributes, where the relation between attributes and quality is not or is only partially known. This meets the definition of uncertainty with regard to outcomes; the alternatives are known. When choice is between two alternatives, the take-the-best heuristic can be used (Gigerenzer, 1996b). This strategy examines attributes lexicographically, that is, they are ordered by their validity, defined as the percentage of correct inferences made by that attribute alone. If the two options have identical values for the attribute, it is ignored and the next attribute is examined. Search through attributes stops as soon as one is found that discriminates, that is, when the difference in value exceeds the aspiration level of zero. Once such an attribute is found, no further attributes are examined, and the decision is based solely on the one

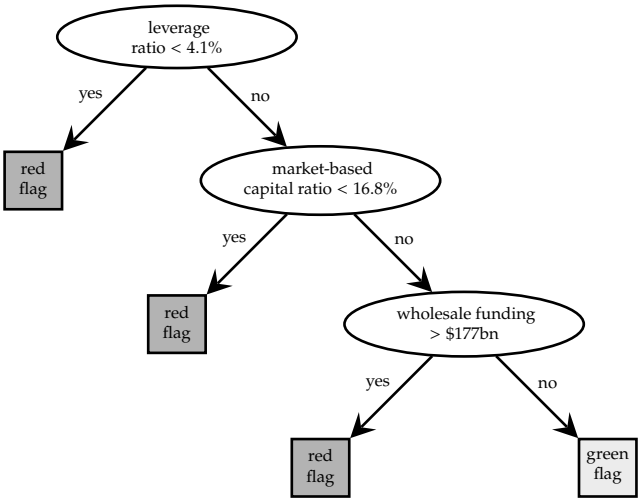

Figure 3.2: *Fast-and-frugal tree for bank classification.*

discriminating attribute. Whereas take-the-best is limited to decisions with binary attributes, δ-inference (Luan et al., 2014) can be used when attributes are continuous.

Testing the performance of take-the-best and δ-inference shows that these heuristics are surprisingly powerful in prediction. Across 20 different data sets, ranging from school dropout rates to property prices, take-the-best was less accurate than multiple regression in choice from seen alternatives (data fitting) but more accurate in choice from unseen alternatives, that is, in out-of-sample testing (Czerlinski, Gigerenzer, & Goldstein, 1999). Similarly, comparing take-the-best with machine learning models such as classification-and-regression trees (CART) and support vector machines shows that take-the-best can match or even outperform these in an out-of-sample setting, while using less information (Brighton & Gigerenzer, 2008, 2012). Luan et al. (2014) and Luan, Reb, and Gigerenzer (2019) found that δ-inference yields better out-of-sample performance than linear regression and machine learning models such as random forest across 20 additional

datasets, irrespective of whether the aspiration level was set optimally (based on past data) or to its most robust value.

When alternatives abound, direct comparisons may no longer be feasible. With more than three alternatives available, the number of possible direct comparisons exceeds the number of alternatives. In these cases, efficiency would not require agents to compare alternatives with one another but instead to assess them individually. This is done, for instance, by the elimination-by-aspects heuristic, which forms aspiration levels for each attribute and examines them in the order of their importance (A. Tversky, 1972). For each attribute, it eliminates all alternatives that do not meet its aspiration level until a single alternative remains, which is then chosen. The structure of elimination-by-aspects resembles the class of consideration-set heuristics that have been proposed as a strategy for dealing with large sets of alternatives in research on consumer decisions: Rather than examining all available alternatives in detail, consumers heuristically exclude options from detailed analysis. The resulting consideration set is then submitted to detailed examination at a second stage (e.g., Hauser, Toubia, Evgeniou, Befurt, & Dzyabura, 2010; Hauser & Wernerfelt, 1990; Marewski, Gaissmaier, Schooler, Goldstein, & Gigerenzer, 2010). Hauser (2014) reviews various heuristics that have been proposed for the formation of consideration sets, including satisficing approaches. These heuristic models are highly predictive of how consumers form small consideration sets (Dzyabura & Hauser, 2011; Yee, Dahan, Hauser, & Orlin, 2007). Consideration sets bear a resemblance to the choice rules discussed earlier (Manzini & Mariotti, 2007; Manzini et al., 2013), with the difference that the work on choice rules makes specific assumptions about preference orders, whereas consideration set heuristics do not assume a specific preference ordering. This difference reflects diverging assumptions about the decision environment: The notion of preference orders assumes a situation of risk where alternatives can be at least partially ranked because all necessary information is available to the agent. In contrast, consideration set heuristics have been devised for situations where the set of alternatives is simply so large that one cannot meaningfully deduce an order.

Note that all heuristics listed in Table 3.3 can effectively deal with incommensurability between attributes and do not need to integrate them onto a single cardinal scale as in expected utility theory. Search proceeds on an attribute-by-attribute basis where the value of an

attribute is evaluated with respect to an aspiration level. This is also the case for the aspiration-level heuristic, which searches by alternatives and evaluates these attribute by attribute. Also note that not all heuristics that deal with decisions under uncertainty are satisficing heuristics. For instance, $\frac{1}{N}$, which allocates an investment equally over N assets, relies neither on an aspiration level nor on search. Its rationale is to reduce error from estimating asset weights and the covariance matrix (DeMiguel, Garlappi, & Uppal, 2009).

Heuristics for well-defined but intractable problems

The above examples address uncertainty. However, even if the complete set of alternatives, possible future states of the world, and consequences is known, intractability surprisingly quickly makes rational choice infeasible and a heuristic strategy becomes an effective solution. Gabaix, Laibson, Moloche, and Weinberg (2006) show this by comparing two different conditions that resemble such a mundane task as selecting a television characterized by a few attributes. The first condition is simple: Facing three alternatives with only one attribute each and a stochastic payoff, the decision maker sequentially explores the attributes and the values that realize, and in turn the value of an alternative. In this condition, information acquisition is costly, and the agent can stop acquiring information at any time. The Gittins-Weitzman index solves the problem optimally by establishing a complete sequence with which to explore the products and their attributes (Gittins, 1979; Weitzman, 1979). The more complex condition is characterized not by three but by eight alternatives with not one but nine attributes. In the experiment, the attributes are positive or negative numerical values, which facilitates commensurability and which participants get to know during a sequential search process. Gabaix et al. (2006) report that the complex task is computationally intractable for the rational choice model because the problem suffers from the curse of dimensionality (see also González-Valdés & de Dios Ortuzar, 2018; Salant, 2011). However, they show that participants in a laboratory experiment employed the directed cognition heuristic, a simple, myopic strategy that looks just one step ahead instead of solving the complete sequence. It can solve not only the simple problem but also the complex problem that rational choice cannot address. The heuristic proceeds as follows: First, it compares the value of stopping search immediately and the expected value

of stopping immediately after the next attribute. This is a myopic calculation because it does not incorporate the possible consequences of continuing search beyond the realization of the next attribute. Second, the heuristic inspects the attribute with the highest expected value. Third, the first two steps are repeatedly executed until the costs of searching and inspecting another attribute outweigh the expected value of the next most attractive attribute, which represents the aspiration level.

Myopic search is a general tool for finding good solutions to intractable problems. Examples are scheduling problems in transportation, where the task is to find the shortest route through cities, beginning and ending at the same city. For 50 cities, finding the shortest route would require searching through more than 1062 possible routes. Heuristics such as the nearest neighbor algorithm — move to the nearest unvisited city — can provide excellent solutions where the best one cannot be found (Lucci & Kopec, 2015).

Behavioral Theory of the Firm

The use of an aspiration level as a heuristic decision strategy was already at the core of Simon's dissertation, published in 1947 under the title "Administrative Behavior". Here, Simon was primarily concerned with firms; a more general decision context is found in his seminal paper of 1955. Nonetheless, it was "for his pioneering research into the decision making process within economic organizations" that Simon was awarded the 1978 Nobel Memorial Prize in Economic Sciences. His work was developed further (March & Simon, 1958) and ultimately inspired the "behavioral theory of the firm" (Cyert & March, 1963). In its analysis of the fundamental decisions of the firm, such as price, output, and resource allocations, it lays "an explicit emphasis on the actual process of decision making as its basic research commitment" (Cyert & March, 1963, p.19). This stands in sharp contrast to traditional economic theory that focuses on market level outcomes and classically models firms as rational actors.

The theoretical foundations on which behavioral theory of the firm builds in order to understand the actual decision processes are a) satisficing instead of maximization, where the first alternative that is satisfactory with respect to an aspiration level is chosen; b) search for information when not all possible outcomes of any choice

alternative can be anticipated; c) the use of robust rules in the face of uncertainty, circumventing predictions about the distant future. Its quest is to align models as closely as possible with the empirical observations of both the output that organizations produce and the process they use (for a review, see Gavetti, Greve, Levinthal, & Ocasio, 2012). The theory has risen to prominence particularly in the domains of organization and strategy (March, 1981, 1991; Winter, 2000).

The behavioral theory of the firm predicts that an aspiration level is a function of recent performance, past historical aspiration levels, and recent performance of other firms. To understand organizational learning processes, models of aspiration formation have been developed that are clearly rooted in the tradition of bounded rationality. In economics, models of this form have a long history as adaptive expectations (e.g., Chow, 1989; Sterman, 1987) or adaptive learning (e.g., Jacobs & Jones, 1980). The main difference is that in expectation models, aspiration levels are not explicitly modeled but subsumed in an overall function. An ecology of learning organizations and competition yields a continuous adaptation process. Search is initially local; if no solution is found that yields satisfactory performance, a broader search ensues (Levinthal & March, 1981). If this does not yield a satisfactory outcome, the aspiration level is adapted.

The behavioral theory of the firm has also inspired research in economics, giving rise to the field of evolutionary economics (Nelson & Winter, 1973, 1982), which examines industrial evolution (e.g., Dosi, Nelson, & Winter, 2000). Nelson and Winter (1982) model the evolution of an industry with firms using particular rules or routines in situations where the world is characterized by complexity and uncertainty. Routines emerge as a response to an adaptive process where no optimal solution can be found ex-ante. A firm uses a given routine as long as its output remains above an aspiration level; only when the output falls below this does it engage in exploration for an alternative routine (Winter, 1971). Such behavior necessarily results in companies failing to survive, as shown for instance by Witt (1986). Comparing three algorithms in a multi-period market competition, where the first maximizes expected profits, the second uses an aspiration level and satisfices when setting prices, and the third algorithm is based on simple reinforcement learning, Witt shows that the survival of an algorithm strongly depends on the initial conditions and that optimization does not dominate the other algorithms. Given uncertainty, Heiner (1983, 1989) formally shows

that that firms will adapt only a limited number of simple decision rules. Rules are added solely if they exceed an aspiration level, which refers to the reliability with which the rule will generate profitable future actions. The larger the degree of uncertainty, the fewer the rules.

The rich body of empirical research on firms operating under uncertainty shows that they rely on various forms of satisficing (Artinger, Petersen, Gigerenzer, & Weibler, 2015). Much of this research concerns the actions of firms and the evaluation of performance with regards to an aspiration level. Aspirations determine whether past performance is framed as a success or failure, which influences subsequent strategic decisions (Lant, 1992). The first to provide empirical evidence for the use of an aspiration level in the formation of organizational goals is Lant (1992). In a laboratory experiment, she uses a management game employed by companies for inhouse training that captures the complexity and dynamics that managers frequently face. Teams of MBA students and managers from an executive program compete over multiple rounds of producing and selling two types of consumer products. After teams set their own goals, they make strategic and resource allocation decisions on a range of different variables. The underlying software uses a complex nonlinear algorithm that simulates a competitive market in a multidimensional, interdependent world. Lant (1992) observes the goals, or aspiration levels, that the teams set in terms of their targeted sales as well as their actual performance. Her findings indicate that the teams are best described as satisficing, whereas the rational expectation model receives relatively little support (see also Audia & Greve, 2006; Lant & Shapira, 2008). The results by Lant (1992) have been replicated in a study by Mezias, Chen, and Murphy (2002), who use field rather than laboratory data of decision makers in a financial services company. Following Lant (1992), considerable evidence has accumulated on the use of aspiration levels in evaluating and guiding firm performance in such diverse contexts as R&D expenditures, outsourcing, and firm growth (e.g., Audia, Locke, & Smith, 2000; Baum, Rowley, Shipilov, & Chuang, 2005; Berg, 2014; Bolton, 1993; de Boer, Gaytan, & Arroyo, 2006; Greve, 1998, 2008; D. Miller & Chen, 1994). Blettner, He, Hu, and Bettis (2015) study the process of aspiration adaptation in a longitudinal data set from the news magazine industry. They find that when constructing an aspiration level, organizations largely focus on their own past performance and whether they achieve the goals they

set for themselves. When they are close to bankruptcy, however, they start focusing on competitors' performance.

In summary, considering the search process is an important element in evaluating the quality of individuals' and firms' decisions. Given search, empirical evidence shows that agents do choose the alternative that is "the best so far". Given alternatives with two or more potentially incommensurable attributes, heuristics dispense with the need to integrate these by applying aspiration levels. Heuristics can be powerful tools to make good decisions given uncertainty or intractability. But what exactly are the conditions under which heuristics perform well?

3.3 A Framework for Integration

As we have argued in the previous section, satisficing strategies are investigated within two distinct research traditions. These traditions differ in the assumed decision environment (risk versus uncertainty) and arrive at different conclusions about the rationality of satisficing: Although models of satisficing under risk without search conclude that satisficing is not commendable because it produces systematic mistakes, satisficing strategies given uncertainty or intractability are often found to yield better decisions than those by alternative models. There are two possible explanations for such divergent findings that allow for integration of these two different traditions: (i) cost-benefit trade-off in information acquisition and in information processing, and (ii) the bias-variance trade-off, which provides a framework to understand what type of strategy is best given the information available, beyond the costs of search and computation.

3.3.1 Cost-Benefit Trade-off

Operating on the basis of limited information can be a result of the costs of information search. Given such costs, it can be beneficial to stop search early and make a decision based on a limited sample such that people deliberately decide not to know (for a review in an applied context, see Hertwig & Engel, 2016). Since at least the work of von Hayek (1945), economists have studied the trade-off between learning costs and decision quality. Agents in general do indeed choose the better alternatives or, as Simon would put it, the best so far (Bearden & Connolly, 2007; Caplin et al., 2011; Reutskaja et al.,

2011). Costs of information acquisition can also rationalize some apparent mistakes in the field. For instance, Chetty, Looney, and Kroft (2009) find that agents buy unnecessarily expensive products due to search costs. Similarly, costs for information acquisition also explain why agents visit only a relatively small number of websites before an online purchase (Santos, Hortaçsu, & Wildenbeest, 2012). Search costs feature in several models in the form of information cost functions (e.g., Caplin & Dean, 2015; Sims, 2003; Verrecchia, 1982). These models have in common that they allow determining the optimal trade-off between benefits and costs of search. At the same time, however, they make some strong assumptions about agents' knowledge and the degree of complexity of the environment.

Similarly, operating with a simple decision strategy can result from the costs of computation. As shown, for instance, by Gabaix et al. (2006) and Salant (2011) and other work mainly in computer science (for a review, see Lucci & Kopec, 2015), computing an optimal solution can entail substantial costs or even quickly prove to be infeasible. People are inherently sensitive to such costs of computation. Payne, Bettman, and Johnson (1988) show that when the cost of computation is increased by introducing time pressure, agents switch to a simpler strategy. Experimental work concludes that humans tend to trade off benefit and cost of cognitive effort rationally (Kool & Botvinick, 2014; Lieder, Griffiths, & Hsu, 2018; Lieder et al., 2014). This suggests that apparent mistakes can be rationalized in terms of computational costs. Formal models that trade off computational benefits and costs have been developed particularly in computer science and specifically in research on artificial intelligence (for a review, see Gershman, Horvitz, & Tenenbaum, 2015). Such models start with a meta-level analysis of the ideal balance between computational effort and the quality of alternatives. Yet, as Gershman et al. (2015) stress, calculating an optimal solution in terms of costs and benefits of computation is frequently challenging. Few decision problems admit an analytic solution, and many problems are computationally intractable. In addition, Conlisk (1996) points to the problem of infinite regress: If computation is costly, then optimizing computation is costly ad infinitum.

Like search costs, computational costs provide a reason why satisficing and simple strategies that require little information can perform so well, albeit trading a reduction in costs for a reduction in accuracy. At the same time, section 3.2.2 reported a number of

findings where heuristics perform on par or even outperform complex alternative models from, among others, operations research and machine learning, while reducing costs of search and computation at the same time. The performance criterion in many cases was predictive accuracy, without accounting for search and computational costs, which would have boosted the performance advantage of the heuristic even more. This calls for an alternative explanation.

3.3.2 Bias-Variance Trade-off

An alternative reason why heuristics can perform so well concerns the trade-off between bias and variance. This trade-off explains why a model with fewer parameters can yield more accurate predictions than does a nested model with more parameters. The trade-off provides the explanation why a simple model such as a heuristic can outperform a complex model simultaneously in terms of accuracy and other performance metrics.

A model's predictive accuracy is grounded in a basic statistical relation between bias and variance. Suppose the task is to predict y, the value of an unseen item, based on its observables x. Value y was generated by an unknown function $y(x, Q)$, which combines the observables using a fixed but unknown parameter set Q. A model $m(x, q)$ is used to predict the value y. To this end, a learning sample is randomly drawn from the population of items generated by $y(x, Q)$ and is used to calibrate q, the parameter set of the model $m(x, q)$. Based on this parameter set, a prediction about y is made, say $m(x, \hat{q})$. The error in prediction can be assessed in many ways, among which the mean squared error is a common choice (Geman, Bienenstock, & Doursat, 1992). The mean squared error in predicting item can be decomposed as follows:

$$\text{error} = \text{bias}^2 + \text{variance} + \epsilon \tag{3.2}$$

where ϵ denotes the irreducible error that cannot be eliminated even if the generating function $y(x, Q)$ were known. An example of ϵ is unsystematic measurement error. In contrast to ϵ, bias is a systematic error of model $m(x, q)$. To better understand this, suppose there are L possible independent learning samples of a given size, and each is used to fit the parameters of model $m(x, q)$ and calculate a prediction about y, yielding a set of predictions $m(x, \hat{q_1}), m(x, \hat{q_2})...m(x, \hat{q_L})$.

Bias is the difference between the average of these predictions and the expectation of the generating function:

$$\text{bias}^2 = \{E_n[m(\boldsymbol{x}, \hat{\boldsymbol{q}}_i)] - E[y(\boldsymbol{x}, \boldsymbol{Q})]\}^2, \tag{3.3}$$

where $E_n[m(\boldsymbol{x}, \hat{\boldsymbol{q}}_i)]$ denotes the expectation with respect to different learning samples and $E[y(\boldsymbol{x}, \boldsymbol{Q})]$ denotes the expectation with respect to unsystematic error. Error from bias reflects a systematic misspecification of model $m(x, q)$ relative to the generating function $y(x, Q)$. Variance, in contrast, arises from a lack of knowledge of the model's parameter values and the need to infer them from a limited sample of data. Formally, variance refers to the average variation of an individual prediction $m(x, q_i)$ around the average prediction:

$$\text{variance} = E_n[\{m(\boldsymbol{x}, \hat{\boldsymbol{q}}_i) - E_n[m(\boldsymbol{x}, \hat{\boldsymbol{q}}_i)]\}^2], \tag{3.4}$$

Error from variance reflects the sensitivity of the model to idiosyncrasies of the sample used for calibrating its parameters. An illustration is provided in the upper panel of Figure 3.3. In it, the center of the target is the expectation of the unknown generating function, the best possible prediction. The black dots represent different predictions of $m(x, q)$ based on different learning samples that yield different estimates of q, and the gray dot is the average of these predictions. The model on the left has low bias and accurate average predictions but high variance, and its predictions fluctuate strongly with parameter estimates; the model on the right has higher bias but lower variance in predictions.

In general, a model that aims to maximize predictive accuracy needs to control both the bias and the variance components of the prediction error. Bias is reduced to the degree that model $m(x, q)$ resembles the generating function $y(x, Q)$. One means of achieving lower bias is to add free parameters to the model. However, each additional parameter increases (or at best keeps constant) the variance component of the prediction error, provided that these models are nested. Thus, there is a trade-off between bias and variance. The lower panel of Figure 3.3 shows the point of the minimum prediction error for two different sample sizes given a set of free parameters estimated from the data. To the left of the point, including fewer parameters implies higher total error; to the right of it, more parameters imply higher total error. The exact location of the point is determined by factors affecting variance, such as the mathematical nature of the

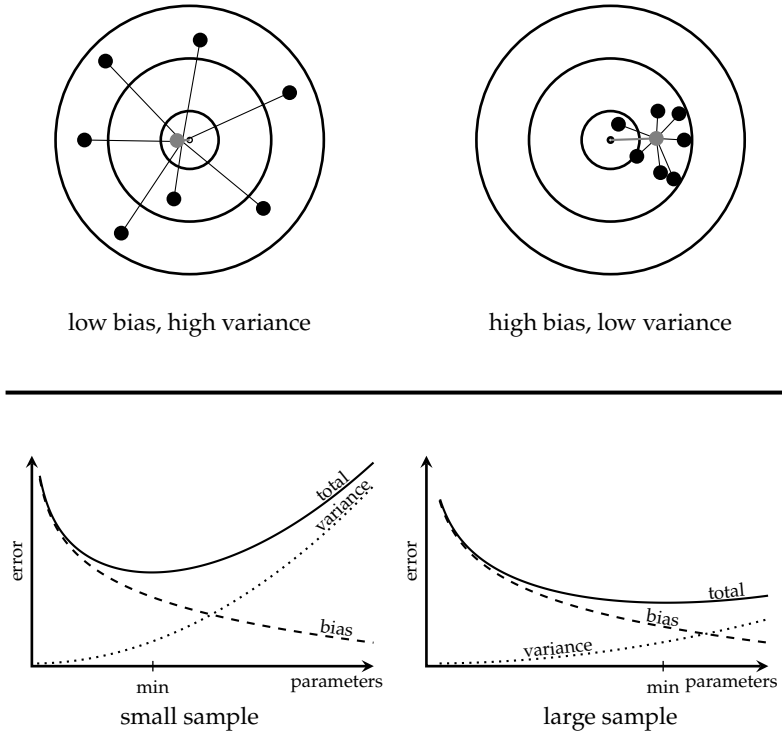

Figure 3.3: *Bias and variance. Upper panel: The center of the target is the expectation of the unknown generating function, the best possible prediction. The black dots represent different predictions of $m(\boldsymbol{x}, \boldsymbol{p})$ based on different learning samples that yield different estimates of $\boldsymbol{p}$. The gray dot is the average of these predictions. The model on the left has a low bias and accurate average predictions but a high variance, and predictions fluctuate strongly with parameter estimates. The model on the right has higher bias but low variance in predictions. Lower panel: Bias diminishes but variance increases in the number of parameters. There exist a number of parameters that minimize total error, the position of which depends on many factors, including the size of the learning sample.*

parameters (additive, multiplicative, etc.) and the size of the sample used to calibrate the parameters. With more data available, the model can "afford" more parameters while keeping its bias and variance in balance.

The trade-off between bias and variance is well known in statistics

and machine learning. However, empirical economics is divided in its view (e.g., Yatchew 1998; Varian 2014; Mullainathan and Spiess 2017). On the one hand, large parts of empirical economics are concerned with accurately estimating model parameters. Here, the goal is to obtain unbiased parameter estimates to evaluate, for instance, policy interventions and only rarely to predict (but see also Kleinberg et al. 2015). Overfitting is usually addressed by using tools such as the Bayesian information criterion or regularization. Yet controlling for overfitting does not seem to be a perfect safeguard. For instance, Artinger et al. (2018) and Luan, Reb, and Gigerenzer (2019) competitively test heuristics against regularized regression and random forests, models designed to be robust to overfitting. They find that the heuristics predict better than the more complex models, particularly in smaller samples. On the other hand, parts of microeconomics and most of behavioral economics focus on comparing theories. This line of work rarely conducts out-of-sample tests of these theories but instead typically compares the in-sample fit (e.g., Starmer 2005; D. Friedman et al. 2014). As the bias-variance trade-off illustrates, this practice can be misleading. Instead, theory comparison should follow two principles set out by Friedman in 1953:

> (i) Theories need to be tested on prediction, not on how well they fit data: "The ultimate goal of a positive science is the development of a 'theory' or 'hypothesis' that yields valid and meaningful (i.e., not truistic) predictions about phenomena not yet observed" (M. Friedman, 1953, p.7).
> (ii) Theories should be tested competitively by comparing their predictions: "The question whether a theory is realistic 'enough' can be settled only by seeing whether it yields predictions that are good enough for the purpose in hand or that are better than predictions from alternative theories" (M. Friedman, 1953, p.41).

Given an uncertain environment, agents — like scholars who seek high accuracy — require strategies that balance bias and variance (Brighton & Gigerenzer, 2015; Gigerenzer & Brighton, 2009). We examine now what this conclusion implies for satisficing in different informational environments.

3.3.3 Satisficing and the Bias-Variance Trade-off

Environments of risk allow for constructivist analysis. Under risk, the only error one can make is due to bias, whereas error due to variance is by definition not possible. With full information about the decision environment, scholars and agents are able to derive the optimal path of action. No inference about the decision environment is required and variance from estimation becomes irrelevant. Consequently, the prediction error, and hence quality of choice, depends solely on the bias of the decision strategy. In these situations, satisficing is suboptimal compared with unbiased strategies such as mathematical optimization.

Environments of uncertainty require a different approach, the analysis of the ecological rationality of strategies (Selten, 2001; V. L. Smith, 2008). In an environment of uncertainty, where the problem is ill-defined and the exhaustive set of states of the world and their consequences is not knowable or foreseeable, the best strategy cannot be foreseen. Thus, the study of ecological rationality under uncertainty compares strategies to determine which one likely performs better in what environment and, importantly, why. To do that, it analyzes both sources of error, bias and variance. The best-performing decision strategy does not reduce bias to zero, but finds a balance between the two kinds of errors. Here is the place for heuristics, whose simplicity can reduce error due to variance, such as by needing fewer free parameters. Given uncertainty, the optimal trade-off between bias and variance is hard to estimate. Instead, it requires two methodological approaches: competitive testing of strategies, and testing on out-of-sample prediction.

Consider first the ecological rationality analysis of bias. So far, we have assumed that lower exposure to variance is paid for with increased bias. However, the amount of bias depends crucially on the statistical structure of the decision environment. When the structure of the heuristic matches the structure of the decision environment, the bias of a heuristic can be surprisingly low. To illustrate, consider the hiatus heuristic defined earlier, which predicts whether a customer will make future purchases or not (Table 3.3). Can we identify conditions under which the bias of this heuristic is the same as that of linear models? Consider a simple example. Assume a linear strategy with n binary attributes $x_1, x_2, ..., x_n$ with values of either $+1$ or -1, where the positive value indicates future purchases. All

of the weights of the attributes are positive and, like beta weights, reflect the additional contribution to the higher ranked attributes. The linear rule infers that the customer will make future purchases if $y > 0$, otherwise not. If the following condition holds, the bias of the hiatus heuristic (or similar one-reason heuristics) is the same as that of a linear model:

> *Dominant attribute condition:* The ordered weights $w_1, w_2, ..., w_n$ form a dominant cue structure if they satisfy the inequality constraint:
>
> $$w_1 > \sum_{k=2}^{n} w_k \tag{3.5}$$

If there is a dominant attribute, the heuristic performs as well as and even better than the linear model because the latter incurs further error from variance (Artinger et al., 2018). Similarly, it can be shown analytically that the take-the-best heuristic has the same bias as a linear model when the weights of the linear model are non-compensatory (Martignon & Hoffrage, 2002).

> *Non-compensatory attributes condition:* Each weight, w_j, exceeds the sum of lesser weights:
>
> $$w_j > \sum_{k=j+1}^{n} w_k \tag{3.6}$$

In these situations, the choice made by a linear decision rule is determined exclusively by the attribute with the highest weight (see also Hogarth & Karelaia, 2007). The take-the-best heuristic exploits such an environmental property by employing a non-compensatory, lexicographic decision strategy. This property leads take-the-best and a linear decision rule to have identical bias in non-compensatory environments. In addition to identical bias, lower variance leads heuristics to outperform linear models in prediction.

How often do these conditions hold in the real world? Şimşek (2013) examines 51 natural environments, ranging from car and house prices to the salaries of college professors. For 93% of the paired comparisons in half of the data sets, she finds that linear models yielded the same decisions as lexicographic models that decided on the basis of the first discriminating attribute.

As to variance, the larger the sample size and the smaller the number of free parameters, the lower the error by variance in general. Heuristic strategies, including satisficing heuristics, tend to have fewer free parameters than do estimation-and-optimization strategies. Heuristics and optimization strategies are rarely nested, which complicates a-priori comparisons of their exposure to variance. The results of competitive performance tests, however, suggest that satisficing heuristics can be less susceptible to differences in training data and offer more robust predictions. It is important to realize that these conclusions are relative to the amount of training data used to compare these strategies. Şimşek and Buckmann (2015) show for 63 data sets that heuristics can be effective when training data is limited but their advantage vanishes as the size of the learning sample increases. The work by DeMiguel et al. (2009) illustrates that the necessary sample size can be very large. For portfolio selection, they compare the $1/N$-heuristic, which allocates an investment equally over N assets, with the mean-variance model (Markowitz 1952) and a range of modern variants (see also Wang, Wu, Yang, Wang, & Wu, 2015). They showed for a training data of 10 years that none of the sophisticated models was able to consistently outperform $1/N$. Indeed, for $N = 25$ assets, the mean-variance model requires about 250 years of stock data to outperform the simple $1/N$-heuristic, and for $N = 50$ assets, 500 years are required, assuming that the same stocks, and the stock market itself, still exist.

The informational requirements of parameter-rich models are an important factor for understanding their performance under uncertainty. The derivation of these models under assumptions of sufficient information often ignores the important role of variance. With remarkable prescience, Simon (1981, p.44) alludes to this issue, writing: "Although uncertainty does not [...] make intelligent choice impossible, it places a premium on robust adaptive procedures instead of strategies that work well only when finely tuned to precisely known environments." In many situations of uncertainty, both agents and scholars cannot obtain an exhaustive representation of the decision environment at hand and therefore cannot determine the optimal course of action. To find the best possible strategy therefore involves a competitive test of different strategies, as already highlighted by M. Friedman (1953).

Experts, such as car dealers or marketing managers, can generate high-performing, simple heuristics. With sufficiently large datasets,

there are also statistical techniques available to generate such simple heuristics. Jung et al. (2020) for instance develop a technique based on regularization that generates transparent and easy to understand heuristics that perform en-par with black-box machine learning models such as random forest. Regularization, for instance using the Stein estimator (James & Stein, 1961), the Lasso estimator (Tibshirani, 1996), or ridge regression (Hoerl & Kennard, 1970), reduces the exposure of the models to error due to variance. Specifically the Stein estimator, which is a biased estimator of the mean, can be shown to dominate the ordinary least squares approach in terms of a strictly better mean squared error.

There is an important caveat. The bias-variance dilemma assumes a stable population from which repeated samples are drawn. An example would be an agent playing a lottery with unknown probabilities who is given the opportunity to sample each alternative before choosing one. Here, agents may be able to estimate the relevant parameters of their decision environment before applying an optimization strategy, such as any form of utility maximization. Yet many situations contain more radical forms of uncertainty. The decision environment may be unstable and leave the agent without reliable learning samples for parameter estimation or the causal structure of the environment may be unknown. The problem may also suffer from computational intractability. Under these circumstances, agents are forced to employ some form of simplification in order to obtain a mathematical representation of the decision problem from which they can optimize. Because optimization strategies are only optimal relative to their assumptions or the sample they were estimated in, there is no guarantee that the decisions they yield are indeed optimal. Simon (1979) succinctly observes in his Nobel Prize speech that "decision makers can satisfice either by finding optimal solutions for a simplified world, or by finding satisfactory solutions for a more realistic world. Neither approach, in general, dominates the other, and both have continued to co-exist in the world of management science". These alternative methodologies correspond to the two traditions of satisficing: optimal solutions assuming a situation of risk, and heuristic solutions assuming an uncertain world.

3.3.4 Integrating Two Research Traditions using the Bias-Variance Trade-off

This article set out to review advances in our understanding of satisficing and identified two research traditions. We refer to these as satisficing under risk and satisficing under uncertainty. Satisficing has been studied under both conditions, albeit in two largely unconnected literatures. Both have their roots in Simon's 1955 article, the first in the appendix, the second in the main text. For both classes of models the motivation is the same, namely to introduce aspiration levels and search in order to reflect what actual decision makers do. Models of satisficing under risk use rational choice theory as a starting point and then proceed to make modifications in line with a satisficing strategy to account for additional behavioral variance. Typically that requires strong assumptions about what decision makers know or can know about the future in order to use the optimization calculus. Moreover, in many of these models, only aspiration levels are considered, search is ignored, and satisficing is viewed as a failure to act rationally. Models of risk with aspiration levels and optimal search need to make additional strong assumptions about what can be known, such as complete information about the relevant probability distributions (e.g., the probability distribution of offers for all future time periods; see Simon, 1955) and the future cost of search in order to calculate an optimal stopping point.

Satisficing under uncertainty refers to situations in which these assumptions are not met or cannot be met. Uncertainty includes situations where the exhaustive and mutually exclusive set of alternatives, or future states, cannot be known, meaning that the relevant probability distributions are also unknown. To model decision making under uncertainty, we have used the bias-variance trade-off, and its key insights are that good models need to have not only low bias but also low variance and that there is a trade-off between the two sources of error. Variance arises in situations where unknown parameters need to be estimated; it does not arise in situations of risk, where the parameters (including probability distributions) are assumed to be known. Heuristics can reduce variance by using few free parameters, or even none (as in the hiatus heuristic with a fixed hiatus or in $1/N$), and thus can lead to more accurate predictions or decisions than more complex models. At the same time, they are fast, transparent, and reduce search costs. In situations of uncertainty, satisficing heuristics

are no longer sub-optimal, they can be the best one can do. But deciding which heuristic to choose in which situation requires careful study of the match between heuristic and environment, that is, by an analysis of their ecological rationality.

Thus, both traditions have their relevance, but in two quite different classes of problems. Against this dualism, one might argue that in situations of uncertainty, one could use the bias-variance dilemma to calculate the optimal trade-off between bias and variance in the same way as when using satisficing under risk, where the optimal stopping point is a trade-off between expected costs and benefits of further search. That would indeed reduce uncertainty to risk. To calculate the bias, however, one would need to know the "true" function that generates the data, which cannot be known in situations of uncertainty. Similarly, one could argue that all situations should be treated as ones of uncertainty. For instance, even in apparently certain conditions, unforeseeable events may happen or rules might be gamed. Yet that attempt towards reduction would be equally mistaken: Models are not unrealistic per se; they can be realistic for one class of problems and not for others. Moreover, both traditions have a different yet complementary approach to rationality, corresponding to Smith's (2008) distinction between constructivist rationality and ecological rationality.

Although Simon's original article planted the seeds for two different research traditions, their diverging methodologies and conclusions are not contradictory. Instead, the review has shown how these differences parsimoniously follow from the classes of decision environments they address. Under risk, where all the relevant information is known, optimization strategies are superior to satisficing strategies and rational choice theory constitutes an obvious starting point. Given intractability when the problem is well-defined, or uncertainty when the problem is ill-defined, no single class of decision strategies is generally superior to another and models are derived from observation rather than function.

3.3.5 A Revised Understanding of Heuristics and Biases

One common view characterizes heuristics solely in terms of their bias (e.g., A. Tversky & Kahneman, 1974). This view does not distinguish between risk and uncertainty and routinely concludes that agents' lack of computational capabilities leads them to make

decisions that are not in their own interest. Generalizing the results of such studies to decision making under uncertainty when the problem is ill-defined is more complicated than commonly assumed. The bias-variance trade-off implies that some degree of bias can be appropriate when the relative scarcity of data leads to potentially unbiased but unreliable parameter estimates. Under such conditions, biased decision strategies can yield better decisions than do unbiased ones. For example, Harry Markowitz used the $1/N$ heuristic for his retirement investments in place of calculating a mean-variance portfolio. Without an empirical test of the predictive accuracy of the $1/N$ heuristic, one likely would conclude that Markowitz relied on a suboptimal strategy. Would one also attribute this to cognitive limitations, as is usually the case for other agents' apparently biased decisions (see also D. Friedman, 1998; D. Friedman, Pommerenke, Lukose, Milam, & Huberman, 2007; Loomes, Starmer, & Sugden, 2003)?

Several phenomena often interpreted in the literature as biases (in the sense of systematic errors of judgment) have been shown to correspond to correct judgments in situations of uncertainty. One group of phenomena comprises judgments of randomness, such as coaches' alleged hot-hand fallacy (J. B. Miller & Sanjurjo, 2018) and people's alleged erroneous intuitions about chance, including the belief in the law of small numbers (Hahn & Warren, 2009). In earlier studies, these judgments were compared to known population probabilities rather than, correctly, to sample probabilities. A second group of phenomena include so-called errors in judgments of low versus high risk, such as overestimation of small risks and underestimation of large risks (Hertwig, Pachur, & Kurzenhäuser, 2005) and overconfidence (Pfeifer, 1994), which again look like systematic errors (bias) but are in fact largely due to unsystematic error (variance). In general, people appear to be quite sensitive to the difference between risk and uncertainty, including small and large samples (Gigerenzer, 2018; Hertwig & Pleskac, 2010).

Expanding on Simon's observation that agents use simple strategies because of the mind's computational limitations, we propose that humans use heuristics because they can yield good decisions under uncertainty or intractability. Under uncertainty, their performance needs to be judged according to their ecological rationality, that is, by their success in achieving a defined criterion, not by principles of logic or consistency. Consistency and success are two different criteria,

which are sometimes uncorrelated (Berg et al., 2016). Moreover, a survey of violations of consistency found little to no evidence that in an uncertain world, coherence violations incur material costs, or if they do, that people would fail to learn (Arkes, Gigerenzer, & Hertwig, 2016). Whereas risky environments allow for general verdicts on the rationality of specific strategies, such generalizations are misguided in uncertain environments.

3.4 Discussion

The term satisficing has been used to mean many things. Sometimes, it is seen as sub-optimal, while at other times it seems optimal, or at least better than other strategies. In this article, we addressed this situation and have argued that there are two different research traditions that contend with two different types of problems, risk and uncertainty. In situations of risk, satisficing is suboptimal if search is ignored. However, when agents need to search for information, as considered by Simon (1955), satisficing can help to solve the cost-benefit trade-off. In situations of uncertainty, satisficing can also help to solve the bias-variance trade-off, which can lead to more accurate decisions than with more complex strategies. Although both research traditions examine distinct classes of environments, they can learn from each other. Specifically, we identify four methodological principles and related areas for fruitful future research on satisficing.

First, we encourage future research in both traditions to study how well decision theories make predictions, such as out-of-sample. We found very few studies that tested their decision models beyond data fitting, despite Friedman's (1953) endorsement of predictive accuracy. The bias-variance decomposition demonstrates how in-sample tests give undue advantage to parameter-rich models that prove inexpedient in predicting behavior.

Second, models of satisficing under uncertainty are currently limited in scope. To date, few of these models specify how aspiration levels are formed and how they are adjusted, although this process appears to be of theoretical importance, particularly in decisions guided by preferences (but see Blettner et al., 2015; Cyert & March, 1963). The satisficing heuristics presented here were tested on inference problems for methodological convenience, although most of these strategies can be applied to decisions involving preferences.

V. L. Smith (1962) already highlighted the disconnect between rational choice models that correctly predict market outcomes in a context where agents do not have the information available that the models assume. Yet little is known how the aggregate level of analysis interacts with the individual level when agents face uncertainty (but see Artinger & Gigerenzer, 2016; Gode & Sunder, 1993; Jamal & Sunder, 2001).

Third, models of satisficing under risk remain largely within the realm of as-if theories, where additional parameters are added to a rational choice model to better account for behavior (e.g., Kahneman & Tversky, 1979; Köszegi & Rabin, 2006). The rational benchmark under risk also depends on the assumed structure of preferences and is not synonymous with rational choice theory. Models of the decision process rather than of decision outcomes could help uncover agents' preference structures, which in turn helps build a coherent theory of choice under risk.

Finally, we encourage further study of decisions given uncertainty, particularly when the problem is ill-defined and the exhaustive set of states of the world and their consequences is not knowable or foreseeable. Over the past years, technological advances have contributed considerably to the rise of machine learning, a branch of computer science distinctly concerned with building algorithms for decisions without full access to information. The rise of this field testifies to the fact that the problem of decisions under uncertainty poses larger problems than commonly acknowledged in normative economics. We therefore encourage competitive tests of specific decision strategies in their natural decision environments. A systematic use of competitive tests can in turn lead to formal analyses of the structural characteristics of environments that regulate the relative advantages of one class of strategies over another. We believe that this research strategy presents a feasible path to a normative theory for decisions under uncertainty.

Chapter 4

How Do Taxi Drivers Terminate Their Shifts when Earnings Are Hard To Predict?

This chapter is published as a working paper and being prepared for submission as: Artinger, F., Gigerenzer, G. & Jacobs, P. (2020), How Do Taxi Drivers Terminate Their Shifts when Earnings Are Hard To Predict?.

In the late 1990s, Colin Camerer, Linda Babcock, Richard Thaler, and George Loewenstein (1997) proposed that taxi drivers conclude shifts after reaching their daily income target. This target-income hypothesis has initiated a fierce debate because it implies that increased hourly earnings lead to shorter shifts as the daily income target is reached more quickly. Yet if drivers can predict an increase in hourly earnings on a given day, this should, all else equal, incentivize them to work longer. Alternatively, if drivers can predict a decrease in hourly earnings, they should stop labor earlier and consume more leisure. To be able to adjust their labor supply in this fashion, however, drivers need to be able to predict future hourly earnings with sufficient precision. Expected utility models assume that drivers act as if they can make such accurate predictions, based on the assumption of "rational expectations." In this paper, we examine the extent to which hourly earnings can actually be predicted by forecasting algorithms, including machine learning techniques. We find that earnings are hard to predict, and assess how drivers decide when to terminate a shift under these circumstances. Finally, we discuss the implications for modeling human behavior under uncertainty, specifically the intertemporal substitution of labor and leisure.

Camerer and colleagues presented an analysis of the wage elasticity of New York City cab drivers' labor supply based on shift data. The wage elasticity describes the relation between percentage

changes in working hours and percentage changes in income. Taxi drivers were deemed useful to study because their hourly earnings fluctuate, implying that their shift earnings and lengths can be used to estimate their wage elasticity. Common sense suggests that this elasticity should be positive: With higher wages, drivers should be willing to work longer as the opportunity costs of leisure increase. Surprisingly, Camerer and colleagues found a negative elasticity across different samples of taxi drivers. In light of these findings, the authors speculated that drivers may use a daily income target.

In response to the findings by Camerer et al. (1997), expected utility theory was modified to include an income target. By this interpretation, drivers reach a decision by comparing the utility of terminating the shift with the expected utility of continuing the shift and then choosing the path that yields higher utility. Unlike neoclassical theory, the utility function includes a particular income target as a reference point. At this reference point, the marginal utility of income decreases, which implies that additional income above the reference point is valued less than below it. For example, building on the seminal model of reference-dependent preferences by Köszegi and Rabin (2006), Crawford and Meng (2011) describe a utility function that includes both an income and a duration target. Their model retains the neoclassical utility framework and adds or subtracts utility based on whether income or working hours are below or above their respective reference points.

Both the negative wage elasticity and its explanation using reference-dependent utility models have attracted considerable criticism. Most prominently, Farber (2005) criticized the focus on shift duration and earnings and reported results from an analysis of each trip within a shift. Using a large dataset of New York City cab drivers, Farber (2008) was able to estimate a model of shift termination that takes the end of each trip as a decision point to decide whether to terminate a shift or not. Farber finds that the probability of shift termination increases discontinuously at a reference income, but also notes that these reference points have little effect on decisions because they are either too high or too unstable. Following this example, subsequent analyses have derived and estimated shift termination models from neoclassical or reference-dependent utility theory. For example, Crawford and Meng (2011) estimated the parameters of their model using Farber's trip data. Because the neoclassical model is nested within the reference-dependent utility model, the authors

used statistical tests on their parameter estimates to conclude that their model offers better fit of the data than neoclassical theory.

Most recently, Farber (2015) used the full record of taxi trips in New York City over a period of five years to examine several aspects of the reference-dependent utility model. His analysis addresses three points. First, he decomposes the variance in average hourly earnings to demonstrate that aggregate hourly earnings are, on average, possible to predict with sufficient precision. In models assuming that the reference point is set at the anticipated level of earnings, this implies that there is little room for reference dependence to play a role as it only affects deviations from the reference point. Second, he estimates a shift termination model and finds that termination probabilities are weakly related to previous earnings. Third, his wealth of data permits him to estimate elasticities for individual drivers. which he finds to be mostly positive. Farber concludes that there is little evidence that reference dependence is an important factor determining the labor supply of NYC taxi drivers.

Overall, the literature on taxi drivers' labor supply has focused on a specific paradigm, both theoretically and empirically. Theoretically, the literature has primarily studied the distinction between reference-dependent and neoclassical utility models. This is presumably due to several factors, including the prominence of utility models in economics and the analytical convenience of nested models. Empirically, the existing literature has initially used aggregate estimations of the wage elasticity but then started to examine individual drivers' behaviors in more detail. However, the general approach has remained the same: A sample of data is fitted using a regression model, and the confidence intervals of one or two variables are used to distinguish between competing models of behavior.

To date, the literature has neglected the more fundamental question of whether utility theory provides the best framework to model taxi drivers' shift decisions. To shed light on this issue, we first ask whether individual drivers can indeed predict hourly earnings with meaningful precision. As we will elaborate in the second section, if such predictability fails, the assumption of rational expectations can still be made but yielding successful decisions is no longer guaranteed. In order to investigate to what extent prediction is possible, we use a dataset of 3,500 drivers and more than 11 million trips in Hamburg, Germany, which we describe in the third section. In the fourth section, we initially replicate Farber's earlier decomposition

of hourly earnings at the aggregate level and then proceed with an analysis at the individual level. We then compare the predictive accuracies of various prediction models commonly used in machine learning, including regularized regression. Across both analyses we find that the predictability of hourly earnings is minimal. Based on these results, we hypothesize that drivers' shift choices are not consistent with the predictions from utility models.

An alternative to utility models are satisficing models in the form of simple heuristic models. For an uncertain environment where the agent cannot determine with sufficient precision the full distribution of possible consequences, Savage (1954), one of the founding fathers of expected utility theory, already suggested that the theory no longer applies. Instead, Herbert Simon (1955) pointed out that agents' behavior is described by satisficing models that dispense with estimating utilities for the possible courses of actions. Heuristic satisficing models are characterized by two components, a search process and an aspiration level (Artinger, Gigerenzer, & Jacobs, 2020b). The bare income target hypothesis proposed by Camerer et al. (1997) constitutes such an satisficing model without the need to integrate it into a utility model. By this model, drivers formulate a daily aspiration level on shift earnings and terminate their shift after reaching the aspiration level. The fifth section of this paper reports a competitive test of different utility and satisficing models for predicting when drivers end their shifts. We find that the overwhelming majority of shifts and drivers are best predicted by satisficing models with aspiration levels on shift earnings or duration. In contrast, fewer than one percent of drivers are best predicted by utility maximization. The final section discusses these findings and broader implications for modeling behavior under uncertainty.

4.1 An Alternative to Rational Expectations

The dominant paradigm to model the intertemporal substitution of labor and leisure is based on utility theory. Taxi drivers face the problem of deciding when to terminate their labor on a given day and are modeled as agents who compare the utilities of terminating and continuing the shift. These models share the assumption that leisure and income are commensurable and that agents are able to map their subjective value of these two goods on a single continuous

utility function. This function allows agents to calculate the utilities of both terminating and continuing the shift.

Whereas the utility of terminating a shift can be computed from cumulative shift earnings and shift duration, the utility of continuing the shift is based on predicted shift income and duration. If drivers knew the probability distribution of their future earnings, they could be modeled as forming "rational expectations" about these earnings and using mathematical optimization to maximize their expected utility. John Muth (1961) was the first to coin the term "rational" expectations whereby agents use the relevant economic theory in forming expectations about macroeconomic variables. To date, the term has seen different definitions as the concept has been applied to diverse settings. In the case of taxi drivers, expectations are formed about individual drivers' earnings rather than macroeconomic outcomes and are based on empirical induction rather than theoretical deduction. That is, drivers need to predict their future earnings rather than derive their value from a theory. We argue that if such prediction is not possible with sufficient precision, the formation of rational expectations does not guarantee high decision quality. To unfold this argument, we consider two different definitions of "rational expectations."

By one rather stringent definition, expectations are rational if they are unbiased and based on all information (e.g., Samuelson & Nordhaus, 1998). However, this definition does not necessarily imply that agents make the most accurate prediction possible. To clarify this point, we include a brief decomposition of prediction error into bias and variance (Geman et al., 1992). This distinction is well known in empirical sciences, including machine learning, where prediction is often a central task.

Suppose the task is to predict y, the value of an unseen item, based on its observables $\boldsymbol{x}$ and a model $m(\boldsymbol{x}, \boldsymbol{p})$ that is trained on a random sample of training data of size n. The error in prediction is measured by the root mean squared error, RSME, and can be decomposed into

$$\text{error} = b^2 + v + e, \tag{4.1}$$

where b denotes the bias in the predictions, v their variance, and e denotes the irreducible error (Geman et al., 1992). To understand bias and variance, recall that there exist L possible learning samples of size n, each of which yields its own predictions $\hat{y}_1, \hat{y}_i, ..., \hat{y}_L$. Bias is

defined as

$$b^2 = \{E_n[\hat{y}_i] - E[y]\}^2, \tag{4.2}$$

where $E_n[\cdot]$ denotes the expectation with respect to different learning samples of size n and $E[\cdot]$ denotes the expectation with respect to unsystematic error. Bias refers to the difference between the average of the possible learning samples and the true value and reflects a misspecification of the model. In contrast, variance is defined as

$$v = E_n[\{\hat{y}_i - E_n[\hat{y}_i]\}^2] \tag{4.3}$$

and refers to the variance of possible predictions around their expected value. Variance therefore reflects the model's sensitivity to idiosyncrasies in the learning sample. Artinger et al. (2020b) offer a more detailed exposition.

Although bias and variance are influenced by multiple factors, they vary according to the number of model parameters. In general, bias tends to decrease alongside the number of model parameters, whereas variance tends to increase alongside the number of model parameters (Hastie, Tibshirani, & Friedman, 2001), creating a trade-off between bias and variance. This trade-off has important implications for model selection: Although unbiased models yield the best in-sample fit, unbiased models are not necessarily those yielding the best predictions. Indeed, the bias-variance trade-off implies that models seeking to reduce bias tend to incur excess error from variance. However, for prediction, the right balance of these two kinds of error depends on the particularities of the problem at hand. In many situations, applying an unbiased model that uses all available information does not yield the best possible prediction and does not result in a rational expectation.

By a second, more generous definition, expectations are rational if "agents do the best they can with what they have" when it comes to the formation of expectations (Maddock & Carter, 1982, p.41). This definition permits agents to use any model suitable for a given prediction problem. It assumes that agents are aware which model or algorithm yields the best balance of bias and variance and choose accordingly. Defined in this way, rational expectations ensure that the agent makes the best prediction possible with the available information.

However, even in this more general definition, decision models based on rational expectations do not necessarily yield the best possi-

ble choices. Rational expectations allow utility maximization to be based on the most precise predictions. However, utility maximization does not necessarily yield better choices than alternative decision strategies that do not require predictions. To see this, it is important to remember that the agent does not know the probabilistic structure of the decision problem, but assumes a simplified structure and estimates the relevant parameters. Both these steps, simplification and estimation, are potential sources of errors.

First, given a complex and dynamic environment, the agent needs to simplify the existing structure of the decision problem — either purposefully to maintain computational tractability or inadvertently for lack of information. Any claim of optimality therefore refers to the simplified problem. Depending on the nature and degree of simplification, the optimal choice in the simplified problem may differ considerably from the true best choice (Simon, 1979). Second, the agent needs to estimate or predict the relevant parameters, such as expectations about future values. The quality of these estimates depends on the agent's choice of prediction model but also on the predictability of the decision environment. Even rational expectations can remain imprecise when randomness is high and available information is scarce or irrelevant.

Both the simplification of the problem and the need for parameter estimation imply that rational expectations are no sufficient condition for utility theory to yield the best possible choice. In the case of shift-ending decisions, forming expectations over future earnings is the Achilles heel of utility theory and of any other decision model building on such expectations. Therefore, our first analysis examines the predictability of hourly earnings. If hourly earnings can be predicted with reasonable accuracy, utility maximization with rational expectations can constitute the normative benchmark for shift decisions. However, if earnings are difficult to predict, the adequacy of utility maximization is questionable in this context.

An alternative approach are heuristic satisficing models. The term "heuristics" borrows from the computer science literature, referring to an algorithm that ignores part of the available information to reach its goal. This disregard can take many forms. Whereas heuristic algorithms in computer science can remain calculation-intensive (e.g., Pearl, 1984), heuristic decision models in cognitive science typically ignore large parts of the available information, resulting in strategies that are quick to implement and easy to communicate (e.g.,

Gigerenzer, 1996b; A. Tversky & Kahneman, 1974).

Consider, for example, the task of predicting whether an existing customer will return for a purchase within a one-year time horizon. This classification task is common in marketing practice to target advertising efficiently. The task can be solved using the pareto-NBD model by Schmittlein et al. (1987), which uses the customer's full purchasing history to calculate the probability of a new purchase. In contrast, the hiatus heuristic requires only the time of the last purchase and predicts a return if that last purchase was after some threshold, and no return otherwise. This frugality makes the strategy easily applicable by marketing managers who seem to use it regularly. Wübben and von Wangenheim (2008) find that this strategy can yield better decisions than the pareto-NBD model because it is less exposed to error from variance.

The hiatus heuristic is an example of the broader class of satisficing heuristics. The common denominator of these heuristics is their use of an aspiration level to reach a decision. An aspiration level is defined here as the threshold on one of the variables of interest that satisfies an aspiration and initiates an action (see also Lewin, Dembo, Festinger, & Sears, 1944). In the case of the hiatus heuristic, the aspiration level is the threshold that separates returning customers from those who do not return. Satisficing heuristics were first described by Herbert Simon (1955, 1956) and later identified and studied across a range of decision tasks (for an overview, see Artinger et al., 2020b). The earnings target described by Camerer et al. (1997) can be modeled more directly as a satisficing model.

Unlike utility theory, satisficing heuristics do not require agents to form expectations about the future. We therefore expect these models to be descriptive of behavior when future hourly earnings are difficult to predict. In contrast, when hourly earnings are reasonably predictable, we would expect drivers' behavior to be better predicted by utility models. Our second analysis tests this hypothesis and compares the predictive powers of two different utility models and four different satisficing heuristics. Before both analyses are presented in turn, the following section presents an overview of the data used for analysis.

4.2 Hamburg Taxi Data

For the purpose of this study, we acquired data for taxi shifts and trips from Hamburg, Germany. The data used by Farber (2015) are unfortunately not openly available for the wider scientific community to analyze. These data comprise a sample of 6,998 drivers, 1,138,726 shifts, and 13,822,310 trips, collected electronically through so-called fiscal taximeters. These devices are regular taximeters that use the cellular network to send trip and shift information to a secure server, where tax authorities can access the data and verify tax statements. The system is commercially maintained.

The data we obtained span the period from January 1, 2013 to December 31, 2015. During these three years, participation in the fiscal taximeter program was voluntary; in exchange for participation, companies received free devices. This incentive was substantial, as companies were aware that from 2017 onward the fiscal taximeter would be mandatory nationwide. Indeed, by the end of 2015, two thirds of the 3,200 taxis in Hamburg used the fiscal taximeter (Levy, 2015). Importantly for the purpose of this analysis, the sample is non-random, as companies self-selected into the data sample.

The data consist of two streams of information, one on shifts and one on trips. The shift stream presents a record of the beginning and end of a shift, as drivers log in and out of the taximeter. Thus, a shift can mean any period of time that the driver defines as such. The shift stream is independent of the trip stream, which keeps a record of each trip and its associated data. In Hamburg, trips can result from taxis being hailed on the street, drivers waiting in lines in places of high demand, drivers accepting trips via smartphone apps, or drivers cooperating with telephone centers that allocate trips among members. Together, both streams give an overview of the entire shift period from the start of the shift to the first trip, to the last trip, and to the point in time when the driver concluded the shift. For both shifts and trips we observed the variables described in Table 4.1.

In addition to the taximeter data, we collected data on observables that we suspected may affect daily demand. First, we identified days with events that may have caused surges in the demand for taxis. These were public holidays (32 days in the observation window), soccer games with Germany playing at the 2014 FIFA world cup (14 days), Hamburg's annual harbor fair, the biggest in the city (8 days), strikes in public transportation (17 days), strikes at Hamburg airport

Table 4.1
Selected Variables Used in Analyses

	variable	description	source
shift data	`sbeg`	shift beginning	taxi data
	`send`	shift end	taxi data
	`searn`	shift earnings in EUR	taxi data
	`sdur`	shift duration in minutes	taxi data
	`precip.d`	daily precipitation per square meter	weather record
trip data	`tid`	trip ID	taxi data
	`tbeg`	trip beginning	taxi data
	`tend`	trip end	taxi data
	`tearn`	trip earnings in EUR	taxi data
	`tdur`	trip duration in minutes	taxi data
	`precip.h`	hourly precipitation per square meter	weather record
in common	`did`	driver ID	taxi data
	`sid`	shift ID	taxi data
	`stype`	shift type	taxi data
	`ctype`	company type	taxi data
	`new`	day in new market	public record
	`weekday`	day of the week	public record
(shock)	`worldcup`	day with Germany playing world cup	public record
(shock)	`harbor`	day during annual harbor fair	public record
(shock)	`strike`	day with strike in public transport	public record
(shock)	`airport`	day with strike at Hamburg Airport	public record
(shock)	`holiday`	public holiday	public record
(shock)	`vacation`	day during Hamburg school vacations	public record

(2 days), and school vacations in Hamburg (111 days). Second, we obtained weather data from the Hamburg weather station, recording the amount of rain per square meter, both per day (DWD Climate Data Center, 2018a) and per clock hour (DWD Climate Data Center, 2018b). These data were matched with each shift based on the day of shift beginning and with each trip based on the clock hour of trip end. An overview of the additional variables used in the analyses is given in Table 4.1.

The separation of shift and trip data allowed us to check their consistency. In a first step, we checked each stream for internal plausibility, for example, whether trip beginning was before trip end. In a second step, we combined the two data streams and checked whether the cumulative trip data were consistent with the recorded shift data. For example, we tested whether the first and last trip associated with a given shift fall within the period between shift beginning and shift end or whether the totals of kilometers or

earnings per shift matched the cumulative trip data. If determining an inconsistency, we deleted the corresponding shift and all associated trips from the dataset, unless we deemed the error to be small and fixable. This happened in either one of three cases. The first occurred when the shift data gave the correct trip count and all trips took place between shift beginning and end but the cumulative earnings did not match. In this case, shift earnings were set to the cumulative trip total. Similarly, in the second case the cumulative trip earnings matched shift earnings and all trips took place within the shift period, but the number of trips in the shift data was incorrect. Here, we set the trip count in the shift data to equal the number of trips observed in the trip data. Finally, when all trips fell within the shift period but the trip count in the shift data was off by one, we adjusted the trip count to the number observed in the trip data. In total, we made adjustments to 327,320 shifts with 5,838,654 trips. In all other cases of inconsistency, we deleted the corresponding shift and all associated trips from our data. This happened with 70,178 shifts comprising 937,340 trips.

Two events fall into our observation window that have plausibly changed market incentives. First, on September 16, 2014, the Hamburg Senate raised the taxi fare by about eight percent on average[1], effective from October 1, 2014. As in other German cities, taxi fares in Hamburg are regulated by the local authority and reviewed every few years. The 2014 increase took place in anticipation of the second major change in the taxi market, the introduction of a national minimum wage in Germany. In August 2014, the German government introduced a nationwide minimum wage of 8.50 EUR per hour, effective from January 1, 2015. This minimum wage was long expected and relevant to German taxi markets as many drivers had earned less than minimum wage. The new minimum wage applied to all employed drivers but not to self-employed drivers. Taxi companies with employees responded to the minimum wage in different ways, from incentives to maximize revenue to regulations that declared waiting periods as stand-by time. In addition, some companies were split up, making drivers henceforth self-employed. Because the two changes are likely to affect taxi drivers' behavior, we have split the

[1]Specifically, the base charge increased from 2.90 EUR to 3.20 EUR, the rate for the first four kilometers increased from 2.20 EUR to 2.35 EUR, for the next five kilometers from 1.90 EUR to 2.10 EUR and afterwards from 1.40 EUR to 1.45 EUR.

observation window into two equally long periods. The *old market*, from October 2013 to September 2014, is characterized by pre-increase fares and no minimum wage. The *new market*, from January 2015 to December 2015, is governed by increased fares and a minimum wage. Shifts and trips before October 2013 and in the three months between these two periods were ignored.

We classified shifts as day and night shifts to examine them separately where necessary. Specifically, shifts starting before noon were classified as day shifts, whereas those starting after noon were classified as night shifts. These two shift types have distinct demand profiles. Aggregated across all days, peaks in demand occurred between 7am and 11am, and between 4pm and 10pm. Day shifts tended to cover the morning peak, and night shifts tended to cover the evening peak. Because these peaks occurred at different points during the shifts[2], we decided to examine driver behavior separately for each of them.

We also placed additional restrictions on the data to receive our final dataset. First, we restricted shifts to those lasting less than 24 hours to exclude drivers who own their car and work long hours with many mid-sized breaks. Second, we excluded shifts with fewer than three trips assuming that these were likely concluded prematurely. Third, we removed shifts that started on days on which the observation window either began or ended (January 1, 2013 and December 31, 2015) to be sure that shift data were complete. Fourth, we removed shifts taking place on days with switches to and from daylight saving time, comprising six days during the three-year period. Finally, we restricted our analysis to drivers with more than 25 day shifts or 25 night shifts or both to have enough material for out-of-sample prediction. The final data set consisted of 3,261 drivers with 785,273 shifts and 10,094,685 trips.

Finally, Hamburg taxi drivers work at one of three types of companies: those with multiple cars and multiple drivers, those with one car but multiple drivers, and those with one car and one driver only. We refer to drivers in the last category as single drivers, as these drivers own their taxi and are not subject to the minimum wage legislation or the restrictions of a shift schedule. The working

[2]Figure 4.A in Appendix A shows both the number of shift and trip beginnings across clock hours, separately for day and night shifts. We observe that in day shifts, most drivers appear to start their shifts around the time of the first peak, whereas for night shifts, most drivers start considerably before the evening peak.

arrangements of these drivers most closely resemble those of New York drivers. In total, the data comprise 784 single drivers with 260,224 shifts and 2,993,352 trips. Where appropriate, we examine single drivers separately to see whether findings are an artifact of employed drivers' working conditions.

4.3 Can Drivers Predict Next Hour's Earnings?

The first of our two analyses is an empirical examination of the predictability of hourly earnings. In one form or another, each utility model assumes that drivers compare the utility of terminating with an expected utility of not terminating their shift. That is, these models assume that transitory wage variation can be predicted with sufficient precision, such that drivers can form expectations about their utility if they were to continue driving. We refer to this assumption as the predictability assumption.

The predictability assumption underlies the normative assertion that labor supply elasticities ought to be positive. Suppose that hourly earnings did not vary during the day. Under such a regime, drivers could choose their working hours as soon as they became aware of the hourly earnings for the day. However, when hourly earnings fluctuate in the course of a day, drivers need to decide incrementally whether the next hour will be worth their time. Whether or not drivers are able to concentrate their working time on profitable hours then depends on the predictability of these earnings: If hourly earnings are known with certainty in advance of each hour, drivers can, in principle, compare them with past earnings and decide whether the shift is worth extending. However, if hourly earnings are fully random, drivers trying to allocate working hours to profitable periods find themselves unable to do so. Therefore, the normative assertion that elasticities are positive requires that hourly earnings can be predicted with a sufficient level of precision.

Farber (2015) has presented a systematic analysis of the predictability of hourly income using his data from New York City. To this end, he calculated the hourly earnings for each of the 43,824 clock hours in his observation window, averaged across all of the 8,802 drivers on duty during that hour. He then used two OLS regressions to relate the variance in average hourly earnings to variation in variables that are readily observable. The first of these models regressed average hourly

earnings on a dummy for each year and a dummy for hours after the fare increase. For this model, the variance of the predicted values reflects permanent wage variation. The second model regressed the residuals of the first model on a dummy for each week of the year, a dummy for each hour of the week, and a dummy for public holidays. Faber refers to the variance explained by the second model as transitory but anticipated, whereas the variance left unexplained by both models is transitory and unanticipated. In his findings, almost ninety percent of the variance in average hourly earnings can be anticipated because it is either permanent or transitory but anticipated. He concludes that drivers can predict hourly earnings with a sufficient level of precision.

We argue that this conclusion is unwarranted for two reasons. First, it confuses the aggregate level of analysis with the individual level. From predictable aggregate earnings we cannot conclude predictable individual earnings, which are necessarily harder to predict. Indeed, insurances have evolved around discrepancy between the predictability of aggregate and individual outcomes. Second, the in-sample fit of a statistical model yields limited conclusions about the ability to predict accurately outside of the sample or population. Because drivers can calibrate their regression model only on past data, the model's learning sample may systematically differ from the sample to which it is applied. Accordingly, the accuracy in out-of-sample prediction can be substantially lower than in in-sample fitting. These two shortcomings call for a more detailed examination of the predictability of hourly earnings. We therefore replicate Farber's decomposition analysis and then extend it by an analysis of drivers' ability to predict their hourly earnings.

4.3.1 Variance Explained in Fitting

To replicate the original analysis by Farber (2015), we calculate $\log[\texttt{earn}_{h,i}]$, the natural logarithm of earnings of individual i in clock hour h. We then aggregate across drivers to obtain $\log[\texttt{earn}_h]$, denoting the natural logarithm of earnings during clock hour h averaged across all drivers active during that clock hour. This dependent variable is then modeled as follows

$$\log[\texttt{earn}_h] = \alpha_0 + \alpha_1 \texttt{y}_h + \alpha_2 \texttt{new}_h + \epsilon_h, \tag{4.4}$$

where $\texttt{y}_h$ denotes a dummy for the year and $\texttt{new}_h$ denotes a dummy for the new market conditions with increased fares and minimum wage. The variance of its predicted values $\log[\widehat{\texttt{earn}}_h]$ represents the portion of variance in $\log[\texttt{earn}_h]$ explained by permanent changes in demand. In contrast, the variance of the residuals ϵ_h represent the portion of $\log[\texttt{earn}_h]$ unexplained by permanent changes. These residuals are then modeled as follows,

$$\epsilon_h = \beta_0 + \boldsymbol{\beta}_1 \texttt{w}_h + \boldsymbol{\beta}_2 \texttt{dh}_h + \beta_3 \texttt{holiday}_h + \gamma_h, \tag{4.5}$$

where $\texttt{w}_h$ denotes a vector with 51 dummies for the week of the year, $\texttt{dh}_h$ denotes a vector with 167 dummies for the hour of the week, and $\texttt{holiday}_h$ denotes a dummy for a public holiday. The variance in the predicted values of this regression represent transitory but anticipated variation in $\log[\texttt{earn}_h]$, whereas the variance of the residual γ_h represents transitory and unanticipated variation.

For this analysis, we use data from 2014 and 2015 only. Of the $2 \times 365 \times 24 = 17{,}520$ clock hours between January 1, 2014 to December 31, 2015, there were 15,106 clock hours for which there was at least one driver active so that we could calculate average earnings. During the remaining clock hours, no driver in our data was on shift. In a first step, we estimate equations (4.4) and (4.5) using all available clock hours, yielding the variance decomposition shown in the first line of Table 4.2. In this analysis, the total variance in $\log[\texttt{earn}_h]$ to be explained is 0.023 and around 6 percent of this variance is due to permanent changes, whereas around 62 percent is due to anticipated transitory changes. This leaves 32 percent of the variance unexplained, about three times the share found by Farber (2015).

To understand this result, recall that Farber's data comprise 8,802 drivers, considerably more drivers than ours — for one because New York City is much larger than Hamburg and also because our data are only a subsample of all drivers in Hamburg. For this reason, hourly earnings are averaged over fewer drivers. In some clock hours, only one driver is active. We therefore repeat the analysis, considering only clock hours with a minimum number of active drivers. The results of these analyses are shown from the second line of Table 4.2 onwards. As the required minimum number of drivers per hour increases, we find that the unanticipated portion of the variance decreases. For example, when considering only clock hours with at least 275 drivers active, we have 5,316 hours left for analysis with a total variance of

Table 4.2
Hourly Earnings: Variance Explained in Fitting

			Percent of Total Variance		
				Transitory	
Minimum Number of Drivers per Hour	Hours	Variance	Permanent	Anticipated	Unanticipated
1	15,106	0.02	6.29	61.71	32.00
25	15,022	0.02	6.31	62.72	30.97
50	14,093	0.02	6.06	65.22	28.72
75	13,338	0.02	5.76	67.27	26.97
100	12,573	0.02	5.52	69.42	25.07
125	11,633	0.02	4.99	71.83	23.18
150	10,407	0.02	4.61	74.30	21.10
175	9,187	0.02	4.31	76.63	19.06
200	8,249	0.03	4.32	78.31	17.37
225	7,295	0.03	3.77	80.59	15.64
250	6,335	0.03	3.41	82.68	13.91
275	5,316	0.03	2.89	85.08	12.03
300	4,353	0.03	2.39	86.60	11.01
325	3,565	0.03	2.12	87.64	10.24
350	2,904	0.03	1.84	88.64	9.52
375	2,219	0.03	1.95	89.01	9.05
400	1,568	0.03	0.98	90.14	8.88
individual drivers	3,435,572	0.70	0.04	4.30	95.66

0.029. Of these hours, the analysis above leaves around 12 percent unexplained, similar to the findings reported by Farber (2015). We repeated the analysis with additional variables on demand shocks, but the results remained virtually identical.

The change in the proportion of unexplained variance illustrates our argument that high shares of explained variance result from high levels of aggregation. Such aggregation is useful to estimate overall demand or some measure of aggregate behavior. However, we argue that an assessment of individual drivers' potential to predict future earnings necessarily has to examine the decision environment of individual drivers. With an average of 1.6 to 2.2 trips per clock hour, drivers' hourly earnings depend crucially on the profitability of individual trips. To illustrate, consider three passengers reaching Hamburg central station at 9.24am on the same train, two of whom

need to get to hotels around the corner and one who needs to get to the airport outside of the city center. The exact assignment of passengers to the first three taxis in line outside of the station determines which driver is looking at a profitable 30- to 40-minute trip to the airport and which driver spends five to ten minutes on a short trip and lands up back at the end of the taxi line. Although consequential, this assignment is random in the sense that it depends on a plethora of unknown factors, including which passenger exits the train closer to the escalator or walks faster. One may argue that in the face of such randomness, the aggregate pattern is the best indicator drivers have. Although this may be true, it does not imply that a predictable aggregate pattern is useful for individual drivers. Its usefulness depends on the magnitude of the pattern relative to the magnitude of random noise.

To obtain a better picture of the predictability of individual drivers' hourly earnings, we repeat the variance decomposition once more without any aggregation across drivers. Instead of $\log[\texttt{earn}_h]$, the log earnings averaged across drivers, we use the original variable $\log[\texttt{earn}_{h,i}]$, the log earnings of an individual driver. This analysis comprises earnings of 3,435,572 clock hours of individual drivers, which are modeled as follows, with

$$\log[\texttt{earn}_{h,i}] = \alpha_{10} + \boldsymbol{\alpha}_{11}\texttt{y}_{h,i} + \alpha_{12}\texttt{new}_{h,i} + \epsilon_{h,i} \tag{4.6}$$

modeling permanent variation, and

$$\begin{aligned}\epsilon_{h,i} = \beta_{10} &+ \boldsymbol{\beta}_{11}\texttt{w}_{h,i} + \boldsymbol{\beta}_{12}\texttt{dh}_{h,i} + \beta_{13}\texttt{holiday}_{h,i} \\ &+ \beta_{14}\texttt{vacation}_{h,i} + \beta_{15}\texttt{worldcup}_{h,i} + \beta_{16}\texttt{harbor}_{h,i} \\ &+ \beta_{17}\texttt{airport}_{h,i} + \beta_{18}\texttt{strike}_{h,i} + \gamma_{h,i}\end{aligned} \tag{4.7}$$

modeling anticipated transitory variation, where subscript i denotes the driver and subscript h denotes clock hour, as before. In addition, we have added all additional dummies of demand shocks listed in Table 4.1.

The results of this decomposition are shown at the bottom of Table 4.2. As expected, the total variance to be explained is considerably larger than the variance of average hourly earnings and the portion to be explained by the models in equations (4.6) and (4.7) is much smaller: Jointly, only 4 percent of the variance is due to either permanent or anticipated transitory changes. By implication, 96 percent of

the variance is transitory and unanticipated. This finding contrasts starkly with the results obtained earlier for the aggregate level and appears to be a direct consequence of the level of analysis.

4.3.2 Error in Prediction

As a next step, we recognize that drivers need to generalize from the past to the future. So far, we have examined predictability by explained variance in fitting the full set of observations. When deciding whether to continue a shift, drivers do not have access to future data points — but need to predict outside of their learning sample[3]. Our analysis mirrors this setup and examines the accuracy in predicting future earnings after each trip, based on a hypothetical learning sample of past trips.

To this end, we extract all trips that took place (a) after January 1, 2014, (b) at least one hour after shift beginning, and (c) at least one hour before shift end. For each of these 6,146,600 trips t, taken by driver i, we sum up the earnings of all trips by i finishing in the 60 minutes after t was completed. This quantity gives the variable of interest, which we refer to as next-hour earnings. Note that this approach is not identical to calculating earnings across clock hours, as more than one trip can end per clock hour. Instead, we follow Farber (2005) in assuming that drivers use trip ends rather than full clock hours as decision points for terminating or continuing shifts and try to predict earnings during the subsequent 60 minutes. To obtain the learning sample used for prediction, we create for each trip t a random sample of 1,000 trips taken by any driver in the 180 days prior to t[4]. This sample is specific to each trip t. The learning sample includes trips of other drivers to reflect exchanges among colleagues about demand on different days. For each trip in the learning sample, i is assumed to be aware of realized next-hour earnings and the following covariates: earnings in the hour before, average next-hour earnings after all previous trips of i, current trip number, current shift

[3] In fact, one may argue that drivers need to predict outside of their learning population, depending on the stationarity assumptions one is willing to impose on the demand function. Consequently, we assume stationarity, such that drivers need to predict out-of-sample.

[4] Indeed, we do this separately for trips before and after the fare increase. That is, for trips after the fare increase, the learning sample is restricted to other trips after the fare increase.

Algorithm 4.1
Competitive Test of Models for Predicting Next-Hour Earnings

```
1  foreach trip t do
2  |  1. select random learning sample of 1,000 trips ending before start of t,
   |    irrespective of driver;
3  |  2. record observed next-hour earnings after t;
4  |  foreach model m do
5  |  |  3. calibrate the model on learning sample;
6  |  |  4. use necessary covariates to calculate predicted next-hour earnings
   |  |    for t;
7  |  |  5. calculate residual between predicted and observed next-hour
   |  |    earnings;
8  |  end
9  end
10 6. calculate for each model the root-mean squared residual across all trips.
```

earnings, current shift duration, a dummy for the year, dummies for the week of the year, dummies for the hour of the week, a dummy for a rainy hour, and dummies for the demand shocks listed in Table 4.1. In this way, we obtain 6,146,600 trips and their associated learning samples.

For this analysis, we adopt the perspective of the driver rather than the omniscient analyst. It differs from the previous analysis in two respects. First, it uses a limited sample of past trips to make predictions of future earnings. The error in predicting outside of the learning sample is typically larger than the error in fitting, particularly when the learning sample is small. Second, this analysis extends the number of cues that are available to each driver. Whereas the variables in the previous analysis reflected permanent changes as well as demand cycles and shocks, these data are supplemented here with data of the current shift, such as cumulative shift earnings and the driver's average hourly earnings in the past. With the inclusion of additional variables on the one hand and the change in methodology on the other, it remains unclear how well hourly earnings can be predicted.

As explained in section 4.1, accurate predictions require a good balance of bias and variance. Because the model with the best balance cannot be determined a priori, we test different candidate models competitively by feeding each of them with the same data and comparing the accuracy of their predictions. This approach is common in machine learning applications where models typically

span different, non-nested classes of models, from linear models to non-linear ones such as decision trees or neural networks. The procedure for this analysis is summarized by Algorithm 4.1.

Two of the five candidate models are regression models. The first candidate model, REG, is elastic-net regression of all covariates listed above (Zou & Hastie, 2005). The model is similar to OLS regression but addresses the exposure of OLS to error from variance by means of regularization. Regularization refers to a penalty of model complexity by "shrinking" the estimates, that is by subtracting a function of the OLS coefficients from the estimated value. Compared with OLS regression, the elastic net model has two additional parameters, λ and α, determining the amount and kind of shrinkage, respectively. In the study at hand, these parameters are determined based on 10-fold cross-validation from all data, where $\lambda = 0.001$ and $\alpha = 0.41$ were found to yield the best performance. Because elastic net regression reduces to OLS when $\lambda = 0$, the model used here yields predictions very similar to OLS. The second candidate model we use, LOG, is similar to model REG but first transforms all continuous variables into their logarithmic versions and re-transforms the predicted values back to EUR.

Apart from these computationally intensive models, we also test three simpler models that ignore part of the available data. These final three candidate models capitalize on low variance, and each uses only one variable to predict earnings in the following hour. Specifically, candidate model PAST uses the driver's average next-hour earnings of *all past* trips to predict earnings in the next hour. In contrast, candidate model LAST uses earnings of only the *previous* hour to predict earnings in the next hour. Finally, candidate model MEAN does not use any covariate but instead the average next-hour earnings in the learning sample to predict earnings. In contrast to the regression models above, these models can, in principle, be implemented by drivers using no more than pen and paper and basic arithmetic.

Table 4.3 shows the results. Across all 6,144,973 trips, next-hour earnings to be predicted are on average 16.85 EUR and vary with a standard deviation of 13.87 EUR. The standard deviation can be viewed as the root mean squared error (RMSE) of the mean: If the average across all trips were known and used to predict next-hour earnings for all trips, the RMSE of these predictions would be equal to the standard deviation. For this reason, we use the standard deviation

as a benchmark for the RMSE of other prediction models.

Consider first the two regression models, REG and LOG, shown in columns six and seven of Table 4.3. For the REG model, the RMSE in prediction is 12.41 EUR. Compared with the standard deviation, this equals an 11 percent reduction in the error. In contrast, the LOG model yields a RMSE of 14.29 EUR, considerably higher than the REG model.

Consider next the three simple models in columns eight to ten of Table 4.3. Whereas model LAST yields errors higher than the standard deviation, models PAST and MEAN yield predictions similar in quality to elastic-net regression; the errors of the PAST and REG models are almost indistinguishable. This finding offers a practical illustration of the bias-variance trade-off: Despite the fact that the REG model relies on all of the available data, its predictions are, on average, hardly better than those of the model that ignores most of this information. This result indicates that the additional variables, combined with the linear structure of the regression model, expose it to higher variance in its predictions.

Irrespective of the model, the findings are sobering. Even the best model yields predictions that are only slightly better than the standard deviation of next-hour earnings. This finding implies that the models tested here are not fit to yield useful predictions for the task at hand. As is typical of competitive tests, we cannot rule out the possibility that other models exist that yield better performances than the ones tested here. However, we have made an effort to include models that make extensive use of the available data, require high computational capacities, and take precautions that guard them against excessive error from variance. Therefore, these findings document the difficulty of predicting next-hour earnings, irrespective of the available computational power.

As an additional test of these conclusions, we expand the forecasting window. So far, we have assumed that the task is to predict earnings one hour beyond trip end. However, it may rightly be pointed out that prediction over one hour may be difficult because it is too short a timescale. After all, even drivers who spend one hour taking a short fare and returning to the waiting line may get an airport trip the next time. Therefore, if the forecasting window is expanded and drivers form predictions over a longer time horizon, part of the randomness cancels out. We therefore repeat the analysis and expand the prediction window. Instead of predicting earnings one

Table 4.3
Hourly Earnings: Error in Prediction in EUR

	Next Hour's Earnings			Model RMSE				
Trips	Median	Mean	SD	REG	LOG	PAST	LAST	MEAN
6,144,973	15.90	16.85	13.87	12.41	14.29	12.58	17.25	13.86

hour ahead, we use the same method to predict the average hourly earnings of the next two, three, four, or five hours. Although average next-hour earnings remain stable, the standard deviation is reduced, as are the RMSEs of most models. With a prediction horizon of five hours, the RMSE of the best predicting model improves 16 percent over the standard deviation. This result shows that predicting over longer time horizons allows the models tested here to yield better predictions. At the same time, these predictions remain too poor to conclude that next-hour earnings can be well predicted.

4.3.3 Conclusion: Prediction of Hourly Earnings is Difficult

This first analysis has demonstrated the difficulty for individual taxi drivers to predict their future earnings. It has shown that the difficulty is rooted not in drivers' limitations in computational power but in a lack of predictability of their decision environment. This finding illustrates that rational expectations do not imply accurate predictions. Even if agents were to base their predictions on a state-of-the art machine learning model, the resulting predictions of their individual outcomes would remain prohibitively imprecise.

The low level of predictability casts doubt on utility theory as a model of drivers' shift choices. Because utility maximization with imprecise expectations is unlikely to be rational, there remains little reason to expect shift choices to be consistent with utility theory. At the same time, it remains true that a test of the assumptions cannot replace a test of the theory (M. Friedman, 1953). However, as we have argued above, the existing evidence does not comprise a comparison of utility theory and alternative models in predicting taxi drivers' shift choices. The next section therefore addresses the question of how to best predict drivers' behavior.

4.4 How Are Drivers' Shift Ends Best Predicted?

At the heart of this text is the question of how drivers' shift ends are best predicted. Although Camerer et al. (1997) speculated about a target income, we have not found any test of this hypothesis. Instead, authors have devised and examined different versions of utility models that incorporate target incomes within its system of trade-offs. This narrow selection of models can answer the question of which version of utility theory is most consistent with observed data, but it does not yield conclusions about the usefulness of utility theory over alternative models. In the following, we present two of the utility models examined before and present four alternative models of shift termination..

The six models are then compared on their accuracy in out-of-sample prediction. Previous analyses have unanimously used in-sample fitting to draw conclusions about the descriptive power of different theories (e.g., Chou, 2002; Crawford & Meng, 2011; Farber, 2015; Oettinger, 1999). However, the bias-variance trade-off implies that in-sample fit cannot be used to judge predictive accuracy, as these two are different measures. Following M. Friedman (1953), we argue that the ultimate purpose of theories is prediction of unobserved behavior and focus on an examination of their predictive accuracy. We therefore employ a competitive out-of-sample test that resembles the previous analysis. Rather than focusing on the absolute performance of statistical models, we are now interested in the relative performance of the different behavioral models.

4.4.1 Candidate Strategies

We have selected the set of six models from different strands of the literature and modeling approaches. The first approach, common in economics and much of statistics, models decision outcomes, typically at some level of aggregation. Models of this kind seek to approximate observed outcomes as closely as possible, often using linear models with error terms. The second approach, common in cognitive science and machine learning, models the decision process. Models of this kind usually have no error terms and are deterministic in the sense that they output a specific decision rather than an average or an

approximation[5].

We argue that process models are more natural models of human decisions than outcome models. These models take the information available to the agent as input and try to mirror the decision process. To this end, models need to be defined algorithmically, that is, in the form of a decision strategy. In the case at hand, a decision strategy takes as input the information available to the driver at the end of each trip and decides whether the driver stops after this trip or not. The predicted shift end can then be compared with the observed shift end.

The existing literature on taxi drivers' shift ends proposes utility models, which are often implemented as outcome models (e.g., Farber, 2005). These models specify the utility function, from which analysts can derive statistical models to link the probability of a trip end to environmental factors. The agents' decision process of comparing the utility of ending a shift to the expected utility of continuing the shift is assumed but not explicitly modeled. In particular, the process by which these expectations are formed is often left unspecified. To allow for a fair comparison among models, however, the competitive framework of our analysis requires that models use the same input variables and output a prediction after each trip. Therefore, we had to implement the utility models algorithmically and specify how agents form expectations. Among the many possibilities, ranging from various forms of regression analysis to random forests, we devised an algorithm with no additional parameters that builds on the average of similar past shifts. In light of the fact that the PAST model in the previous analysis yielded almost identical performance to that of the REG model, we selected an algorithm that is conservative insofar as the lack of additional parameters does not expose the strategy to additional variance. In this way, we could devise an algorithmic version of all decision strategies considered here and presented in turn.

S1: Neoclassical Utility The first strategy considered here is the neoclassical utility model of intertemporal substitution, as presented by Crawford and Meng (2011). According to this model, driver *i*

[5]Leo Breiman (2001b) saw a similar distinction in modeling approaches in statistics, which he referred to as algorithmic and data modeling, respectively (see Brighton, 2020, for a discussion).

compares utilities at the end of each trip t and terminates a shift as soon as the utility of terminating exceeds that of continuing. The utility of terminating the shift is calculated as follows,

$$U_t^T = r_t - \frac{\psi}{1+\upsilon} \times d_t^{1+\upsilon}, \tag{4.8}$$

where r_t and d_t denote the shift earnings and duration at the end of trip t, respectively, and index i is suppressed for brevity. This termination utility is compared with the continuation utility, which gives the expected utility from not terminating the shift. The continuation utility is calculated as follows,

$$E_t[U^C] = E_t[r] - \frac{\psi}{1+\upsilon} \times E_t[d]^{1+\upsilon}, \tag{4.9}$$

where $E_t[\cdot]$ denotes the expected value after trip t, r denotes shift earnings, and d denotes shift duration. To calculate the expected values of earnings and duration, driver i consults their recollection of comparable shifts. Among i's previous shifts of type m_t (day or night) with earnings at least equal to r_t and duration at least equal to d_t, similar ones are identified with the same values on demand shock, day of the week, and rain. That is, all previous shifts similar along these dimensions are identified and their shift earnings and duration are averaged to obtain the expected value for the shift at hand. If no previous shift has the same combination of demand proxies, individual variables are removed in reverse order until a recollection set of minimum size $N = 1$ is found. For example, if night shift s takes place on a rainy Friday without demand shock but i has only experienced sunny Fridays without demand shock, these Fridays are used for comparison (rather than, say, rainy Saturdays) because rain is the first variable to be ignored. In some rare cases, there is no comparison set because no previous shift of the same type was as long or as remunerative. In these cases, the expected values are calculated as $E_t[d] = d_t + 60$ and $E_t[r] = r_t + 60 \times \frac{r_t}{d_t}$, where d_t is measured in minutes. If $U_t^T > E_t[U^C]$, the shift is predicted to terminate after t, otherwise the procedure is repeated after the next trip, $t + 1$. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

S2: Reference Utility The second strategy considered here follows the same process as the neoclassical utility strategy but uses a

different utility function to compute U^T and $E_t[U^C]$. The utility function used here is described by Crawford and Meng (2011) and based on Köszegi and Rabin (2006). This function augments neoclassical utility, which they refer to as *consumption utility*, by a *gain/loss utility* that depends on targets for both shift earnings and shift duration, as well as a parameter of loss aversion, λ, which reduces utility when earnings are below the earnings target or hours are above the duration target or both. In addition, a parameter η governs how relevant gain/loss utility is relative to consumption utility.

According to this model, the termination utility is calculated as

$$\begin{aligned} U_t^T = &(1-\eta) \times \left[r_t - \frac{\psi}{1+\upsilon} \times d_t^{1+\upsilon} \right] \\ &+ \eta \times \left[1_{(r_t - R \leq 0)} \times \lambda \times (r_t - R) + 1_{(r_t - R > 0)} \times (r_t - R) \right] \\ &- \eta \times \left[1_{(d_t - D \geq 0)} \times \lambda \times \left[\frac{\psi}{1+\upsilon} \times d_t^{1+\upsilon} - \frac{\psi}{1+\upsilon} \times D^{1+\upsilon} \right] \right] \\ &- \eta \times \left[1_{(d_t - D < 0)} \times \left[\frac{\psi}{1+\upsilon} \times d_t^{1+\upsilon} - \frac{\psi}{1+\upsilon} \times D^{1+\upsilon} \right] \right], \end{aligned} \tag{4.10}$$

where r_t and d_t denote shift earnings and duration at the end of trip t, respectively; R and D denote the earnings and duration targets, respectively; and index i is suppressed for brevity. Again, this termination utility is compared with the continuation utility, calculated as follows,

$$\begin{aligned} E_t[U^C] = &(1-\eta) \times \left[E_t[r] - \frac{\psi}{1+\upsilon} \times E_t[d]^{1+\upsilon} \right] \\ &+ \eta \times \left[1_{(E_t[r] - R \leq 0)} \times \lambda \times (E_t[r] - R) + 1_{(E_t[r] - R > 0)} \times (E_t[r] - R) \right] \\ &- \eta \times \left[1_{(E_t[d] - D \geq 0)} \times \lambda \times \left[\frac{\psi}{1+\upsilon} \times E_t[d]^{1+\upsilon} - \frac{\psi}{1+\upsilon} \times D^{1+\upsilon} \right] \right] \\ &- \eta \times \left[1_{(E_t[d] - D < 0)} \times \left[\frac{\psi}{1+\upsilon} \times E_t[d]^{1+\upsilon} - \frac{\psi}{1+\upsilon} \times D^{1+\upsilon} \right] \right], \end{aligned} \tag{4.11}$$

where $E_t[\cdot]$ denotes the expected value after trip t, r denotes shift earnings, and d denotes shift duration. To calculate the expected values during each shift s, driver i consults their recollection of comparable shifts.

Both versions of utility theory follow the conventional economic approach to decision modeling. Different attributes, here time and money, are brought onto a common scale, utility, and can be traded

off against one another. The exact trade-off is governed by a set of parameters that capture different preferences. In this theory, decisions respond to changes in demand if these changes are anticipated through the expected values of shift earnings and shift length.

S3: Earnings Target The third strategy is the first interpretation of a "raw" income target. In contrast to earlier decision models inspired by the earnings target, this model defines the target on the raw earnings scale, not on a utility scale. By implication, working hours and earnings are incommensurable and cannot be traded off against one another. The earnings target algorithm is defined as follows.

Driver i evaluates current shift earnings r_t at the end of each trip t. If $r_t > \rho_i$, the shift is ended and the end time of t is predicted to be the shift end. Parameter ρ_i is individual to each driver and estimated from the data. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

Strategies S1 – S3 were derived from the existing literature on taxi drivers' shift decisions. To expand the set of strategies tested, we generated additional satisficing strategies by varying the aspiration variable, that is, the variable on which the aspiration level is defined. Whereas the earnings target terminates shifts as soon as cumulative shift earnings exceed their aspiration level, the following three heuristics use time on shift, clock hour, or the hiatus between trips to terminate a shift.

S4: Duration Target The fourth strategy considered here is the duration target. This strategy corresponds to a driver with a fixed shift duration planned at shift start, irrespective of information gathered and demand observed during the shift. The duration target algorithm is defined as follows.

Driver i evaluates current shift duration d_t at the end of each trip t. If $d_t > \delta_i$, the shift is ended and the end time of the previous trip is predicted to be the shift end. Parameter δ_i is individual to each driver and estimated from the data. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

S5: Clock Target The fifth strategy considered here is the clock target. Similar to the duration target, the clock target strategy makes decisions based on time, but uses clock time rather than shift duration. When shifts consistently start at the same time, the two strategies lead to the same prediction. However, when shifts start at different times, ending after a specific shift length implies different clock times. The clock target algorithm is defined as follows.

Driver i evaluates current clock time c_t at the end of each trip t. If $c_t > \chi_i$, the shift is ended and the end time of the previous trip is predicted to be the shift end. Parameter χ_i is individual to each driver and estimated from the data. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

S6: Hiatus Target The sixth and final strategy considered here is the hiatus target. Whereas the hiatus heuristic described in the previous section was defined by the hiatus between two purchases of the same customer, the hiatus target uses the hiatus between two subsequent trips, usually by different customers. The heuristic target algorithm is defined as follows.

Driver i evaluates the previous hiatus between two trips, h_t at the end of each trip t. If $h_t > \eta_i$, the shift is ended and the end time of the previous trip is predicted to be the shift end. Parameter η_i is individual to each driver and estimated from the data. If the predicted shift end is later than 24 hours after shift beginning, the prediction is truncated at 24 hours.

Strategies S3 to S6 are similar in structure but differ in their aspiration variables. The choice of the aspiration variable can be regarded as an expression of which goal the driver prioritizes. For example, an earnings target promises a fixed income, whereas a duration target promises a fixed shift length. Thus, a driver using a duration target may be interpreted as valuing a daily routine more than earning a fixed amount of income. Such drivers may have family obligations that restrict them from extending their shifts beyond a particular duration. With the exception of the hiatus target, the satisficing strategies essentially implement simple goals independent of demand. In contrast, it appears unrealistic that drivers have preferences over the hiatus between trips. Instead, they may use waiting time as a signal

of demand: When demand decreases relative to the number of active taxis, waiting time should increase on average. The hiatus target therefore is the only satisficing strategy that responds to changes in demand.

To corroborate our selection of decision strategies, we conducted a survey among taxi drivers operating in the Hamburg taxi market. Drivers were recruited in two ways. First, taxi companies were approached by email and asked to forward the survey to their employees. Second, taxi drivers were approached in person in Hamburg on a sunny day in May outside the observation window of the data. Participation in the survey was voluntary but incentivized with two prizes of 100 EUR, awarded at random. Overall, 70 drivers participated in the survey. Among other questions, they were asked to verbalize their strategies for calling it a day. For 27 drivers, the answers could not be classified because they were too unspecific (e.g., "too little demand") or because they referred to traffic and tiredness. Because traffic is either idiosyncratic or correlates with clock hour, these answers were ignored. The remaining answers were mapped onto the strategies above, with 5 drivers mentioning target earnings, 6 drivers mentioning shift duration, 7 drivers mentioning clock hour, and 16 drivers mentioning the hiatus as a cue. None of the drivers verbalized the utility maximization strategy. In addition to verifying the existing candidate strategies, our second goal was to elicit strategies of which we had previously been unaware. Overall, 11 drivers mentioned new cues. These strategies included a target number of trips (1 driver), some form of minimum hourly wage (2 drivers), and some form of minimum number of trips per hour (2 drivers). Five drivers mentioned market saturation as a cue, mostly in the form of long lines of cars at taxi stands. Despite these somewhat scattered responses, it appears that there is no widely used strategy that our analysis ignores.

4.4.2 Empirical Approach

The purpose of this analysis is to identify for each driver the decision model that is most predictive of their observed shift ends. In several respects our analysis differs from much of microeconomic analysis. We therefore present the empirical approach of this analysis in some detail and change the terminology from decision strategies to decision models to underline the descriptive question addressed in this section.

The empirical strategy in this section is an extension of that in the preceding section. As before, we use a competitive test to find the model yielding the best prediction for each shift. These predictions are then aggregated to classify drivers by their overall most predictive model. This approach does not require that the strategies are nested within the same class of models, such as linear models. Instead, models can differ in nature, provided they yield comparable outputs.

Algorithm 4.2 gives an overview of the testing procedure. To be able to classify drivers independently, the competitive test is carried out separately for each driver. As a first step, each shift is predicted by all six models and the best-predicting model is found for each shift. To this end, we employ 10-fold cross-validation. By this procedure, the sample of N shifts is divided into ten randomly composed folds of (roughly) equal size $N_f \approx \frac{N}{10}$ (step 1). All six models are then calibrated based on nine of these folds (step 2). For calibration, we use a derivative-free minimization algorithm (Hooke & Jeeves, 1961) that allows for a search of the best-fitting set of parameters within given bounds. Whereas most parameters are left unconstrained, we constrain the search space for all utility parameters to be positive and the weight of gain-loss utility to be $\eta < 1$. The calibrated models are used to calculate each model's predictions for the tenth fold (steps 3–7). For each model, the mean squared errors in these predictions are recorded (step 8). The procedure is performed ten times, each with a different fold used for prediction. With this procedure, predictions can be calculated for all shifts without fitting.

Calculating the mean squared error for the sequential data is not straightforward. The predicted shift end can easily be computed when it lies within the observed shift. Frequently, however, a given model predicts that the driver continues beyond the observed shift end, and after the shift ends, the models lack the cues they require for a prediction. This is a typical problem of stopping decisions. Our solution exploits the size of the data sample. Although driver i ends shift s, we observe other drivers' trips in the aftermath of s. Although we do not observe the full market, the portion we do observe reflects variation in demand independently of i. In the absence of location data, we assume that each of the subsequent trips taken by competing drivers is equally likely to have been assigned to i, had the latter not ended their shift. Under these assumptions, we construct "extended shifts", which are shifts amended by imputed trips from other drivers.

Specifically, for each shift s we look at the final minute of the

Algorithm 4.2
10-Fold Competitive Test of Models Predicting Shift Ends

```
1  forall shift type do
2      forall market do
3          foreach driver d do
4              1. assign shifts randomly across ten folds;
5              foreach fold f do
6                  2. calibrate all models based on all folds except f;
7                  foreach shift s in f do
8                      3. record end of final observed trip in s as actual end of s;
9                      foreach version v do
10                         foreach model m do
11                             foreach trips t do
12                                 4. use fitted models to calculate prediction;
13                                 if prediction is "continue" then
14                                     5. move to next t;
15                                 else
16                                     if end of t is later than 24 hours after
                                         beginning of s then
17                                         6a. use beginning of s + 24 hours as
                                             predicted end of s;
18                                     else
19                                         6b. record end of t as predicted end
                                             of s for model m;
20                                     end
21                                     7. move to next model;
22                                 end
23                             end
24                             8. calculate residual between predicted and actual
                                 end of s;
25                         end
26                     end
27                     9. calculate root-mean squared residual across all 20 versions;
28                     10. identify best-predicting model by smallest RMSR;
29                 end
30             end
31             11. count for each model number of shifts it predicts best;
32             12. order models by count with m1 predicting most shifts and m6 the
                 fewest;
33             if count of m1 > 1.2 × count of m2 then
34                 13. classify d as using m1;
35             else
36                 14. leave d unclassified;
37             end
38         end
39         15. count number of drivers for each model;
40     end
41 end
```

final trip and find all other drivers who fulfill three conditions: i) They are currently on duty, ii) they are currently without passengers, and iii) they complete at least one more trip during their current shift. Of these drivers, we select one at random and impute their

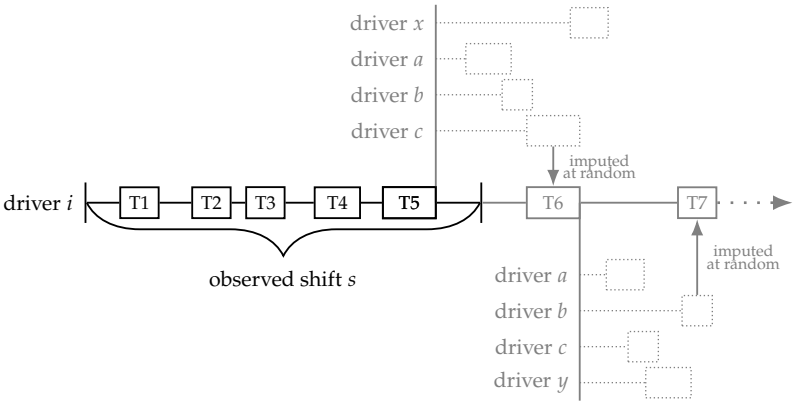

Figure 4.1: *Construction of extended shifts: The shift of driver i lasts from trips T1 to T5; at the end of T5, drivers currently waiting for passengers are found and one of their next trips is chosen at random (here of driver c) and imputed as the next trip of i; at the end of that imputed trip, the procedure is repeated with drivers currently waiting. Note that driver x is busy or off duty at T6 and not considered, whereas driver y is considered at T6 but was previously busy or not on duty.*

next trip as the next trip of s, including trip beginning, trip end, and trip earnings. This procedure is repeated for the final minute of this imputed trip until a total of fifty trips are imputed. The set of drivers for which the trips are imputed varies over time, as different drivers fulfill the above conditions at different points in time. Using this procedure, which is depicted in Figure 4.1, we obtain hypothetical trip sequences beyond observed shift ends. Because each of these sequences consists of randomly imputed trips, the sequences beyond the observed section are random themselves. For this reason, we apply the selection procedure twenty times for each shift. The result is that for each shift s with n observed trips, we obtain twenty versions $s_1, s_2, ..., s_{20}$, each consisting of $n + 50$ trips, the first n of which are equal, whereas the following 50 trips vary. These extended shifts allow us to calculate predictions for all strategies beyond the observed shift end by applying each decision model to all twenty versions of each shift, and averaging the residuals in prediction (step 9). The

Table 4.4
Classifications of Shifts and Drivers

	All Drivers				Single Drivers only			
	Shifts		Drivers		Shifts		Drivers	
Model	Count	Share	Count	Share	Count	Share	Count	Share
neoclassical	69,742	9.4 %	10	0.3 %	23,923	9.6 %	4	0.5 %
reference	57,569	7.7 %	1	<1. %	18,536	7.5 %	1	0.1 %
earnings	193,528	26 %	537	16.5 %	71,208	28.7 %	190	24.2 %
duration	231,738	31.2 %	1,432	44.1 %	70,503	28.4 %	224	28.6 %
clock	147,645	19.9 %	260	8 %	49,265	19.8 %	64	8.2 %
hiatus	43,361	5.8 %	5	0.2 %	15,016	6 %	1	0.1 %
unclassified	—	—	1,004	30.9 %	—	—	300	38.3 %
total	743,583	100 %	3,249	100 %	248,451	100 %	784	100 %

model with the lowest root-mean squared residual is then considered the best predictive model for shift s (step 10).

The results are aggregated to classify drivers by their most predictive decision model. To this end, for each driver d, the number of shifts most accurately predicted by each of the six models is counted. The first model, that is, the model that predicts the simple majority of shifts, is then ranked as the best predicting model for d (steps 11 and 12), provided the simple majority is decisive. To be decisive, the first model needs to predict twenty percent more shifts than the second model. We take this precaution to ensure that drivers for whom two models predict roughly equally well remain unclassified (steps 13 and 14). Finally, classified drivers are counted (step 15).

4.4.3 Classification of Shifts and Drivers

We begin our presentation of results with the classification of shifts and drivers. Table 4.4 gives an overview of these classifications across all four samples we have analyzed separately. Starting with an overview of shift classifications, we then proceed with a discussion of driver classifications.

First, we focus on the classification of shifts. Columns 2 and 3 of Table 4.4 show the number of shifts best predicted by each of the six models, as well as their share. The majority of shifts are best predicted by the duration model, accounting for 31 percent of all shifts, and also by the earnings model, accounting for 26 percent of

shifts[6]. Because many drivers may need to adhere to a shift schedule, we also counted the number of shifts driven by single drivers, who are likely unconstrained by shift schedules. Columns 6 and 7 show that the duration model accounts for a smaller percentage of shifts but also that the majority of single-driver shifts are best predicted by the duration and earnings models.

Of the remaining shifts, most were predicted by the clock model and fewest by the hiatus model. Again, this result holds for both all drivers and the subgroup of single drivers. Like the hiatus model, both utility models account for at most 10 percent of all shifts. Between them, the neoclassical utility model predicts somewhat more shifts than the reference utility model. In contrast to the hiatus and utility models, the clock model typically accounts for about 20 percent of shifts. Overall, it appears that a distinction can be made between the duration, earnings, and clock models jointly predicting around three quarters of shifts and the hiatus and utility models predicting around one quarter of shifts.

Next we examine the classification of drivers. To arrive at these classifications, we calculated for each driver the percentage of shifts best predicted by each of the six models. The vertical bars in Figure 4.2 show for a random sample of 50 drivers the range of these percentages. The letters on top indicate the models predicting the highest number of shifts, which we refer to as the first and second model, respectively. The lower end of the line shows the percentage best predicted by the last model. Differences between these two models are stark and mostly above 25 percentage points.

Drivers were classified by their first model only if it predicted 20 percent — not percentage points — more shifts than the second model; otherwise they were left unclassified. In Figure 4.2, the first 38 drivers were classified according to their first model, whereas the remaining 12 drivers were left unclassified. For these drivers, we deemed the evidence too weak for a classification. In Table 4.4, columns 4 and 5 show the number and share of drivers classified by each model. Around 30 percent of all drivers were left unclassified, with a slightly higher percentage among single drivers.

[6]Note that shift classifications are not independent of drivers: By itself, each shift end is consistent with all six models if model parameters could be chosen freely. However, if each model's parameter values are fixed for each driver, we can identify for each shift the model yielding the best prediction. The counts of these models are presented here.

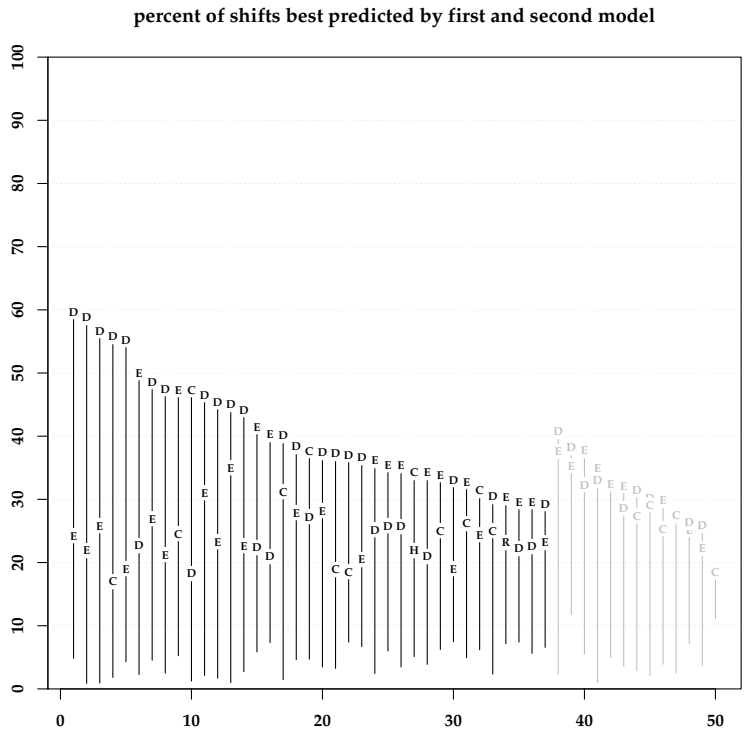

Figure 4.2: *Percentages of shifts predicted by six models for random sample of 50 drivers; letters indicate model with highest and second-highest number of shifts: neoclassical utility (U), reference utility (R), earnings (E), duration (D), clock (C), hiatus (H); drivers classified by their first model shown in black, unclassified drivers in gray.*

Among those drivers we could classify, consider first the earnings and duration models. Again, these two models jointly account for the majority of drivers, including those left unclassified. By these results, the duration model offers the widest description of drivers' shift choices. Consider next the clock model, which accounts for about 10 percent of all drivers. Together with the duration and earnings models, the clock model accounts for around two thirds of drivers. With around 30 percent unclassified, the hiatus model and the utility model each accounts for a fraction of a percent of

drivers. These results hold true in our subsample of single drivers, although the duration model accounts for a smaller percentage of drivers, indicating that part of its descriptive power may be due to shift schedules. Overall, driver classifications therefore mirror the dichotomy of shift classifications, with stronger divergence in the performances of both groups of models.

Table 4.4 shows results aggregated across all drivers, ignoring differences between day and night shifts and between the old and the new market. In Appendix C we produce a more extended version of Table 4.4 that reports results separately for the four samples constructed by crossing shift type and market. Although there are some differences, primarily between day and night drivers, the results do not change qualitatively. In addition, we report more details on the number of shifts consistent with driver classifications, on average between 30 and 40 percent per driver, indicating that shift ends may be best predicted using a combination of models. We therefore examine the second model and find that for most drivers, the majority of shifts are best predicted by the duration and earnings targets but percentages vary. Finally, Appendix B reports histograms of parameter estimates.

4.4.4 Aggregate Outcomes

Given the classifications of shifts and drivers above, we can compare drivers' aggregate outcomes stratified by their best predicting model. We begin with a short discussion of the labor supply elasticity and then turn to average hourly earnings.

To obtain the overall wage elasticity, we follow the IV approach by Farber (2015) and regress `ln(sdur)`, the natural logarithm of shift earnings, on `ln(searn/sdur)`, `rain.d`, `saturday`, and `sunday` separately for day and night shifts. In addition, we add the variable `shock`, a dummy indicating one of the demand shocks listed in Table 4.1. We use other drivers' average hourly earnings on the same day as an instrument for `ln(searn/sdur)`. This approach is similar to Farber's approach, who used a non-overlapping sample to instrument drivers' wages in the remainder of his data set.

The results are shown in Table 4.5, with elasticity estimates of -0.206 and -0.002 for day and night shifts, respectively. For day drivers, we therefore find a negative wage elasticity, despite using an instrumental variable. For night drivers, we find a wage elasticity

Table 4.5
Wage Elasticity of Labor Supply Across All Shifts

	All Drivers			Single Drivers		
Variable	Estimate	SE	p-value	Estimate	SE	p-value
Day Shifts						
intercept	2.797	0.020	<0.001	2.597	0.034	<0.001
log wage	−0.175	0.007	<0.001	−0.074	0.013	<0.001
rainy day	−0.001	0.001	0.493	−0.004	0.002	0.090
Saturday	−0.075	0.003	<0.001	−0.110	0.005	<0.001
Sunday	−0.064	0.002	<0.001	−0.080	0.004	<0.001
shock	−0.011	0.002	<0.001	−0.010	0.003	0.002
Night Shifts						
intercept	2.062	0.018	<0.001	1.856	0.047	<0.001
log wage	0.045	0.006	<0.001	0.150	0.018	<0.001
rainy day	−0.009	0.002	<0.001	−0.013	0.005	0.007
Saturday	−0.037	0.003	<0.001	−0.176	0.008	<0.001
Sunday	0.086	0.003	<0.001	0.016	0.007	0.034
shock	−0.004	0.002	0.072	−0.014	0.006	0.019

very close to zero, implying that drivers on average do not respond to wage increases. The elasticities of single drivers are negative for day shifts and positive for night shifts but small in magnitude.

With heterogeneous drivers, calculating elasticities across all drivers may be misleading and our classification of drivers allows for a more detailed examination. The estimates in Table 4.5 give an aggregate summary across all drivers but also hide a considerable amount of variation between them. To shed light on the relation between best predicting model and demand elasticity, we can use both the classifications of shifts and drivers. The former give a complete picture, whereas the latter offer a clearer picture of the relation between shift termination model and aggregate outcomes. Table 4.6 reports both of these analyses.

Table 4.6
Elasticities And Hourly Earnings by Model

	Across Drivers				Across Shifts			
			Earnings				Earnings	
Model	Drivers	Elasticity	Mean	SD	Shifts	Elasticity	Mean	SD
neoclassical	10	0.25	22.3	7.0	69,742	0.03	19.0	7.9
reference	1	−0.38	24.7	—	57,569	−0.23	19.4	7.4
earnings	537	−0.75	15.8	5.5	193,528	−0.89	17.5	7.1
duration	1,432	0.05	19.5	5.2	231,738	0.02	18.7	8.0
clock	260	0.09	20.6	5.1	147,645	0.11	18.5	7.4
hiatus	5	−0.08	22.7	8.1	43,361	0.03	16.6	9.7

Notes: Across drivers: hourly earnings averaged across consistent shifts and drivers, hourly earnings SD gives SD across drivers, elasticity gives median elasticity estimate from IV regression for each driver; across shifts: hourly earnings averaged across shifts of the same strategy, elasticity obtained from regression using all shifts of the same strategy; for brevity, no separate display for single drivers.

First, we estimate the elasticity separately for each driver by applying the IV approach to those shifts that are consistent with the driver's classification. Column 3 of Table 4.6 reports these results, stratified by model. As expected, the median elasticity of drivers best predicted by the earnings target is negative. In contrast, the median elasticity of drivers best predicted by the duration and clock targets fluctuates around zero and is small in magnitude. This is consistent with expectations, given that shifts in these models are terminated independent of their profitability. Because drivers best predicted by the utility and hiatus models are few in number, we refrain from interpretations of their median elasticities.

In a second analysis, we apply the IV analysis to all shifts, irrespective of driver classifications; the results are shown in column 7 of Table 4.5. Again, we find a negative elasticity for shifts best predicted by the earnings model, whereas those best predicted by the duration and clock models are close to zero on the median. The neoclassical utility model yields an average elasticity parameter of zero, but averages for both utility models differ considerably from the first analysis. We therefore suspect that these results primarily reflect between-driver variability, as these models typically predict only few shifts per driver. For brevity, both of these analyses have ignored differences between day and night shifts and between the old and the new market. Appendix D reports classifications of drivers

separately for these four subsamples.

One final issue concerns the profitability of different models, as measured by the hourly earnings they generated. Mean hourly earnings across drivers are given in columns 4 and 5 of Table 4.6. Drivers best predicted by the earnings model exhibit the lowest mean earnings, whereas those best predicted by the hiatus model exhibit the highest. Columns 8 and 9 of Table 4.6 report means across shifts rather than drivers. By comparison, differences between models are somewhat smaller but the earnings model yields, on average, lower hourly earnings than the duration model. Similarly, the heuristic models appear to yield somewhat lower mean earnings than the utility models. At the same time, these differences are small compared with the variation within each class. For a more detailed illustration, Appendix D reports kernel density plots of mean hourly earnings separately for day and night shifts in the old and new market.

4.4.5 Conclusion: Duration and Earnings Models Most Predictive

The second analysis demonstrates the inability of utility theory to consistently predict drivers' behavior. Although both utility models could predict sizable portions of shift ends, these shifts are spread across many drivers rather than concentrated on drivers who can be described consistently by these models. Instead, those models of drivers as pursuing simple aspirations on earnings or time are best at predicting drivers, despite the fact that these models assume drivers to have constant aspiration levels within each sample of shifts, irrespective of predictable demand shocks. In particular, many drivers are best predicted by some combination of earnings and duration targets. We hypothesize that the individual mix depends on personal circumstances. Across models, wage elasticities, with few exceptions, conformed to expectations. The distributions of average hourly earnings are very similar across models, with heuristic models exhibiting slightly lower modes than utility models in some samples.

4.5 Discussion

This paper set out to predict taxi drivers' earnings and shift ends. For both analyses, we used competitive out-of-sample tests. As we have

argued, this methodology differs from conventional microeconomic analysis but has desirable properties for empirical work. It directly tests models' predictive powers rather than their abilities to adjust to existing data. Furthermore, the competitive approach enables a comparison of different, non-nested models, provided they work on the same input and yield comparable outputs. This methodology has resulted in a picture of taxi drivers' choices that differs from earlier results in several important ways.

Our first analysis found that a variety of statistical models can hardly predict individual drivers' hourly earnings from readily observable variables. Unless we are willing to assume that drivers' predictive talents exceed the power of these models, we may accept the conclusion that drivers' earnings in the next hour appear largely random. These findings contradict those by Farber (2015) and lead to the divergent conclusion that even in reference utility models that set their reference point at the expected level of earnings, there is ample room for the reference point to affect decisions. However, the level of predictability is so low that behavioral models relying on expected earnings, such as neoclassical and reference-dependent utility models, appear unfit for the task at hand. This conclusion is conditioned on the assumption that our set of considered covariates is exhaustive and that drivers do not have access to information beyond these variables. In practice, of course, additional variables may exist that drivers use for making predictions. Such variables include information on the timing of specific events, such as concerts or sports events, or the number of other taxis on shift. That being said, the small progress made with the existing set of variables leaves us skeptical that extensions will lead to improvements so large as to qualitatively change our conclusion that predictions of earnings are difficult in practice and prone to error. Because the merit of utility theory is typically seen in its assumed rationality, our findings furnish no theoretical reasons to assume a priori that utility theory describes drivers' behavior better than any other theory.

Our second analysis found most drivers to be best described by one of three satisficing models that set aspiration levels on earnings, shift duration, or clock time. These aspiration levels are fixed in the sense that drivers may set different aspiration levels for day and night shifts but ignore other factors. Camerer and colleagues (1997) referred to such models as the "strong form of the target income hypothesis," emphasizing their inflexibility. However, given the

small predictive power of observables such as weather or demand shocks, we decided to test models with fixed aspiration levels. These inflexible models yielded considerably better predictions than both utility models we tested.

From these results, we cannot conclude that drivers actually use one model rather than another. Instead, there are two alternatives. First, it is possible that they use a model not tested here. The results of our analysis are namely relative to the selection of candidate models: Although we have made an effort to select promising and relevant candidates, we cannot exclude the existence of better predicting models from the universe of countless possible models. Second, it is theoretically possible that drivers use one of the utility models but their behavior is better predicted by the satisficing models because of the latter's lower exposure to variance. Despite these caveats, our analysis has demonstrated a considerable predictive advantage of satisficing models over utility models.

The results of this study appertain to the Hamburg taxi market, and generalizations must be made with caution. Markets differ along several dimensions, including demand patterns and drivers' goals and working conditions. Our findings indicate that drivers use a set of heuristics that are assumably selected not at random but according to the specifics of each decision environment. For example, Farber (2015) reports positive wage elasticities for the majority of NYC taxi drivers, indicating that differing conditions lead drivers to use different strategies in New York and Hamburg[7]. Therefore, similar analyses of other markets, for taxis or otherwise, would be more appropriate than over-generalizations beyond the domain studied here.

Our findings also imply that positive elasticities are more difficult to attain than is commonly understood. The inaccuracy of predicted earnings virtually prohibits any judgements of future earnings as being worth the time or not. To illustrate, consider a taxi driver on a small island who operates one of few taxis that bring day tourists from the harbor into town and back. The number of tourists varies with a few observables such as weather, day of the week, and time of the year. Because day tourists arrive by ferry, the driver can use the

[7]Similarly, Fehr and Goette (2007) report positive wage elasticities for bike messengers in their experiment. Unlike Hamburg taxi drivers, who spend most of their shifts waiting for passengers, the bike messengers in their study have considerable control over their earnings and the effort they exert.

ferry schedule to predict the timing of passengers across the day and terminate the shift when expected earnings drop too low. Therefore, the driver can focus work on the most profitable hours and attain a positive elasticity. The conditions for Hamburg taxi drivers are quite different. Trips do not necessarily go from harbor to town and vice versa but vary in length and profitability, and timing of passengers cannot be looked up on a schedule. Individual earnings therefore not only depend on overall demand but vary along many factors, including location and the labor supply of other drivers. Predicting market averages therefore helps little in finding the best point during the day to terminate the shift. Although such points may exist, drivers may find themselves unable to identify them.

Nonetheless, positive elasticities are entirely possible. For example, a driver may decide a priori to work three hours a night during the week and six hours on profitable weekend nights. Presumably, this simple strategy generates a positive wage elasticity and illustrates how the wage elasticity is also affected by shift choice and shift beginnings. Indeed, other strategies are conceivable that can generate positive elasticities, such as the hiatus heuristic, which seeks to detect signals of decreasing demand. However, our findings suggest that substantially positive elasticities are difficult to attain based on prediction of future earnings.

The emergent picture of the taxi market is more complicated that initially assumed by Camerer and colleagues. They chose the taxi market for its variability in daily earnings, which could be used to estimate the wage elasticity. However, this plan ignores the fact that the variability in earnings appears to be dominated by factors not readily observable, which makes positive elasticities difficult to attain on the basis of predictions of earnings. Under these circumstances, it appears misguided to assume a positive wage elasticity to be rational and view a negative wage elasticity as evidence against neoclassical theory.

In addition, the focus on elasticities can be inadequate for other reasons. Consider again the driver working three hours on weeknights and six on weekend nights. If the same driver were to stop working on weeknights altogether, the elasticity would likely decrease, seeing as there are fewer comparatively short and unprofitable shifts. At the same time, average wages would likely increase. This discrepancy illustrates that evaluating drivers according to a single indicator can be misleading and does not do justice to the intricacies of preferential

choice. Given the lack of evidence that violations of coherence, the classical criterion for economic rationality, impair choices (Arkes et al., 2016), we advocate a more comprehensive benchmark of ecological rationality that judges decision strategies by their ability to reach defined goals. By this account, an analysis of the rationality of drivers' strategies requires detailed data on drivers' goals. Until such data is available for analysis, assertions of rationality remain speculative.

Finally, our analysis has shown the limits of conventional decision modeling. One of the reasons the income-target hypothesis has been criticized is the counter-intuitive idea that drivers would *prefer* a specific income. On the one hand, why would income beyond a target be less valuable than below the target? On the other hand, the conventional approach models deviations from neoclassical theory through modifications of the preference structure. In this article, we have described an alternative explanation. By this hypothesis, drivers have no particular preference for any given amount of earnings. However, the uncertainty of their decision environment prevents them from calculating optimal paths of action. For that reason, they rely on a "toolbox" of satisficing strategies. The selection of the aspiration variable can be regarded as an expression of which goal the driver deems most important. Yet the aspiration level is not necessarily an expression of preference but the value that yields the best trade-off of time and earnings in the driver's specific decision environment. The empirical challenge lies in understanding the circumstances under which such strategies are rational.

Appendix A Plots of Shift Beginning s

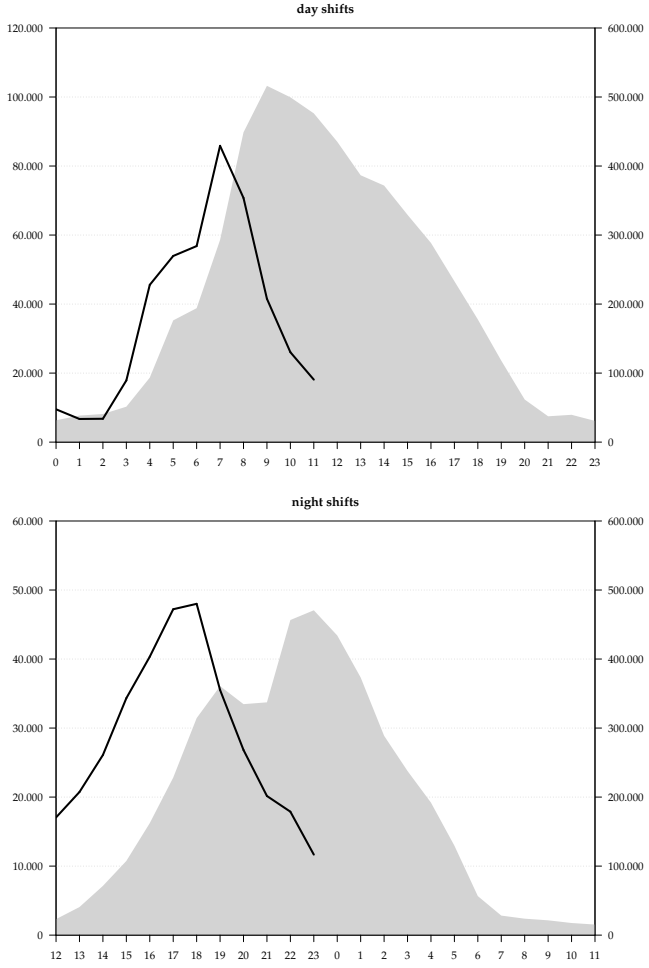

Figure 4.A: *Number of shift beginnings (black, left axis) and trips (gray, right axis) by clock hour.*

Appendix B Parameter Estimates

The classifications reported in Section 4.4 rest crucially on the plausible parametrization of the six models. We therefore include a brief description of the estimated parameter values. Although each model was evaluated for each driver, we examine only drivers we could classify and for each of them, we examine only shifts that are consistent with the driver's classification. These parameters are estimated ten times per driver, once for each of the ten folds across which the driver's shifts were distributed. For each driver we have computed the median parameter value across these ten folds. The distributions of these median values are shown in Figures **??** to **??**, respectively. We display them separately for day and night shifts and for the old and new markets, that is, before and after the fare increase and the introduction of the minimum wage.

In our implementation, neoclassical utility theory has two free parameters that were estimated from the data. First, the disutility of work was restricted in fitting to $0 \leq \theta$ and was estimated for the 6 day drivers in the new market at values between 0.328 and 1.260 and for the 6 drivers in the old market at values between 0.496 and 3.124. Values for night drivers ranged from 0.059 to 1.597 for both markets. Second, the wage elasticity parameter was restricted in fitting to $0 \leq \rho$ and was estimated for most drivers across markets and shift types above one. However, because there were few drivers best described by the neoclassical utility model, these results cannot be used to infer representative parameter estimates. Similarly, there were fewer than 5 drivers best described by the reference utility model across all four samples; we refer the reader to Figures **??** to **??** for distributions of its parameters.

We now turn to the four satisficing models, each of which has one parameter only. For day drivers whose shift ends are best predicted by the earnings model, the parameter estimates follow a bell-shaped distribution, irrespective of old or new market. The central 60 percent of drivers between the 20^{th} and 80^{th} percentiles had estimated targets between 112 EUR and 190 EUR. For night drivers, the distribution was flatter with 60 percent of drivers estimated to have targets between 90 EUR and 197 EUR. Parameter estimates for drivers best predicted by the duration model followed a bell-shaped distribution with 60 percent of estimates between 446 and 603 minutes for day drivers and between 406 and 559 minutes for night drivers, irrespective of

the market. For day drivers best predicted by the clock model, 60 percent of targets were estimated between 1pm and 8pm, with no apparent differences between old and new market. In contrast, 60 percent of night drivers had targets estimated between 9pm and 6am, again with no apparent differences between the two markets. For those day drivers best predicted by the hiatus model, 60 percent of drivers had estimated targets between 47 and 63 minutes. Similarly, 60 percent of night drivers had estimated targets between 34 and 49 minutes..

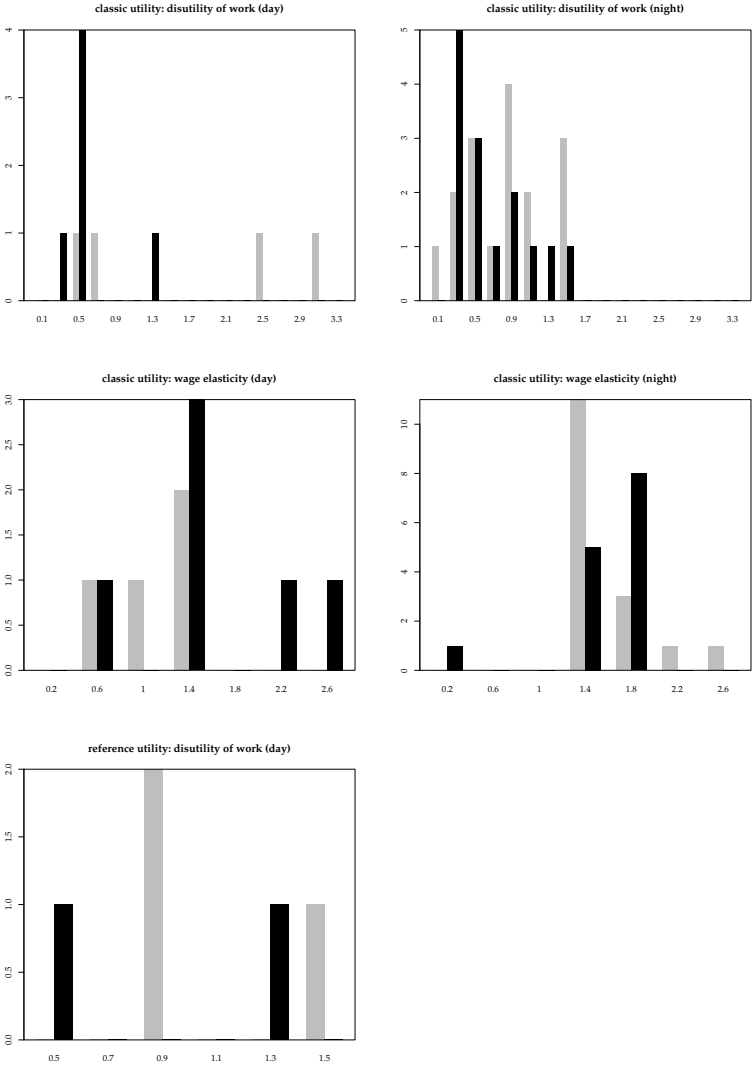

Figure 4.B: *Histograms of median parameter estimates for old market (gray) and new market (black)*

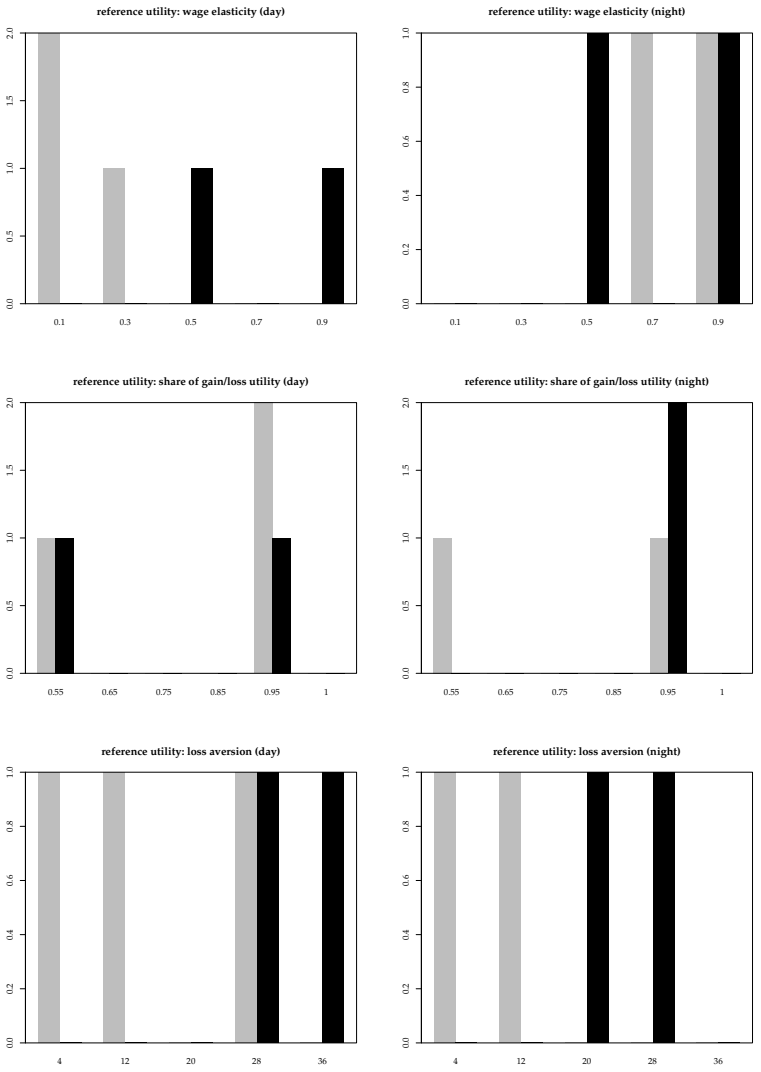

Figure 4.B (cont.): *Histograms of median parameter estimates for old market (gray) and new market (black)*

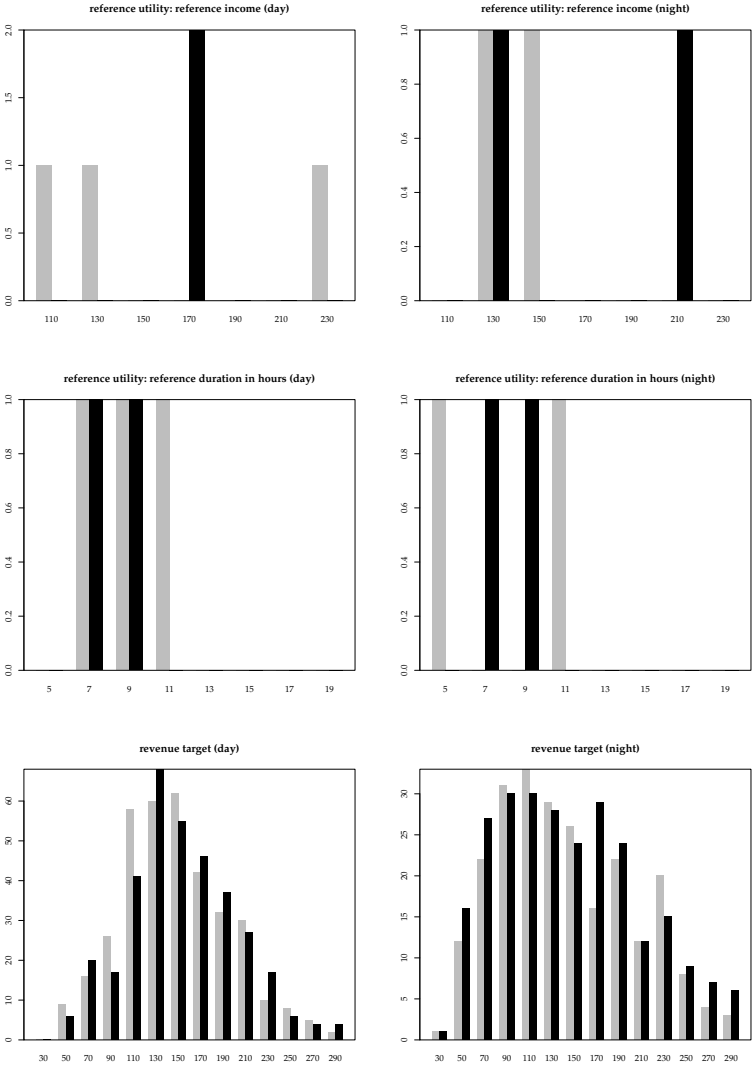

Figure 4.B (cont.): *Histograms of median parameter estimates for old market (gray) and new market (black)*

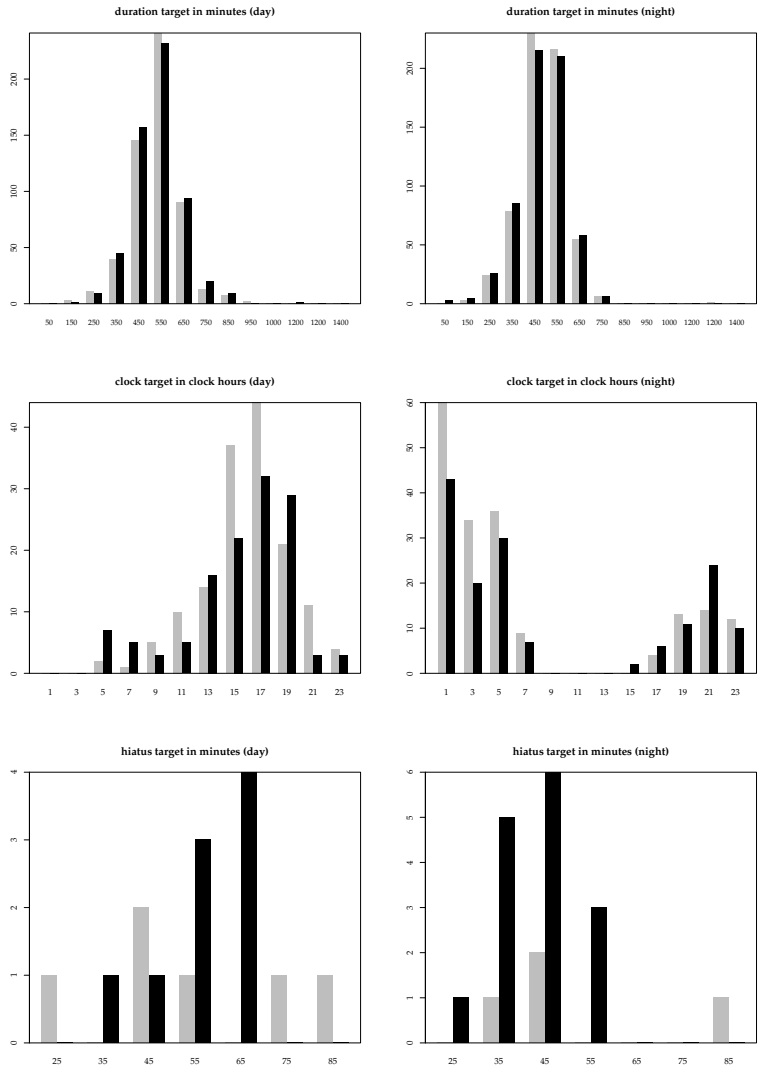

Figure 4.B (cont.): *(Histograms of median parameter estimates for old market (gray) and new market (black)*

Appendix C Shift and Driver Classifications in Detail

Here, we offer more details on the classification of shifts and drivers. Table 4.C gives an extended version of Table 4.4 and reports classifications stratified by shift type and market. Shift classification appear robust across all subsamples, with only minor changes. Driver classifications were carried out separately for each of the four subsamples, implying that drivers may count multiple times if they fall into multiple samples. Unclassified drivers remain stable at around 30 percent. Overall, the duration model accounts for slightly more drivers among night shifts than day shifts, but this conclusion does not hold for single drivers. Difference between the old and the new market are minor.

Table 4.C also shows the absolute number of shifts consistent with their classification. These numbers are considerably lower than the overall absolute number of shifts best predicted by each of the six models, as shown in columns 2 and 5. This discrepancy is not surprising for two reasons. First, for each model, the total number of shifts include shifts of drivers that we could not classify. Second, there is no driver for whom all shifts are best predicted by the same model, so for each driver there are shifts inconsistent with the driver's classification. Indeed, Figure 4.2 illustrated that the first model rarely predicts more than 50 percent of shifts, indicating only a moderate level of consistency among drivers.

To assess consistency, we have added in parentheses to columns 4 and 7 the average of the share of consistent shifts among all shifts. Across all samples of drivers, consistency is around 40 percent for the goal models and between 30 and 35 percent for the utility models. There are several reasons for this finding. First, a larger number of models tested implies that — even by chance — percentages for individual models are reduced. Second, the nature of observational data implies that many personal constraints are unobserved, such as appointments or days with lack of motivation. Such constraints affect shift ends irrespective of the driver's strategy. Third, we hypothesize that many drivers use different strategies in parallel. These drivers predominantly use one strategy but change strategies in some regular fashion. Consider, for instance, a father who needs to pick up his children from sports every Tuesday and Friday at 6pm and is best predicted by the clock model for those days only.

If drivers use strategies in parallel, their second models deserve

Table 4.C
Classifications of Shifts and Drivers By Market and Shift

	All Drivers			Single Drivers only		
		Drivers			Drivers	
Model	Shifts	Count	Consistent	Shifts	Count	Consistent
Day Shifts in Old Market						
neoclassical utility	16,225 (8%)	4 (0%)	168 (35%)	6,336 (9%)	1 (0%)	60 (36%)
reference utility	15,171 (8%)	3 (0%)	93 (29%)	5,625 (8%)	1 (0%)	16 (27%)
earnings	55,124 (28%)	360 (23%)	17,011 (42%)	21,216 (30%)	130 (28%)	7,051 (43%)
duration	61,059 (31%)	554 (36%)	29,795 (42%)	20,906 (29%)	119 (25%)	7,470 (39%)
clock	38,317 (19%)	149 (10%)	7,084 (41%)	13,460 (19%)	40 (9%)	2,412 (38%)
hiatus	10,836 (6%)	6 (0%)	184 (40%)	4,155 (6%)	3 (1%)	121 (38%)
unclassified	—	484 (31%)	—	—	175 (37%)	—
total	196,732	1,560	54,335	71,698	469	17,130
Night Shifts in Old Market						
neoclassical utility	17,142 (11%)	16 (1%)	462 (33%)	4,092 (11%)	6 (2%)	135 (31%)
reference utility	12,092 (8%)	2 (0%)	48 (29%)	2,556 (7%)	2 (1%)	48 (29%)
earnings	36,880 (24%)	239 (16%)	8,875 (43%)	9,762 (27%)	89 (26%)	3,613 (44%)
duration	48,125 (31%)	614 (41%)	27,858 (40%)	9,898 (27%)	85 (25%)	3,916 (37%)
clock	31,955 (21%)	182 (12%)	6,713 (40%)	7,711 (21%)	48 (14%)	1,958 (38%)
hiatus	9,258 (6%)	4 (0%)	96 (35%)	2,350 (6%)	2 (1%)	42 (34%)
unclassified	—	423 (29%)	—	—	105 (31%)	—
total	155,452	1,480	44,052	36,369	337	9,712
Day Shifts in New Market						
neoclassical utility	18,013 (8%)	6 (0%)	410 (34%)	7,920 (9%)	3 (1%)	226 (37%)
reference utility	16,501 (7%)	2 (0%)	49 (30%)	6,840 (7%)	1 (0%)	38 (31%)
earnings	60,450 (27%)	348 (22%)	17,026 (41%)	26,716 (29%)	144 (26%)	8,338 (42%)
duration	70,350 (32%)	568 (36%)	34,965 (42%)	26,959 (29%)	144 (26%)	10,231 (39%)
clock	43,658 (20%)	125 (8%)	6,995 (41%)	18,223 (20%)	45 (8%)	2,904 (40%)
hiatus	12,492 (6%)	9 (1%)	333 (35%)	5,215 (6%)	3 (1%)	161 (37%)
unclassified	—	515 (33%)	—	—	206 (38%)	—
total	221,464	1,573	59,778	91,873	546	21,898
Night Shifts in New Market						
neoclassical utility	18,362 (11%)	14 (1%)	550 (31%)	5,575 (11%)	5 (1%)	205 (32%)
reference utility	13,805 (8%)	2 (0%)	64 (30%)	3,515 (7%)	0 (0%)	0 (0%)
earnings	41,074 (24%)	258 (18%)	10,640 (42%)	13,514 (28%)	102 (26%)	5,194 (44%)
duration	52,204 (31%)	608 (42%)	30,665 (40%)	12,740 (26%)	103 (26%)	5,178 (37%)
clock	33,715 (20%)	153 (10%)	5,986 (39%)	9,871 (20%)	55 (14%)	2,419 (38%)
hiatus	10,775 (6%)	15 (1%)	474 (35%)	3,296 (7%)	6 (2%)	175 (32%)
unclassified	—	414 (28%)	—	—	123 (31%)	—
total	169,935	1,464	48,379	48,511	394	13,171

closer examination. Thus far, our most striking conclusion has been that the majority of drivers are best described by the earnings and duration models. Indeed, of all 1,076 day drivers in the old market we could classify, only 12 had neither of the two models among the first two, and of the 914 best predicted by either of the two models, 624 had the respective other model as their second one. These proportions are similar for the other three samples, although only about half of the night drivers best predicted by either the earnings or duration model

has both model among the first two[8]. Taken together, the first and second model account for about two thirds of the shifts completed by drivers on average.

Next we turn to overlaps between the four samples examined separately so far, beginning with the overlap between day and night drivers. In total,1,054 drivers had sufficiently many day and night shifts to fall into both our samples and for 355 of them, day shifts are best predicted by the duration model. For 147 of these drivers, night shifts were also best predicted by the duration model, with night parameter values on average 4 percent below their day counterparts. For 54 drivers, in contrast, night shifts were best predicted by the earnings model, and for 97 drivers, night shifts could not be classified.

Consider next the overlap between drivers in the old and new markets. In total, 1,411 drivers had a sufficient number of shifts in the old and new market to fall into both of our samples. For 393 of them, shifts in the old market were best predicted by the duration model and for 401 of those, shifts across both markets were best predicted by that model. On average, parameter values for the new market were 0.3 percent lower than for the old market. In contrast, for 52 drivers, night shifts were best predicted by the earnings model and for 148 drivers, night shifts could not be classified.

[8]Specifically, for day drivers in the new market/night drivers in the old market/night drivers in the new market, we find that 22/18/20 drivers out of 1,058/ 1,057/ 1,050 had neither the duration nor the earnings model among the first two and of the 916/853/866 drivers best predicted by either the earnings or duration model, 594/434/474 drivers had both models among their first two.

Appendix D Aggregate Outcomes in Detail

Table 4.D is an extension of Table 4.6 and reports elasticity estimates and mean earnings stratified by shift type and market. Differences between subsamples are large for the utility and hiatus models that account for very few drivers only. For the three most predictive models, median elasticity estimates of drivers are fairly stable: The median for drivers best predicted by the earnings target consistently falls between −1 and −1 but fluctuates around zero for the duration and clock models. These conclusions also hold for elasticity estimates across shifts. For the earnings model, we find a moderate increase from the old market to the new and for the duration model, we find slightly more positive elasticities for night than for day shifts, but these remain close to zero. Estimates for the utility and the hiatus models fluctuate considerably, strengthening our suspicion that these results are strongly affected by between-driver variance.

Table 4.D also reports mean hourly earnings across the four subsamples. Overall, hourly earnings tend to be higher during night shifts than day shifts. Nonetheless, we find only minor differences between mean earnings of different models — between day and night shifts and between the old and the new market. As before, the earnings model consistently exhibits the lowest mean hourly earnings among the three best predicting models, sometimes the lowest overall. Calculated across shifts, the utility models exhibit somewhat higher earnings than other models, but standard deviations remain large. For a better illustration of differences, Figure 4.D shows the kernel density estimates of the distribution of earnings across drivers predicted by the earnings and duration models in gray and of those predicted by the utility model with highest mean earnings in black. Across all four subsamples, differences between the models exist but appear small in comparison with the variance within each group of drivers.

Table 4.D
Elasticities And Hourly Earnings by Model By Market and Shift

	Across Drivers				Across Shifts			
			Earnings				Earnings	
Model	Drivers	Elasticity	Mean	SD	Shifts	Elasticity	Mean	SD
Day Shifts in Old Market								
neoclassical utility	4	−0.27	18.5	3.0	16,225	−0.23	16.0	6.9
reference utility	3	−0.90	15.3	5.0	15,171	−0.32	16.5	6.4
earnings target	360	−0.66	15.1	5.1	55,124	−0.72	16.0	6.3
duration target	554	0.00	16.0	4.8	61,059	−0.01	15.8	6.6
clock target	149	0.00	18.3	4.7	38,317	0.04	16.3	6.4
hiatus target	6	−0.04	16.3	4.0	10,836	0.05	14.9	7.4
Night Shifts in Old Market								
neoclassical utility	16	0.19	25.1	3.2	17,142	0.27	19.9	7.5
reference utility	2	2.06	13.1	4.6	12,092	0.00	20.4	7.2
earnings target	239	−0.72	14.1	5.5	36,880	−0.68	17.6	7.2
duration target	614	0.11	20.5	5.0	48,125	0.14	19.7	8.3
clock target	182	0.26	19.9	5.0	31,955	0.39	19.2	7.4
hiatus target	4	1.30	19.1	6.5	9,258	0.67	16.6	7.9
Day Shifts in New Market								
neoclassical utility	6	1.15	21.6	6.4	18,013	0.03	18.1	7.5
reference utility	2	−0.38	21.4	—	16,501	−0.38	18.6	6.8
earnings target	348	−0.86	16.6	6.0	60,450	−1.01	17.4	6.7
duration target	568	−0.01	18.5	4.8	70,350	0.01	17.9	6.9
clock target	125	0.09	20.5	5.8	43,658	0.13	17.9	6.8
hiatus target	9	−1.40	17.6	3.8	12,492	0.07	16.3	7.5
Night Shifts in New Market								
neoclassical utility	14	0.43	23.2	6.7	18,362	0.33	21.7	8.3
reference utility	2	0.60	20.5	—	13,805	−0.04	22.6	8.0
earnings target	258	−0.83	17.1	7.1	41,074	−0.95	19.6	8.0
duration target	608	0.11	23.0	5.3	52,204	0.14	22.3	8.9
clock target	153	0.03	21.2	5.8	33,715	0.32	21.1	8.2
hiatus target	15	0.01	20.5	5.7	10,775	0.23	18.8	14.0

Notes: Across drivers: hourly earnings averaged across consistent shifts and drivers, hourly earnings SD gives SD across drivers, elasticity gives median elasticity estimate from IV regression for each driver; across shifts: hourly earnings averaged across shifts of the same strategy, elasticity obtained from regression using all shifts of the same strategy; for brevity, no separate display for single drivers.

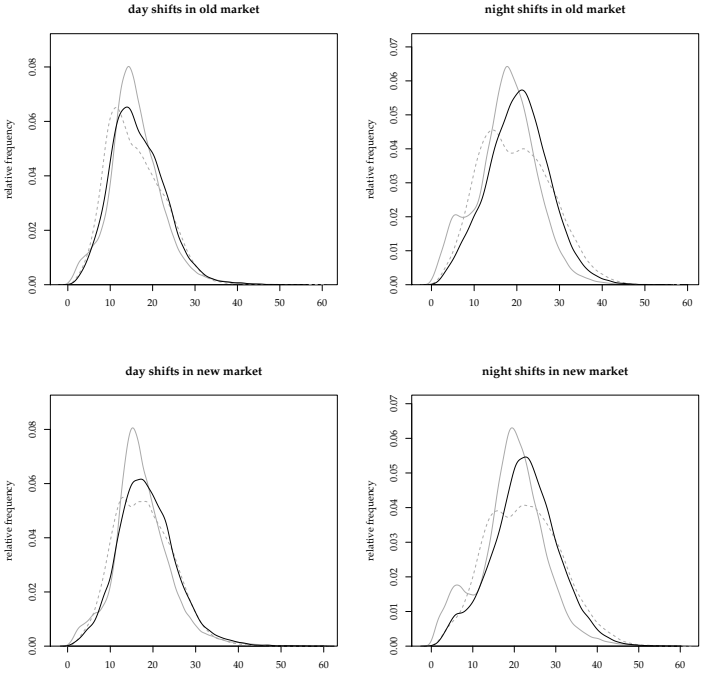

Figure 4.D: *Plots of kernel density estimates for distribution of mean hourly shift earnings for earnings model (gray, solid), duration model (gray, dashed), and the utility model with highest mean earnings (black); includes all shifts, rare outliers above 60 EUR excluded.*

References

Åstebro, T., & Elhedhli, S. (2006). The Effectiveness of Simple Decision Heuristics: Forecasting Commercial Success for Early-Stage Ventures. *Management Science*, *52*(3), 395–409. doi: 10.1287/mnsc.1050.0468

Aikman, D., Galesic, M., Gigerenzer, G., Kapadia, S., Katsikopoulos, K. V., Kothiyal, A., ... Neumann, T. (2020). Taking uncertainty seriously: simplicity versus complexity in financial regulation. *Industrial and Corporate Change*.

Akerlof, G. (1970). The market for 'lemons': Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, *84*(3), 488–500. doi: 10.2307/1879431

Aleskerov, F., Bouyssou, D., & Monjardet, B. (2007). *Utility Maximization, Choice and Preference* (2nd ed.). Berlin: Springer.

Amitani, Y. (2015). The natural frequency hypothesis and evolutionary arguments. *Mind & Society*, *14*(1), 1-19. doi: 10.1007/s11299-014-0155-7

Anscombe, F. J., & Aumann, R. J. (1963). A definition of subjective probability. *Annals of mathematical statistics*, *34*(1), 199–205. doi: 10.1214/aoms/1177704255

Antes, G. (n.d.). *former director of cochrane germany.* (personal communication)

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, *3*(1), 20.

Artinger, F., & Gigerenzer, G. (2016). Heuristic Pricing in an Uncertain Market: Ecological and Constructivist Rationality. *working paper*. doi: 10.2139/ssrn.2938702

Artinger, F., Gigerenzer, G., & Jacobs, P. (2020a). How do taxi drivers

terminate their shifts when earnings are hard to predict? *In this thesis*.

Artinger, F., Gigerenzer, G., & Jacobs, P. (2020b). Satisficing: Integrating two traditions. *In this thesis*.

Artinger, F., Kozodi, N., Wangenheim, F., & Gigerenzer, G. (2018). Recency: Prediction with smart data. In *American marketing association winter conference proceedings* (p. 29:L–2).

Artinger, F., Petersen, M., Gigerenzer, G., & Weibler, J. (2015). Heuristics as adaptive decision strategies in management. *Journal of Organizational Behavior*, *36*(S1), S33–S52.

Audia, P. G., & Greve, H. R. (2006). Less Likely to Fail: Low Performance Firm Size, and Factory Expansion in the Shipbuilding Industry. *Management Science*, *52*(1), 83–94. doi: 10.1287/mnsc.1050.0446

Audia, P. G., Locke, E. A., & Smith, K. G. (2000). The paradox of success: An archival and a laboratory study of strategic persistence following radical environmental change. *Academy of Management Journal*, *43*(5), 837–853. doi: 10.2307/1556413

Ayal, S., & Beyth-Marom, R. (2014). The effects of mental steps and compatibility on Bayesian reasoning. *Judgment and Decision Making*, *9*, 226-242.

Barbey, A. K., & Sloman, S. A. (2007). Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences*, *30*, 241-297. doi: 10.1017/S0140525X07001653

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*, 211-233. doi: 10.1016/0001-6918(80)90046-3

Bar-Hillel, M. (1984). Representativeness and fallacies of probability judgment. *Acta Psychologica*, *55*, 91-107. doi: 10.1016/0001-6918(84)90062-3

Barton, A., Mousavi, S., & Stevens, J. R. (2007). A statistical taxonomy and another “chance" for natural frequencies. *Behavioral and Brain Sciences*, *30*, 255-256. doi: 10.1017/S0140525X07001665

Baum, J. A. C., Rowley, T. J., Shipilov, A. V., & Chuang, Y.-T. (2005). Dancing with strangers: Aspiration performance and the search for underwriting syndicate partners. *Administrative Science Quarterly*, *50*(4), 536–575. doi: 10.2189/asqu.50.4.536

Bearden, J. N., & Connolly, T. (2007). Multi-attribute sequential search. *Organizational Behavior and Human Decision Processes*, *103*(1), 147–158. doi: 10.1016/j.obhdp.2006.10.006

Ben-Shlomo, Y., Collin, S. M., Quekett, J., Sterne, J. A., & Whiting, P. (2015). Presentation of diagnostic information to doctors may change their interpretation and clinical management: A web-based randomised controlled trial. *PLoS One*, *10*(7), e0128637. doi: 10.1371/journal.pone.0128637

Berg, N. (2014). Success from satisficing and imitation: Entrepreneurs' location choice and implications of heuristics for local economic development. *Journal of Business Research*, *67*(8), 1700–1709. doi: 10.1016/j.jbusres.2014.02.016

Berg, N., Biele, G., & Gigerenzer, G. (2016). Consistent bayesians are no more accurate than non-bayesians: economists surveyed about psa. *Review of Behavioral Economics*, *3*(2), 189–219.

Bernoulli, D. (1738/1954). Exposition of a new theory on the measurement of risk. *Econometrics*, *22*(1), 23-36. doi: 10.2307/1909829

Biernaskie, J. M., Walker, S. C., & Gegear, R. J. (2009). Bumblebees learn to forage like Bayesians. *The American Naturalist*, *174*(3), 413-423. doi: 10.1086/603629

Binder, K., Krauss, S., & Bruckmaier, G. (2015). Effects of visualizing statistical information: An empirical study on tree diagrams and 2×2 tables. *Frontiers in Psychology*, *6*, 1-9. doi: 10.3389/fpsyg.2015.01186

Blettner, D. P., He, Z.-L., Hu, S., & Bettis, R. A. (2015). Adaptive aspirations and performance heterogeneity: Attention allocation among multiple reference points. *Strategic Management Journal*, *36*(7), 987–1005. doi: 10.1002/smj.2260

Böcherer-Linder, K., & Eichler, A. (2016). The impact of visualizing nested sets: An empirical study on tree diagrams and unit squares. *Frontiers in Psychology*, *7*, 2026. doi: 10.3389/fpsyg.2016.02026

Bolton, M. K. (1993). Organizational innovation and substandard performance: when is necessity the mother of innovation? *Organization Science*, *4*(1), 57–75. doi: 10.1287/orsc.4.1.57

Borch, K. (1968). Decision rules depending on the probability of ruin. *Oxford Economic Papers*, *20*(1), 1–10. doi: 10.1093/oxfordjournals.oep.a041075

Bramwell, R., West, H., & Salmon, P. (2006). Health professionals' and service users' interpretation of screening test results: Experimental study. *BMJ*, *333*, 284–286A. doi: 10.1136/bmj.38884.663102.AE

Brase, G. L. (2007). Omissions, conflations, and false dichotomies: Conceptual and empirical problems with the Barbey & Sloman account. *Behavioral and Brain Sciences*, *30*, 258-259. doi: 10.1017/S0140525X07001690

Brase, G. L. (2008). Frequency interpretation of ambiguous statistical information facilitates Bayesian reasoning. *Psychonomic Bulletin & Review*, *15*, 284-289. doi: 10.3758/PBR15.2.284

Brase, G. L. (2009a). How different types of participant payments alter task performance. *Judgment and Decision Making*, *4*(5), 419-428.

Brase, G. L. (2009b). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, *23*, 369-381. doi: 10.1002/acp.1460

Brase, G. L. (2014). The power of representation and interpretation: Doubling statistical reasoning performance with icons and frequentist interpretations of ambiguous numbers. *Journal of Cognitive Psychology*, *26*, 81-97. doi: 10.1080/20445911.2013.861840

Brase, G. L., Fiddick, L., & Harries, C. (2006). Participant recruitment methods and statistical reasoning performance. *Quarterly Journal of Experimental Psychology*, *59*, 965-976. doi: 10.1080/02724980543000132

Brase, G. L., & Hill, W. T. (2015). Good fences make for good neighbors but bad science: A review of what improves Bayesian reasoning and why. *Frontiers in Psychology*, *6*, 1-9. doi: 10.3380/fpsyg.2015.00340

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231. doi: 10.1214/ss/1009213726

Brighton, H. (2020). Statistical Foundations of Ecological Rationality. *Economics: The Open-Access, Open-Assessment E-Journal*, *14*(2020-2), 1–32. doi: 10.5018/economics-ejournal.ja.2020-2

Brighton, H., & Gigerenzer, G. (2008). Bayesian brains and cognitive mechanisms: Harmony or dissonance. *The probabilistic mind: Prospects for Bayesian cognitive science, ed. N. Chater & M. Oaksford*, 189–208. doi: 10.1093/acprof:oso/9780199216093.003.0009

Brighton, H., & Gigerenzer, G. (2012). Are rational actor models "rational" outside small worlds. *Evolution and Rationality: Decisions, Co-operation, and Strategic Behavior*, 84–109.

Brighton, H., & Gigerenzer, G. (2015). The bias bias. *Journal of Business Research*, *68*(8), 1772–1784. doi: 10.1016/j.jbusres.2015.01.061

Brown, D. B., De Giorgi, E., & Sim, M. (2012). Aspirational Preferences and Their Representation by Risk Measures. *Management Science*, *58*(11), 2095–2113. doi: 10.1287/mnsc.1120.1537

Brown, D. B., & Sim, M. (2009). Satisficing measures for analysis of risky positions. *Management Science*, *55*(1), 71–84. doi: 10.1287/mnsc.1080.0929

Bunn, F., Trivedi, D., Alderson, P., Hamilton, L., Martin, A., & Iliffe, S. (2014). The impact of cochrane systematic reviews: a mixed method evaluation of outputs from cochrane review groups supported by the uk national institute for health research. *Systematic reviews*, *3*(1), 125. doi: 10.1186/2046-4053-3-125

Camerer, C., Babcock, L., Loewenstein, G., & Thaler, R. (1997). Labor supply of New York City cabdrivers: one day at a time. *Quarterly Journal of Economics*, *112*(2), 407–441. doi: 10.1162/003355397555244

Caplin, A., & Dean, M. (2011). Search, choice, and revealed preference. *Theoretical Economics*, *6*(1), 19–48. doi: 10.3982/TE592

Caplin, A., & Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, *105*(7), 2183–2203. doi: 10.1257/aer.20140117

Caplin, A., Dean, M., & Martin, D. (2011). Search and Satisficing. *American Economic Review*, *101*(7), 2899–2922. doi: 10.1257/aer.101.7.2899

Central Intelligence Agency. (2020). *World factbook, field listing: languages.* `https://www.cia.gov/library/publications/resources/the-world-factbook/fields/402.html`. (Accessed: April 2020)

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*, *140*(4), 980-1008. doi: 10.1037/a0035661

Chapman, G. B., & Liu, J. (2009). Numeracy, frequency, and Bayesian reasoning. *Judgment and Decision Making*, *4*, 34-40.

Charnes, A., & Cooper, W. W. (1963). Deterministic Equivalents for Optimizing and Satisficing under Chance Constraints. *Operations Research*, *11*(1), 18–39. doi: 10.1287/opre.11.1.18

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811-23. doi: 10.1002/wcs.79

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models

of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, *10*(7), 287-91. doi: 10.1016/j.tics.2006.05.007

Chetty, R., Looney, A., & Kroft, K. (2009). Salience and taxation: Theory and evidence. *American Economic Review*, *99*(4), 1145–1177. doi: 10.1257/aer.99.4.1145

Chou, Y. K. (2002). Testing alternative models of labour supply: evidence from taxi drivers in Singapore. *Singapore Economic Review*, *47*(1), 17–47. doi: 10.1142/S0217590802000389

Chow, G. (1989). Rational versus adaptive expectations in present value models. *The Review of Economics and Statistics*, *71*(3), 376–84. doi: 10.2307/1926893

Cole, W. G. (1988). Three graphic representations to aid Bayesian inference. *Methods of Information in Medicine*, *27*, 125-132. doi: 10.1055/s-0038-1635532

Conlisk, J. (1996). Why bounded rationality? *Journal of Economic Literature*, *34*, 669–700.

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. doi: 10.1016/0010-0277(95)00664-8

Cosmides, L., & Tooby, J. (2008). Can a general deontic logic capture the facts of human moral reasoning? How the mind interprets social exchange rules and detects cheaters. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (p. 53-119). Cambridge, MA: MIT Press.

Crawford, V. P. (2013). Boundedly rational versus optimization-based models of strategic thinking and learning in games. *Journal of Economic Literature*, *51*(2), 512–27. doi: 10.1257/jel.51.2.512

Crawford, V. P., & Meng, J. (2011). New York City Cab Drivers' Labor Supply Revisited: Reference-Dependent Preferences with Rational-Expectations Targets for Hours and Income. *American Economic Review*, *101*, 1912–1932.

Cyert, R. M., & March, J. G. (1963). *A Behavioral Theory of the Firm*. Englewood Cliffs: Prentice-Hall.

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In *Simple heuristics that make us smart* (pp. 97–118). New York: Oxford University Press.

de Boer, L., Gaytan, J., & Arroyo, P. (2006). A satisficing model of outsourcing. *Supply Chain Management: An International Journal*, *11*(5), 444-455. doi: 10.1108/13598540610682462

DeGroot, M. H. (1970). *Optimal statistical decisions*. New York: McGraw-Hill.

Dembo, T. (1931). Der Ärger als dynamisches Problem. *Psychologische Forschung*, *15*, 1–144. doi: 10.1007/BF00406043

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, *22*(5), 1915–1953. doi: 10.1093/rfs/hhm075

Diecidue, E., & van de Ven, J. (2008). Aspiration level, probability of success and failure, and expected utility. *International Economic Review*, *49*(2), 683–700. doi: 10.1111/j.1468-2354.2008.00494.x

Domurat, A., Kowalczuk, O., Idzikowska, K., Borzymowska, Z., & Nowak-Przygodzka, M. (2015). Bayesian probability estimates are not necessary to make choices satisfying Bayes' rule in elementary situations. *Frontiers in Psychology*, *6*, 1-14. doi: 10.3389/fpsyg.2015.01194

Dosi, G., Nelson, R. R., & Winter, S. G. (2000). *The nature and dynamics of organizational capabilities* (G. Dosi, R. R. Nelson, & S. G. Winter, Eds.). Oxford: Oxford University Press. doi: 10.1093/0199248540.001.0001

Dudey, T., & Todd, P. M. (2001). Making good decisions with minimal information: Simultaneous and sequential choice. *Journal of Bioeconomics*, *3*(2-3), 195—-215. doi: 10.1023/A:1020542800376

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. doi: 10.1111/j.0006-341X.2000.00455.x

DWD Climate Data Center. (2018a). *Historical daily station observations (temperature, pressure, precipitation, sunshine duration, etc.) for Germany.* (version v006)

DWD Climate Data Center. (2018b). *Historical hourly station observations of precipitation for Germany.* (version v006)

Dzyabura, D., & Hauser, J. R. (2011). Active machine learning for consideration heuristics. *Marketing Science*, *30*(5), 801–819. doi: 10.1287/mksc.1110.0660

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (p. 249-267). Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511809477.019

Ellsberg, D. (1961). Risk, ambiguity, and the savage axioms. *Quarterly Journal of Economics*, 643–669. doi: 10.2307/1884324

Etner, J., Jeleva, M., & Tallon, J.-M. (2012). Decision theory under ambiguity. *Journal of Economic Surveys*, *26*(2), 234–270. doi: 10.1111/j.1467-6419.2010.00641.x

Evans, J. S. B. T., & Elqayam, S. (2007). Dual-processing explains base-rate neglect, but which dual-process theory and how? *Behavioral and Brain Sciences*, *30*, 261-262. doi: 10.1017/S0140525X07001720

Evans, J. S. B. T., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, *77*, 197-213. doi: 10.1016/S0010-0277(00)00098-6

Farber, H. S. (2005). Is tomorrow another day? The labor supply of New York City cabdrivers. *Journal of Political Economy*, *113*(1), 46–82. doi: 10.1086/426040

Farber, H. S. (2008). Reference-dependent preferences and labor supply: the case of New York City taxi drivers. *American Economic Review*, *98*(3), 1069–1082. doi: 10.1257/aer.98.3.1069

Farber, H. S. (2015). Why you can't find a taxi in the rain and other labor supply lessons from cab drivers. *Quarterly Journal of Economics*, *130*(4), 1975–2026. doi: 10.3386/w20604

Fehr, E., & Goette, L. (2007). Do workers work more if wages are high? Evidence from a randomized field experiment. *American Economic Review*, *97*(1), 298–317. doi: 10.1257/aer.97.1.298

Ferguson, E., & Starmer, C. (2013). Incentives, expertise, and medical decisions: Testing the robustness of natural frequency framing. *Health Psychology*, *32*, 967-977. doi: 10.1037/a0033720

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. doi: 10.1037/0096-3445.129.3.399

Fiedler, K., & von Sydow, M. (2015). Heuristics and biases: Beyond Tversky and Kahneman's (1974) judgment under uncertainty. In M. W. Eysenck & D. Groome (Eds.), *Cognitive psychology: Revisiting the classical studies* (p. 146-161). Los Angeles, CA: Sage.

Fishburn, P. C. (1977). Mean-Risk Analysis with Risk Associated with Below-Target Returns. *American Economic Review*, *67*(2),

116–126.

Fontanari, L., Gonzalez, M., Vallortigara, G., & Girotto, V. (2014). Probabilistic cognition in two indigenous mayan groups. *Proceedings of the National Academy of Sciences*, *111*, 17075-17080. doi: 10.1073/pnas.1410583111

Friederichs, H., Ligges, S., & Weissenstein, A. (2014). Using tree diagrams without numerical values in addition to relative numbers improves students' numeracy skills: A randomized study in medical education. *Medical Decision Making*, *34*, 253-257. doi: 10.1177/0272989X13504499

Friedman, D. (1998). Monty Hall's Three Doors: Construction and Deconstruction of a Choice Anormaly. *American Economic Review*, *88*(4), 933–946.

Friedman, D., Pommerenke, K., Lukose, R., Milam, G., & Huberman, B. a. (2007). Searching for the sunk cost fallacy. *Experimental Economics*, *10*(1), 79–104. doi: 10.1007/s10683-006-9134-0

Friedman, M. (1953). *Essays in Positive Economics*. Chicago: The University Press of Chicago.

Gabaix, X. (2014). A sparsity-based model of bounded rationality. *Quarterly Journal of Economics*, *129*(4), 1661–1710. doi: 10.1093/qje/qju024

Gabaix, X., Laibson, D., Moloche, G., & Weinberg, S. (2006). Costly information acquisition: Experimental analysis of a boundedly rational model. *American Economic Review*, *96*(4), 1043–1068. doi: 10.1257/aer.96.4.1043

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28*(2), 210-216. doi: 10.1037/a0014474

Galesic, M., Gigerenzer, G., & Straubinger, N. (2009). Natural frequencies help older adults and people with low numeracy to evaluate medical screening tests. *Medical Decision Making*, *29*, 368-371. doi: 10.1177/0272989X08329463

Gallagher, J. (2014). Learning about an infrequent event: evidence from flood insurance take-up in the united states. *American Economic Journal: Applied Economics*, 206–233. doi: 10.1257/app.6.3.206

Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids. *Current Directions in Psychological Science*, *22*(5), 392-399. doi: 10.1177/0963721413491570

Garcia-Retamero, R., & Hoffrage, U. (2013). Visual representation of statistical information improves diagnostic inferences in doctors and their patients. *Social Science & Medicine*, *83*, 27-33. doi: 10.1016/j.socscimed.2013.01.034

Gardner, J. W. (1940). The use of the term 'Level of Aspiration'. *Psychological Review*, *47*(1), 59–68. doi: 10.1037/h0059521

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability* (Vol. 174). Freeman.

Gavetti, G., Greve, H. R., Levinthal, D., & Ocasio, W. (2012). The Behavioral Theory of the Firm: Assessment and Prospects. *The Academy of Management Annals*, *6*(1), 1–40.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, *4*, 1–58. doi: 10.1162/neco.1992.4.1.1

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273–278. doi: 10.1126/science.aac6076

Gigerenzer, G. (1996a). On narrow norms and vague heuristics: A reply to Kahneman and Tversky. *Psychological Review*, *103*, 592-596. doi: 10.1037/0033-295X.103.3.592

Gigerenzer, G. (1996b). Why do frequency formats improve Bayesian reasoning? Cognitive algorithms work on information, which needs representation. *Behavioral and Brain Sciences*, *19*, 23-24. doi: 10.1017/S0140525X00041248

Gigerenzer, G. (1998). Ecological intelligence: An adaptation for frequencies. In D. D. Cummins & C. Allen (Eds.), *The evolution of mind* (p. 9-29). New York, NY: Oxford University Press.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. Retrieved from `http://linkinghub.elsevier.com/retrieve/pii/S1053535704000927` doi: 10.1016/j.socec.2004.09.033

Gigerenzer, G. (2015). On the supposed evidence for libertarian paternalism. *Review of Philosophy and Psychology*, *6*(3), 361-383. doi: 10.1007/s13164-015-0248-1

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3-4), 303–336. doi: 10.1561/105.00000092

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics In Cognitive Science*, *1*,

107-143. doi: 10.1111/j.1756-8765.2008.01006.x

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. doi: 10.1111/j.1539-6053.2008.00033.x

Gigerenzer, G., & Goldstein, D. G. (2011). The recognition heuristic: A decade of research. *Judgment and Decision Making*, *6*(1), 100-121.

Gigerenzer, G., Hertwig, R., & Pachur, T. (2011). *Heuristics: The foundations of adaptive behavior.* (G. Gigerenzer, R. Hertwig, & T. Pachur, Eds.). Oxford University Press. doi: 10.1093/acprof:oso/9780199744282.001.0001

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*, 684-704. doi: 10.1037/0033-295X.102.4.684

Gigerenzer, G., & Hoffrage, U. (2007). The role of representation in Bayesian reasoning: Correcting common misconceptions. *Behavioral and Brain Sciences*, *30*, 264-267. doi: 10.1017/S0140525X07001756

Gigerenzer, G., & Selten, R. (2001). Rethinking Rationality. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality - the adaptive toolbox* (pp. 1–12). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/1654.001.0001

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance: How probability theory changed science and everyday life*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511720482

Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.

Gigerenzer, G., Todd, P. M., & The ABC Research Group. (2012). *Ecological Rationality - Intelligence in the World*. New York: Oxford University Press.

Gilbert, J. P., & Mosteller, F. (1966). Recognizing the maximum of a sequence. *Journal of the American Statistical Association*, *61*, 35–73. doi: 10.1080/01621459.1966.10502008

Gilboa, I., & Schmeidler, D. (1995). Case-based decision theory. *Quarterly Journal of Economics*, *110*(3), 605–639. doi: 10.2307/2946694

Gilboa, I., & Schmeidler, D. (2001). Reaction to price changes and aspiration level adjustments. *Review of Economic Design*, *6*(2), 215–223. doi: 10.1007/PL00013704

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511808098

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, *78*, 247-276. doi: 10.1016/S0010-0277(00)00133-5

Girotto, V., & Gonzalez, M. (2002). Chances and frequencies in probabilistic reasoning: Rejoinder to hoffrage, gigerenzer, krauss, and martignon. *Cognition*, *84*, 353-359. doi: 10.1016/S0010-0277(02)00051-3

Girotto, V., & Pighin, S. (2015). Basic understanding of posterior probability. *Frontiers in Psychology*, *6*, 680. doi: 10.3389/fpsyg.2015.00680

Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*(2), 148–164. doi: 10.1111/j.2517-6161.1979.tb01068.x

Gode, D. K., & Sunder, S. (1993). Allocative Efficiency of Markets with Zero-Intelligence Traders: Market as a Partial Substitute for Individual Rationality. *Journal of Political Economy*, *101*(1), 119. doi: 10.1086/261868

Goldstein, D. G., McAfee, R. P., Suri, S., & Wright, J. R. (2020). Learning when to stop searching. *Management Science*, *66*(3), 1375–1394. doi: 10.1287/mnsc.2018.3245

González-Valdés, F., & de Dios Ortuzar, J. (2018). The stochastic satisficing model: A bounded rationality discrete choice model. *Journal of Choice Modelling*, *27*, 74–87. doi: 10.1016/j.jocm.2017.11.002

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology*, *37*, 620-629. doi: 10.1037//0012-1649.37.5.620

Greve, H. R. (1998). Performance aspirations, and risky change organizational change. *Administrative Science Quarterly*, *43*(1), 58–86. doi: 10.2307/2393591

Greve, H. R. (2008). A behavioral theory of firm growth: Sequential attention to size and performance goals. *Academy of Management Journal*, *51*(3), 476–494. doi: 10.5465/amj.2008.32625975

Grüne-Yanoff, T., & Hertwig, R. (2016). Nudge versus boost: How coherent are policy and theory? *Minds and Machines*, *26*(1),

149-183. doi: 10.1007/s11023-015-9367-9

Güth, W. (2010). Satisficing and (un)bounded rationality—A formal definition and its experimental validity. *Journal of Economic Behavior & Organization*, *73*(3), 308–316. doi: 10.1016/j.jebo.2010.01.003

Güth, W., Levati, M. V., & Ploner, M. (2010). Satisficing in strategic environments: a theoretical approach and experimental evidence. *The Journal of Socio-Economics*, *39*(5), 554–561. doi: 10.1016/j.socec.2009.07.010

Hacking, I. (2006). *The emergence of probability* (2nd ed.). Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511817557

Hafenbrädl, S., & Hoffrage, U. (2015). Toward an ecological analysis of Bayesian inferences: How task characteristics influence responses. *Frontiers in Psychology*, *6*, 1-15. doi: 10.3389/fpsyg.2015.00939

Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: why three heads are better than four. *Psychological Review*, *116*(2), 454. doi: 10.1037/a0015241

Haller, H. (1985). The principal-agent problem with a satisficing agent. *Journal of Economic Behavior & Organization*, *6*, 359–379. doi: 10.1016/0167-2681(85)90004-6

Handel, B., & Schwartzstein, J. (2018). Frictions or mental gaps: what's behind the information we (don't) use and when do we care? *Journal of Economic Perspectives*, *32*(1), 155–78. doi: 10.1257/jep.32.1.155

Harstad, R. M., & Selten, R. (2013). Bounded-Rationality Models: Tasks to Become Intellectually Competitive. *Journal of Economic Literature*, *51*(2), 496–511. doi: 10.1257/jel.51.2.496

Hastie, P., Tibshirani, R., & Friedman, J. (2001). *The Elements of Statistical Learning* (2nd ed.). New York, NY: Springer. doi: 10.1007/978-0-387-21606-5

Hauser, J. R. (2014). Consideration-set heuristics. *Journal of Business Research*, *67*(8), 1688–1699. doi: 10.1016/j.jbusres.2014.02.015

Hauser, J. R., Toubia, O., Evgeniou, T., Befurt, R., & Dzyabura, D. (2010). Disjunctions of Conjunctions, Cognitive Simplicity, and Consideration Sets. *Journal of Marketing Research*, *47*(3), 485–496. doi: 10.1509/jmkr.47.3.485

Hauser, J. R., & Wernerfelt, B. (1990). An Evaluation Cost Model of Consideration Sets. *Journal of Consumer Research*, *16*(4), 393–408.

doi: 10.1086/209225

Heath, C., Larrick, R. P., & Wu, G. (1999). Goals as reference points. *Cognitive Psychology*, *38*(1), 79–109. doi: 10.1006/cogp.1998.0708

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. doi: 10.1002/jrsm.5

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486. doi: 10.1037/1082-989X.3.4.486

Heiner, R. (1983). The origin of predictable behavior. *American Economic Review*, *76*(4), 560–595.

Heiner, R. (1989). The origin of predictable dynamic behavior. *Journal of Economic Behavior & Organization*, *12*(2), 233–257. doi: 10.1016/0167-2681(89)90057-7

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*, 61-83. doi: 10.1017/S0140525X0999152X

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*(8), 534-9. doi: 10.1111/j.0956-7976.2004.00715.x

Hertwig, R., & Engel, C. (2016). Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science*, *11*(3), 359–372. doi: 10.1177/1745691616635594

Hertwig, R., & Erev, I. (2009). The description-experience gap in risky choice. *Trends in Cognitive Sciences*, *13*(12), 517-523. doi: 10.1016/j.tics.2009.09.004

Hertwig, R., Hoffrage, U., & the ABC Research Group. (2013). *Simple heuristics in a social world*. New York, NY: Oxford University Press.

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(4), 621. doi: 10.1037/0278-7393.31.4.621

Hertwig, R., & Pleskac, T. J. (2010). Decisions from experience: Why small samples? *Cognition*, *115*(2), 225–237. doi: 10.1016/j.cognition.2009.12.009

Hey, J. D., Permana, Y., & Rochanahastin, N. (2017). When and how to satisfice: an experimental investigation. *Theory and Decision*,

83(3), 337–353. doi: 10.1007/s11238-017-9600-5

Higgins, J., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. doi: 10.1002/sim.1186

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ*, *327*(7414), 557–560. doi: 10.1136/bmj.327.7414.557

Hill, W. T., & Brase, G. L. (2012). When and for whom do frequencies facilitate performance? On the role of numerical literacy. *Quarterly Journal of Experimental Psychology*, *65*, 2343-2368. doi: 10.1080/17470218.2012.687004

Hill, W. T., & Brase, G. L. (2015). *Natural frequencies improve diagnostic test result comprehension when using one, two, and three cues.* Unpublished manuscript.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.1970.10488634

Hoffrage, U., & Gigerenzer, G. (1998). Using natural frequencies to improve diagnostic inferences. *Academic Medicine*, *73*, 538-540. doi: 10.1097/00001888-199805000-00024

Hoffrage, U., & Gigerenzer, G. (2004). How to improve the diagnostic inferences of medical experts. In E. Kurz-Milcke & G. Gigerenzer (Eds.), *Experts in science and society.* (p. 249-268). New York, NY: Kluwer Academic/Plenum Publishers. doi: 10.1007/b105826

Hoffrage, U., Gigerenzer, G., Krauss, S., & Martignon, L. (2002). Representation facilitates reasoning: What natural frequencies are and what they are not. *Cognition*, *84*, 343-352. doi: 10.1016/S0010-0277(02)00050-1

Hoffrage, U., Hafenbrädl, S., & Bouquet, C. (2015). Natural frequencies facilitate diagnostic inferences of managers. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00642

Hoffrage, U., Krauss, S., Martignon, L., & Gigerenzer, G. (2015). Natural frequencies improve Bayesian reasoning in simple and complex inference tasks. *Frontiers in Psychology*, *6*, 1-14. doi: 10.3389/fpsyg.2015.01473

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating statistical information. *Science*, *290*, 2261-2262. doi: 10.1126/science.290.5500.2261

Hogarth, R. M., & Karelaia, N. (2007). Heuristic and linear models of judgment: matching rules and environments. *Psychological*

Review, *114*(3), 733–758. doi: 10.1037/0033-295X.114.3.733

Hooke, R., & Jeeves, T. A. (1961). A "Direct Search" solution of numerical and statistical problems. *Journal of the ACM (JACM)*, *8*(2), 212–229. doi: 10.1145/321062.321069

Hume, D. (1739/2000). *A treatise of human nature*. Oxford University Press. doi: 10.1093/oseo/instance.00046221

Ioannidis, J. P. A., Stuart, M. E., Brownlee, S., & Strite, S. A. (2017). How to survive the medical misinformation mess. *European Journal of Clinical Investigation*, *47*(11), 795-802. doi: 10.1111/eci.12834

Jackson, D., White, I. R., & Riley, R. D. (2012). Quantifying the impact of between-study heterogeneity in multivariate meta-analyses. *Statistics in Medicine*, *31*(29), 3805–3820. doi: 10.1002/sim.5453

Jacobs, R. L., & Jones, R. A. (1980). Price expectations in the United States: 1947-75. *American Economic Review*, 269–277.

Jamal, K., & Sunder, S. (2001). Why do biased heuristics approximate Bayes rule in double auctions? *Journal of Economic Behavior and Organization*, *46*(4), 431–435. doi: 10.1016/S0167-2681(01)00173-1

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth berkeley symposium on mathematical statistics and probability* (p. 361).

Jamison, D. T., & Lau, L. J. (1973). Semiorders and the theory of choice. *Econometrica*, *41*(5), 901–912. doi: 10.2307/1913813

Jefferson, T., Demicheli, V., Rivetti, D., & Deeks, J. (1999). Cochrane reviews and systematic reviews of economic evaluations. *Pharmacoeconomics*, *16*(1), 85–89. doi: 10.2165/00019053-199916001-00011

Johnson, E. D., & Tubau, E. (2013). Words, numbers, & numeracy: Diminishing individual differences in Bayesian reasoning. *Learning and Individual Differences*, *28*, 34-40. doi: 10.1016/j.lindif.2013.09.004

Johnson, E. D., & Tubau, E. (2015). Comprehension and computation in Bayesian problem solving. *Frontiers in Psychology*, *6*, 1-19. doi: 10.3389/fpsyg.2015.00938

Johnson-Laird, P. N., Legrenzi, P., Girotto, V., Legrenzi, M. S., & Caverni, J. P. (1999). Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, *106*, 62-88. doi: 10.1037/0033-295X.106.1.62

Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2020). Simple rules to guide expert classifications. *Journal of the Royal*

Statistical Society: Series A (Statistics in Society), *183*(3), 771-800. doi: 10.1111/rssa.12576

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*(3), 430-454. doi: 10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, *80*(4), 237-251. doi: 10.1037/h0034747

Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, *47*(2), 263–292. doi: 10.2307/1914185

Kahneman, D., & Tversky, A. (1982). On the study of statistical intuitions. *Cognition*, *11*, 123-141. doi: 10.1016/0010-0277(82)90022-1

Katsikopoulos, K., Şimşek, O., Buckmann, M., & Gigerenzer, G. (2020). *Classification in the wild*. MIT Press.

Keynes, J. M. (1921). *Treatise on Probability*. London: Macmillan and Co.

Khan, A., Breslav, S., Glueck, M., & Hornbaek, K. (2015). Benefits of visualization in the mammography problem. *International Journal of Human-Computer Studies*, *83*, 94-113. doi: 10.1016/j.ijhcs.2015.07.001

Khare, R., Leaman, R., & Lu, Z. (2014). Accessing biomedical literature in the current information landscape. In *Biomedical literature mining* (pp. 11–31). Springer. doi: 10.1007/978-1-4939-0709-0_2

King, M., & Kay, J. (2020). *Radical Uncertainty: Decision-Making for an Unknowable Future* (1st ed.). The Bridge Street Press.

Kleiter, G. D. (1994). Natural sampling: Rationality without base rates. In G. H. Fischer & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (p. 375-388). New York, NY: Springer. doi: 10.1007/978-1-4612-4308-3_27

Knight, F. H. (1921). *Risk, Uncertainty, and Profit*. Boston: Houghton Mifflin.

Kochetova-Kozloski, N., Messier, W. F. J., & Eilifsen, A. (2011). Improving auditors' fraud judgments using a frequency response mode. *Contemporary Accounting Research*, *28*, 837-858. doi: 10.1111/j.1911-3846.2011.01067.x

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*, 1-17. doi: 10.1017/S0140525X00041157

Konheim-Kalkstein, Y. L. (2008). *Facilitation of Bayesian decision making.*

(Doctoral dissertation). University of Minnesota, Twin Cities, MN.

Kool, W., & Botvinick, M. (2014). A labor/leisure tradeoff in cognitive control. *Journal of Experimental Psychology: General*, *143*(1), 3-18. doi: 10.1037/2333-8113.1.S.3

Köszegi, B., & Rabin, M. (2006). A model of reference-dependent preferences. *Quarterly Journal of Economics*, *121*(4), 1133–1165. doi: 10.1093/qje/121.4.1133

Krauss, S., Martignon, L., & Hoffrage, U. (1999). Simplifying Bayesian inference: The general case. In L. Magnani, N. J. Nersessian, & P. Thagard (Eds.), *Model-based reasoning in scientific discovery* (p. 165-179). New York, NY: Kluwer Academic/Plenum Publishers. doi: 10.1007/978-1-4615-4813-3_11

Krishnan, K. (1977). Incorporating Thresholds of Indifference in probabilistic choice models. *Management Science*, *23*(11), 1224–1233. doi: 10.1287/mnsc.23.11.1224

Krynski, T. R., & Tenenbaum, J. B. (2007). The role of causality in judgment under uncertainty. *Journal of Experimental Psychology: General*, *136*, 430-450. doi: 10.1037/0096-3445.136.3.430

Kunreuther, H. (1976). Limited knowledge and insurance protection. *Public Policy*, *24*(2), 227–261.

Kurzenhäuser, S., & Hoffrage, U. (2002). Teaching Bayesian reasoning: An evaluation of a classroom tutorial for medical students. *Medical Teacher*, *24*, 516-21. doi: 10.1080/0142159021000012540

Lagnado, D. A., & Shanks, D. R. (2007). Dual concerns with the dualist approach. *Behavioral and Brain Sciences*, *30*, 271-272. doi: 10.1017/S0140525X0700180X

Lant, T. (1992). Aspiration Level Adaptation: An empirical exploration. *Management Science*, *38*(5), 623–644. doi: 10.1287/mnsc.38.5.623

Lant, T., & Shapira, Z. (2008). Managerial reasoning about aspirations and expectations. *Journal of Economic Behavior & Organization*, *66*(1), 60–73. doi: 10.1016/j.jebo.2007.03.006

Lau, J., Ioannidis, J. P., Terrin, N., Schmid, C. H., & Olkin, I. (2006). Evidence based medicine: The case of the misleading funnel plot. *BMJ*, *333*(7568), 597-600. doi: 10.1136/bmj.333.7568.597

Leach, J. R. (2002). *Information selection in a simulated medical diagnosis task: The effects of external representations and compeltely natural sampling*. (Doctoral dissertation). Bowling Green State University, Bowling Green, OH.

Lesage, E., Navarrete, G., & De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. *Thinking & Reasoning*, *19*, 27-53. doi: 10.1080/13546783.2012.713177

Levinthal, D., & March, J. G. (1981). A model of adaptive organizational search. *Journal of Economic Behavior and Organization*, *2*(4), 307–333. doi: 10.1016/0167-2681(81)90012-3

Levy, S. (2015). Bitte aussteigen [please exit here]. *Die Zeit*, *2015*(13).

Lewin, K. (1935). *A dynamic theory of personality*. New York: McGraw-Hill.

Lewin, K., Dembo, T., Festinger, L., & Sears, P. S. (1944). Level of Aspiration. In J. McVicker Hunt (Ed.), *Personality and the behavior disorders* (pp. 333–378). New York: The Ronald Press Company.

Lichtman, A. J. (2020). *Predicting the next president: The keys to the white house*. Rowman & Littlefield Publishers.

Lieder, F., Griffiths, T. L., & Hsu, M. (2018). Overrepresentation of extreme events in decision making reflects rational use of cognitive resources. *Psychological Review*, *125*(1), 1. doi: 10.1037/rev0000074

Lieder, F., Plunkett, D., Hamrick, J. B., Russell, S. J., Hay, N., & Griffiths, T. (2014). Algorithm selection by rational metareasoning as a model of human strategy selection. In *Advances in neural information processing systems* (pp. 2870–2878).

Lindsey, S., Hertwig, R., & Gigerenzer, G. (2003). Communicating statistical DNA evidence. *Jurimetrics*, *43*(2), 147-163.

Lioukas, S. (1984). Thresholds and Transitivity in stochastic consumer choice: A multinomial logit analysis. *Management Science*, *30*(1), 110–122. doi: 10.1287/mnsc.30.1.110

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*(1), 37-44. doi: 10.1177/0272989x0102100105

Loomes, G., Starmer, C., & Sugden, R. (2003). Do anomalies disappear in repeated markets. *Economic Journal*, *113*(486). doi: 10.1111/1468-0297.00108

Luan, S., Reb, J., & Gigerenzer, G. (2019). Ecological rationality: Fast-and-frugal heuristics for managerial decision making under uncertainty. *Academy of Management Journal*, *62*(6), 1735–1759. doi: 10.5465/amj.2018.0172

Luan, S., Schooler, L. J., & Gigerenzer, G. (2011). A signal-detection analysis of fast-and-frugal trees. *Psychological Review*, *118*(2),

316–38. doi: 10.1037/a0022684

Luan, S., Schooler, L. J., & Gigerenzer, G. (2014). From perception to preference and on to inference: An approach-avoidance analysis of thresholds. *Psychological Review*, *121*(3), 501–25. doi: 10.1037/a0037025

Lucci, S., & Kopec, D. (2015). *Artificial intelligence in the 21st century*. Stylus Publishing, LLC.

Luce, R. D. (1956). Semiorders and a theory of utility discrimination. *Econometrica*, *24*(2), 178–191. doi: 10.2307/1905751

Macchi, L. (2000). Partitive formulation of information in probabilistic problems: Beyond heuristics and frequency format explanations. *Organizational Behavior and Human Decision Processes*, *82*, 217-236. doi: 10.1006/obhd.2000.2895

Maddock, R., & Carter, M. (1982). A child's guide to rational expectations. *Journal of Economic Literature*, *20*(1), 39–51.

Malmendier, U., & Nagel, S. (2011). Depression babies: do macroeconomic experiences affect risk taking? *Quarterly Journal of Economics*, *126*(1), 373–416. doi: 10.1093/qje/qjq004

Mandel, D. R. (2007). Nested sets theory, full stop: Explaining performance on Bayesian inference tasks without dual-systems assumptions. *Behavioral and Brain Sciences*, *30*, 275-276. doi: 10.1017/S0140525X07001835

Mandel, D. R. (2014). The psychology of Bayesian reasoning. *Frontiers In Psychology*, *5*, 1144. doi: 10.3389/fpsyg.2014.01144

Manski, C. F. (2017). Optimize, satisfice, or choose without deliberation? a simple minimax-regret assessment. *Theory and Decision*, *83*(2), 155–173. doi: 10.1007/s11238-017-9592-1

Manzini, P., & Mariotti, M. (2007). Sequentially Rationalizable Choice. *American Economic Review*, *97*(5), 1824–1839. doi: 10.1257/aer.97.5.1824

Manzini, P., & Mariotti, M. (2012). Choice by lexicographic semiorders. *Theoretical Economics*, *7*(1), 1–23. doi: 10.3982/TE679

Manzini, P., Mariotti, M., & Tyson, C. J. (2013). Two-stage threshold representations. *Theoretical Economics*, *8*, 875–882. doi: 10.3982/TE1048

March, J. G. (1981). Footnotes to Organizational Change. *Administrative Science Quarterly*, *26*(4), 563–577. doi: 10.2307/2392340

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organizational Science*, *2*(1), 71–87. doi: 10.1287/orsc.2.1.71

March, J. G., & Simon, H. A. (1958). *Organizations*. New York: John Wiley & Sons.

Marewski, J. N., Gaissmaier, W., Schooler, L. J., Goldstein, D. G., & Gigerenzer, G. (2010). From recognition to decisions: extending and testing recognition-based models for multialternative inference. *Psychonomic Bulletin & Review*, *17*(3), 287–309. doi: 10.3758/PBR.17.3.287

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review*, *118*(3), 393-437. doi: 10.1037/a0024143

Markose, S. M. (2005). Computability and evolutionary complexity: markets as complex adaptive systems (cas). *The Economic Journal*, *115*(504), F159–F192. doi: 10.1111/j.1468-0297.2005.01000.x

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*(1), 29–71. doi: 10.1023/A:1015516217425

Martignon, L., Katsikopoulos, K. V., & Woike, J. K. (2008). Categorization with limited resources: A family of simple heuristics. *Journal of Mathematical Psychology*, *52*(6), 352–361. doi: 10.1016/j.jmp.2008.04.003

McDowell, M., Gigerenzer, G., Wegwarth, O., & Rebitschek, F. G. (2019). Effect of tabular and icon fact box formats on comprehension of benefits and harms of prostate cancer screening: a randomized trial. *Medical Decision Making*, *39*(1), 41–56. doi: 10.1177/0272989X18818166

McNair, S. (2015). Beyond the status-quo: Research on Bayesian reasoning must develop in both theory and method. *Frontiers in Psychology*, *6*, 1-3. doi: 10.3389/fpsyg.2015.00097

McNair, S., & Feeney, A. (2014). When does information about causal structure improve statistical reasoning? *Quarterly Journal of Experimental Psychology*, *67*, 625-645. doi: 10.1080/17470218.2013.821709

McNair, S., & Feeney, A. (2015). Whose statistical reasoning is facilitated by a causal structure intervention? *Psychonomic Bulletin & Review*, *22*(1), 258-264. doi: 10.3758/s13423-014-0645-y

Mellers, B. A., & McGraw, A. P. (1999). How to improve Bayesian reasoning: Comment on Gigerenzer and Hoffrage (1995). *Psychological Review*, *106*, 417-424. doi: 10.1037/0033-295X.106.2.417

Mezias, S. J., Chen, Y.-R., & Murphy, P. R. (2002). Aspiration-Level

Adaptation in American Financial Services Study Organization: A. *Management Science*, *48*(10), 1285–1300. doi: 10.1287/mnsc.48.10.1285.277

Micallef, L., Dragicevic, P., & Fekete, J.-D. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Transactions On Visualization and Computer Graphics*, *18*, 2536-2545. doi: 10.1109/TVCG.2012.199

Miller, D., & Chen, M.-J. (1994). Sources and consequences of competitive inertia: A study of the US airline industry. *Administrative Science Quarterly*, 1–23. doi: 10.2307/2393492

Miller, J. B., & Sanjurjo, A. (2018). Surprised by the hot hand fallacy? a truth in the law of small numbers. *Econometrica*, *86*(6), 2019–2047. doi: 10.3982/ECTA14943

Misuraca, R., Carmeci, F. A., Pravettoni, G., & Cardaci, M. (2009). Facilitating effect of natural frequencies: Size does not matter. *Perceptual and Motor Skills*, *108*, 422-430. doi: 10.2466/PMS.108.2.422-430

Moro, R., Bodanza, G. A., & Freidin, E. (2011). Sets or frequencies? How to help people solve conditional probability problems. *Journal of Cognitive Psychology*, *23*, 843-857. doi: 10.1080/20445911.2011.579072

Mulley, A. G., & Wennberg, J. E. (2011). Reducing unwarranted variation in clinical practice by supporting clinicians and patients in decision making. In G. Gigerenzer & J. A. Muir Gray (Eds.), *Better doctors, better patients, better decisions: Envisioning health care 2020* (pp. 29–44). Cambridge, MA: MIT Press. doi: 10.7551/mitpress/9780262016032.003.0003

Muth, J. F. (1961). Rational expectations and the theory of price. *Econometrica*, *29*(3), 315–335. doi: 10.2307/1909635

Nash, J. (1950). The bargaining problem. *Econometrica*, *18*(2), 155–162. doi: 10.2307/1907266

Navarrete, G., Correia, R., Sirota, M., Juanchich, M., & Huepe, D. (2015). Doctor, what does my positive test mean? From Bayesian textbook tasks to personalized risk communication. *Frontiers in Psychology*, *6*, 1-6. doi: 10.3389/fpsyg.2015.01327

Navarrete, G., & Santamaria, C. (2011). Ecological rationality and evolution: The mind really works that way? *Frontiers In Psychology*, *2*, 251. doi: 10.3389/fpsyg.2011.00251

Navarro-Martinez, D., Loomes, G., Isoni, A., Butler, D., & Alaoui, L. (2018). Boundedly rational expected utility theory. *Journal of Risk*

and Uncertainty, *57*(3), 199–223. doi: 10.1007/s11166-018-9293-3

Neace, W. P., Michaud, S., Bolling, L., Deer, K., & Zecevic, L. (2008). Frequency formats, probability formats, or problem structure? A test of the nested-sets hypothesis in an extensional reasoning task. *Judgment and Decision Making*, *3*, 140-152.

Nelson, R. R., & Winter, S. G. (1973). Toward an Evolutionary Theory of Economic Capabilities. *American Economic Review*, *63*(2), 440–449.

Nelson, R. R., & Winter, S. G. (1982). *An Evolutionary Theory of Economic Change*. Cambridge: Harvard University Press.

Obrecht, N. A., Anderson, B., Schulkin, J., & Chapman, G. B. (2012). Retrospective frequency formats promote consistent experience-based Bayesian judgments. *Applied Cognitive Psychology*, *26*, 436-440. doi: 10.1002/acp.2816

O'Brien, D., Roazzi, A., & da Graca B. B. Dias, M. (2004). Reasoning about conditional probabilities: The evidence for the frequency hypothesis has relied on flawed comparisons. *Estudos de Psicologia*, *9*, 35-43. doi: 10.1590/S1413-294X2004000100005

Oechssler, J. (2002). Cooperation as a result of learning with aspiration levels. *Journal of Economic Behavior & Organization*, *49*(3), 405–409. doi: 10.1016/S0167-2681(02)00013-6

Oettinger, G. S. (1999). An empirical analysis of the daily labor supply of stadium vendors. *Journal of Political Economy*, *107*(2), 360–392. doi: 10.1086/250063

Ottley, A., Peck, E. M., Harrison, L. T., Afergan, D., Ziemkiewicz, C., Taylor, H. A., ... Chang, R. (2016). Improving Bayesian reasoning: The effects of phrasing, visualization, and spatial ability. *IEEE Transactions On Visualization and Computer Graphics*, *22*, 529-538. doi: ieeecomputersociety.org/10.1109/TVCG.2015.2467758

Papadimitriou, C. H., & Steiglitz, K. (1998). *Combinatorial optimization: algorithms and complexity*. Courier Corporation.

Papi, M. (2012). Satisficing choice procedures. *Journal of Economic Behavior & Organization*, *84*(1), 451–462. doi: 10.1016/j.jebo.2012.04.017

Papi, M. (2013). Satisficing and maximizing consumers in a monopolistic screening model. *Mathematical Social Sciences*, *66*(3), 385–389. doi: 10.1016/j.mathsocsci.2013.08.005

Papi, M. (2018). Price competition with satisficing consumers. *International Journal of Industrial Organization*, *58*, 252–272. doi:

10.1016/j.ijindorg.2017.09.001

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition*, *14*(3), 534. doi: 10.1037/0278-7393.14.3.534

Payne, J. W., Laughhunn, D. J., & Crum, R. (1980). Translation of gambles and aspiration level effects in risky choice behavior. *Management Science*, *26*(10), 1039–1060. doi: 10.1287/mnsc.26.10.1039

Payne, J. W., Laughhunn, D. J., & Crum, R. (1981). Note—further tests of aspiration level effects in risky choice behavior. *Management Science*, *27*(8), 953–958. doi: 10.1287/mnsc.27.8.9539

Pazgal, A. (1997). Satisficing Leads to Cooperation in Mutual Interests. *International Journal of Game Theory*, *26*, 439–453. doi: 10.1007/BF01813884

Pearl, J. (1984). *Intelligent search strategies for computer problem solving*. Addision Wesley.

Persson, A., & Ryals, L. (2014). Making customer relationship decisions: Analytics v rules of thumb. *Journal of Business Research*, *67*(8), 1725–1732. doi: 0.1016/j.jbusres.2014.02.019

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2006). Comparison of two methods to detect publication bias in meta-analysis. *JAMA*, *295*(6), 676–680. doi: 10.1001/jama.295.6.676

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29-46. doi: 10.1037/h0024722

Pfeifer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior and Human Decision Processes*, *58*(2), 203–213. doi: 10.1006/obhd.1994.1034

Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346-354. doi: 10.1037/h0023653

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rabin, M. (2013). Incorporating limited rationality into economics. *Journal of Economic Literature*, *51*(2), 528–43. doi: 10.1257/jel.51.2.528

Rakoczy, H., Cluever, A., Saucke, L., Stoffregen, N., Graebener, A.,

Migura, J., & Call, J. (2014). Apes are intuitive statisticians. *Cognition*, *131*, 60-68. doi: 10.1016/j.cognition.2013.12.011

Real, L. (1991). Animal choice behavior and the evolution of cognitive architecture. *Science*, *253*(5023), 980-986. doi: 10.1126/science.1887231

Real, L., & Caraco, T. (1986). Risk and foraging in stochastic environments. *Annual Review of Ecology and Systematics*, *17*, 371-390. doi: 10.1146/annurev.es.17.110186.002103

Reis, R. (2006). Inattentive consumers. *Journal of Monetary Economics*, *53*(8), 1761–1800. doi: 10.1016/j.jmoneco.2006.03.001

Research America. (2017). *U.S. investments in medical and health research and development, 2013-2016* (Tech. Rep.). Research America. (Available at `https://www.researchamerica.org/sites/default/files/RA-2017_InvestmentReport.pdf` (accessed 2 January 2019))

Reutskaja, E., Nagel, R., Camerer, C. F., & Rangel, A. (2011). Search Dynamics in Consumer Choice under Time Pressure: An Eye-Tracking Study. *American Economic Review*, *101*, 900–926. doi: 10.1257/aer.101.2.900

Rothschild, M. (1974). Searching for the lowest price when the distribution of prices is unknown. *Journal of Political Economy*, *82*(4), 689–711. doi: 10.1086/260229

Rubinstein, A., & Salant, Y. (2006). A model of choice from lists. *Theoretical Economics*, *1*, 3–17.

Ruscio, J. (2003). Comparing Bayes's theorem to frequency-based approaches to teaching Bayesian reasoning. *Teaching of Psychology*, *30*, 325-328.

Salant, Y. (2011). Procedural Analysis of Choice Rules with Applications to Bounded Rationality. *American Economic Review*, *101*, 724–748. doi: 10.1257/aer.101.2.724

Samuels, R. (2007). Varieties of dual-process theory for probabilistic reasoning. *Behavioral and Brain Sciences*, *30*, 280-281. doi: 10.1017/S0140525X07001884

Samuelson, P. A., & Nordhaus, W. D. (1998). *Economics* (16th ed.). McGraw-Hill.

Santos, B. D. L., Hortaçsu, A., & Wildenbeest, M. R. (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review*, *102*(6), 2955–2980. doi: 10.1257/aer.102.6.2955

Sauermann, H., & Selten, R. (1962). Anspruchanpassungstheorie.

Zeitschrift für die gesamte Staatswissenschaft, *118*, 577–597.
Savage, L. J. (1954). *The foundations of statistics*. New York, NY: Wiley.
Schmittlein, D. C., Morrison, D. G., & Colombo, R. (1987). Counting Your Customers: Who Are They and What Will They Do Next? *Management Science*, *33*(1), 1–24. doi: 10.1287/mnsc.33.1.1
Schulze, C., & Hertwig, R. (2016). *Statistical intuitions: Smart babies, stupid adults?* Manuscript submitted for publication.
Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*(11), 966-72. doi: 10.7326/0003-4819-127-11-199712010-00003
Schünemann, H. J., Oxman, A. D., Vist, G. E., Higgins, J. P. T., Deeks, J. J., Glasziou, P. P., & Guyatt, G. D. (2008). Interpreting results and drawing conclusions. In J. P. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions* (pp. 359–387). Cambridge, MA: Cochrane. doi: 10.1002/9780470712184.ch12
Sedlmeier, P., & Gigerenzer, G. (2001). Teaching Bayesian reasoning in less than two hours. *Journal of Experimental Psychology: General*, *130*, 380-400. doi: 10.1037//0096-3445.130.3.380
Selten, R. (1998). Aspiration Adaptation Theory. *Journal of Mathematical Psychology*, *42*(2/3), 191–214. doi: 10.1006/jmps.1997.1205
Selten, R. (2001). What is Bounded Rationality. In G. Gigerenzer & R. Selten (Eds.), *Bounded rationality - the adaptive toolbox* (pp. 13–36). Cambridge: MIT Press.
Shannon, C. E. (1950). A chess-playing machine. *Scientific American*, *182*(2), 48–51. doi: 10.1038/scientificamerican0250-48
Shubik, M. (1958). Models of Man: Social and Rational by H.A. Simon. *Journal of Political Economy*, *66*(3), 273–274. doi: 10.1086/258048
Siegler, R. S., Thompson, C. A., & Schneider, M. (2011). An integrated theory of whole number and fractions development. *Cognitive Psychology*, *62*(4), 273-296. doi: 10.1016/j.cogpsych.2011.03.001
Siegrist, M., & Keller, C. (2011). Natural frequencies and Bayesian reasoning: The impact of formal education and problem context. *Journal of Risk Research*, *14*, 1039-1055. doi: 10.1080/13669877.2011.571786
Şimşek, O. (2013). Linear decision rule as aspiration for simple decision heuristics. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 2904–2912). Curran

Associates, Inc.

Şimşek, O., & Buckmann, M. (2015). Learning from small samples: An analysis of simple decision heuristics. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 3159–3167). Curran Associates, Inc.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118. doi: 10.2307/1884852

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(3), 129–138. doi: 10.1037/h0042769

Simon, H. A. (1957). Rationality and administrative decision making. In *Models of man* (pp. 196–206). New York, NY: John Wiley & Sons, Ltd.

Simon, H. A. (1959). Theories of decision-making in economics and behavioral science. *American Economic Review*, *49*(3), 253–283.

Simon, H. A. (1979). Rational Decision Making in Business Organizations. *American Economic Review*, *69*(4), 493–513.

Simon, H. A. (1981). *The Sciences of the Artificial* (2nd ed.). London: MIT Press.

Simon, H. A. (1985). Human nature in politics: The dialogue of psychology with political science. *American political science review*, *79*(2), 293–304. doi: 10.2307/1956650

Simon, H. A. (1989). The scientist as problem solver. *Complex information processing: The impact of Herbert A. Simon*, 375–398. doi: 10.21236/ADA240569

Simon, H. A. (1996). *Letter to john conlisk.* Retrieved from `http://digitalcollections.library.cmu.edu/awweb/awarchive?type=file&item=48878`

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, *50*(3), 665–690. doi: 10.1016/S0304-3932(03)00029-1

Sirota, M., & Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies. *Studia Psychologica*, *53*, 151-161.

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic Bulletin & Review*, *21*, 198-204. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., & Juanchich, M. (2014). The effect of

iconicity of visual displays on statistical reasoning: Evidence in favor of the null hypothesis. *Psychonomic Bulletin & Review*, *21*, 961-8. doi: 10.3758/s13423-013-0555-4

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015a). How to train your Bayesian: A problem-representation transfer rather than a format-representation shift explains training effects. *Quarterly Journal of Experimental Psychology*, *68*, 1-9. doi: 10.1080/17470218.2014.972420

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015b). Now you Bayes, now you don't: Effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic Bulletin & Review*, *22*(5), 1465-1473. doi: 10.3758/s13423-015-0810-y

Sirota, M., Vallée-Tourangeau, G., Vallée-Tourangeau, F., & Juanchich, M. (2015). On Bayesian problem-solving: Helping Bayesians solve simple Bayesian word problems. *Frontiers in Psychology*, *6*, 1-4. doi: 10.3389/fpsyg.2015.01141

Sloman, S. A., Over, D., Slovak, L., & Stibel, J. M. (2003). Frequency illusions and other fallacies. *Organizational Behavior and Human Decision Processes*, *91*, 296-309. doi: 10.1016/S0749-5978(03)00021-9

Slovic, P., Kunreuther, H., & White, G. F. (1974). Decision processes, rationality and adjustment to natural hazards: A review of some hypotheses. In *Natural hazards: Local, national and global, edited by gilbert white* (pp. 187–205). Oxford University Press.

Smith, A. (1776/1976). *An inquiry into the nature and causes of the wealth of nations* (Vol. 1). Clarendon Press. doi: 10.1093/oseo/instance.00043218

Smith, A. (1779/1981). *The theory of moral sentiments* (D. Raphael & A. Macfie, Eds.). Liberty Fund.

Smith, V. L. (1962). An experimental Study of Competitive Market Behavior. *Journal of Political Economy*, *70*(2), 111–137. doi: 10.1086/258609

Smith, V. L. (2008). *Rationality in economics: Constructivist and ecological forms*. Cambridge University Press.

Sobel, D. M., & Munro, S. E. (2009). Domain generality and specificity in children's causal inference about ambiguous data. *Developmental Psychology*, *45*, 511-524. doi: 10.1037/a0014944

Sobel, D. M., Tenenbaum, J. B., & Gopnik, A. (2004). Children's causal inferences from indirect evidence: Backwards blocking

and Bayesian reasoning in preschoolers. *Cognitive Science*, *28*, 303-333. doi: 10.1016/j.cogsci.2003.11.001

Solow, R. M. (1958). Models of Man - Social and Rational, by Herbert A . Simon. *The Review of Economics and Statistics*, *40*(1), 81–84. doi: 10.2307/1926487

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*(3), 355–374.

Stanovich, K. E., & West, R. F. (2000). Advancing the rationality debate. *Behavioral and Brain Sciences*, *23*(05), 701–717. doi: 10.1017/S0140525X00623439

Stedman, M. R., Curtin, F., Elbourne, D. R., Kesselheim, A. S., & Brookhart, M. A. (2011). Meta-analyses involving cross-over trials: Methodological issues. *International Journal of Epidemiology*, *40*(6), 1732–1734. doi: 10.1093/ije/dyp345

Sterman, J. D. (1987). Systems Simulation. Expectation formation in behavioral simulation models. *Behavioral Science*, *32*(3), 190–211. doi: 10.1002/bs.3830320304

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, *69*(3), 213–225. doi: 10.1086/258464

Stiglitz, J. E., & Weiss, A. (1987). *Macro-economic equilibrium and credit rationing* (Tech. Rep.). National Bureau of Economic Research. doi: 10.3386/w2164

Stüttgen, P., Boatwright, P., & Monroe, R. T. (2012). A Satisficing Choice Model. *Marketing Science*, *31*(6), 878–899. doi: 10.1287/mksc.1120.0732

Telser, L. G. (1973). Searching for the Lowest Price. *American Economic Review*, *63*(2), 40–49.

Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New Haven, CT: Yale University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Todd, P. M., & Brighton, H. (2015). Building the theory of ecological rationality. *Minds and Machines*, 1-22. doi: 10.1007/s11023-015-9371-0

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*, *16*(3), 167-171. doi: 10.1111/j.1467-8721.2007.00497.x

Todd, P. M., Hertwig, R., & Hoffrage, U. (2005). Evolutionary

cognitive psychology. In D. M. Buss (Ed.), *The handbook of evolutionary psychology* (p. 776-802). Hoboken, NJ: Wiley. doi: 10.1002/9780470939376.ch27

Trafimow, D. (2007). Why the empirical literature fails to support or disconfirm modular or dual-process models. *Behavioral and Brain Sciences*, *30*, 283-284. doi: 10.1017/S0140525X07001926

Tsai, J., Miller, S., & Kirlik, A. (2011). Interactive visualizations to improve Bayesian reasoning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 385–389. doi: 10.1177/1071181311551079

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48. doi: 10.1037/h0026750

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*(4), 281–299. doi: 10.1037/h0032955

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: heuristics and biases. *Science*, *185*, 1124–1131. doi: 10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453–458. doi: 10.1126/science.7455683

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315. doi: 10.1037/0033-295X.90.4.293

Tversky, A., & Kahneman, D. (1992). Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*, *5*, 297–323. doi: 10.1007/BF00122574

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (p. 79-112). Massachusetts, MA: MIT Press.

Tversky, B. (2011). Visualizing thought. *Topics in Cognitive Science*, *3*(3), 499-535. doi: 10.1111/j.1756-8765.2010.01113.x

Useem, J., Brennan, A., LaValley, M., Vickery, M., Ameli, O., Reinen, N., & Gill, C. J. (2015). Systematic differences between cochrane and non-cochrane meta-analyses on the same topic: a matched pair analysis. *PloS one*, *10*(12), e0144980.

Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2015). Interactivity fosters Bayesian reasoning without instruction. *Journal of Experimental Psychology: General*, *144*(3), 581-603. doi: 10.1037/a0039161

Vallée-Tourangeau, G., Abadie, M., & Vallée-Tourangeau, F. (2016). *Interactivity fosters Bayesian reasoning without instruction* [Data file and code book]. Retrieved from `osf.io/2ur8f`

Vallée-Tourangeau, G., Sirota, M., Juanchich, M., & Vallée-Tourangeau, F. (2015). Beyond getting the numbers right: What does it mean to be a "successful" Bayesian reasoner? *Frontiers in Psychology*, *6*, 712. doi: 10.3389/fpsyg.2015.00712

van Rooij, R. (2010, October). Revealed preference and satisficing behavior. *Synthese*, *179*(S1), 1–12. doi: 10.1007/s11229-010-9826-z

van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293-309. doi: 10.1037/met0000025

van Houwelingen, H. C., Zwinderman, K. H., & Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine*, *12*, 2273–2284. doi: 10.1002/sim.4780122405

Varian, H. R. (1980). A model of sales. *American Economic Review*, *70*(4), 651–659.

Verrecchia, R. E. (1982). Information acquisition in a noisy rational expectations economy. *Econometrica*, *50*(6), 1415–1430. doi: 10.2307/1913389

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*, 1–48. doi: 10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W.-L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. doi: 10.1002/jrsm.11

von Neumann, J., & Morgenstern, O. (1944). *Theory of Games and Economic Behavior* (3rd ed.). Princeton, NJ: Princeton University Press.

von Hayek, F. A. (1937). Economics and knowledge. *Economica*, *4*(13), 33-54. doi: 10.2307/2548786

von Hayek, F. A. (1945). The use of knowledge in society. *American Economic Review*, *35*(4), 519-530.

von Mises, R. (1957). *Probability, statistics, and truth*. Allen Unwin.

Wald, A. (1948). *Sequential Analysis* (2nd ed.). New York: Wiley.

Wall, K. D. (1993). A model of decision making under bounded rationality. *Journal of Economic Behavior and Organization*, *21*, 331–352. doi: 10.1016/0167-2681(93)90030-S

Wang, Y., Wu, C., Yang, L., Wang, Y., & Wu, C. (2015). Hedging with Futures: Does Anything Beat the Naïve Hedging Strategy? *Management Science*. doi: 10.1287/mnsc.2014.2028

Wegwarth, O., & Gigerenzer, G. (2018). US gynecologists' estimates and beliefs regarding ovarian cancer screening's effectiveness 5 years after release of the PLCO evidence. *Scientific reports*, *8*(1), 1–9. doi: 10.1038/s41598-018-35585-z

Wegwarth, O., Wagner, G. G., & Gigerenzer, G. (2017). Can facts trump unconditional trust? evidence-based information halves the influence of physicians' non-evidence-based cancer screening recommendations. *PloS one*, *12*(8), e0183024. doi: 10.1371/journal.pone.0183024

Weisz, G., Cambrosio, A., Keating, P., Knaapen, L., Schlich, T., & Tournay, V. J. (2007). The emergence of clinical practice guidelines. *The Milbank Quarterly*, *85*(4), 691–727. doi: 10.1111/j.1468-0009.2007.00505.x

Weitzman, M. L. (1979). Optimal Search for the Best Alternative. *Econometrica*, *47*(3), 641–654. doi: 10.2307/1910412

Wierzbicki, A. P. (1982). A mathematical basis for satisficing decision making. *Mathematical Modelling*, *3*, 391–405. doi: 10.1016/0270-0255(82)90038-0

Winter, S. G. (1971). Satisficing, Selection, and the innovating remnant. *Quarterly Journal of Economics*, *85*(2), 237–261. doi: 10.2307/1880703

Winter, S. G. (2000). The satisficing principle in capability learning. *Strategic Management Journal*, *21*, 981–996. doi: 10.1002/1097-0266(200010/11)21:10/11<981::AID-SMJ125>3.0.CO;2-4

Witt, U. (1986). Firms' market behavior under imperfect information and economic natural selection. *Journal of Economic Behavior & Organization*, *7*(3), 265–290. doi: 10.1016/0167-2681(86)90032-6

Woolf, B. (1955). On estimating the relation between blood group and disease. *Annals of Human Genetics*, *19*(4), 251–253. doi: 10.1111/j.1469-1809.1955.tb01348.x

World Bank. (2020). *World development indicators.* `http://datatopics.worldbank.org/world-development-indicators/`. (Accessed: April 2020)

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(8), 1274. doi: 10.1037/xlm0000374

Wübben, M., & von Wangenheim, F. (2008). Instant Customer Base Analysis: Managerial Heuristics Often 'Get It Right'. *Journal of Marketing*, *72*, 82–93. doi: 10.1509/jmkg.72.3.082

Yamagishi, K. (2003). Facilitating normative judgments of conditional probability: Frequency or nested sets? *Experimental Psychology*, *50*, 97-106. doi: 10.1026//1618-3169.50.2.97

Yee, M., Dahan, E., Hauser, J. R., & Orlin, J. (2007). Greedoid-Based Noncompensatory Inference. *Marketing Science*, *26*(4), 532–549. doi: 10.1287/mksc.1060.0213

Yeung, T. M., & Mortensen, N. J. (2012). Assessment of the quality of patient-orientated internet information on surgery for diverticular disease. *Diseases of the Colon & Rectum*, *55*(1), 85–89. doi: 10.1097/DCR.0b013e3182351eec

Zahner, D., & Corter, J. E. (2010). The process of probability problem solving: Use of external visual representations. *Mathematical Thinking and Learning*, *12*(2), 177-204. doi: 10.1080/10986061003654240

Zhu, L. Q., & Gigerenzer, G. (2006). Children can solve Bayesian problems: The role of representation in mental computation. *Cognition*, *98*, 287-308. doi: 10.1016/j.cognition.2004.12.003

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x

List of Figures

List of Tables

Declaration

Information about co-authors can be found in the introduction to each chapter. Resources other than those listed as references were not used.

I testify through my signature that all information that I have provided about resources used in the writing of my doctoral thesis, about the resources and support provided to me as well as in earlier assessments of my doctoral thesis correspond in every aspect to the truth.

_____

Perke Jacobs, May 2021