Simple Heuristics That Make Us Smart

Gerd Gigerenzer

Peter M. Todd

and the ABC Research Group

New York Oxford

Oxford University Press

1999

Oxford University Press

Oxford New York
Athens Auckland Bangkok Bogotá Buenos Aires Calcutta
Cape Town Chennai Dar es Salaam Delhi Florence Hong Kong Istanbul
Karachi Kuala Lumpur Madrid Melbourne Mexico City Mumbai
Nairobi Paris São Paulo Singapore Taipei Tokyo Toronto Warsaw

and associated companies in
Berlin Ibadan

Copyright © 1999 by Oxford University Press, Inc.

Published by Oxford University Press, Inc.
198 Madison Avenue, New York, New York 10016

Oxford is a registered trademark of Oxford University Press

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior permission of Oxford University Press.

Library of Congress Cataloging-in-Publication Data
Gigerenzer, Gerd.
Simple heuristics that make us smart / Gerd Gigerenzer, Peter M. Todd, and the ABC Research Group.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-19-512156-2
1. Heuristic. I. Todd, Peter M. II. ABC Research Group.
III. Title.
BD260.G54 1999
128'.33—dc21 98-51084

9 8 7 6 5 4 3 2 1

Printed in the United States of America
on acid-free paper

5

How Good Are Simple Heuristics?

Jean Czerlinski
Gerd Gigerenzer
Daniel G. Goldstein

> Psychology has forgotten that it is a science of organism-environment relationships, and has become a science of the organism. . . . This . . . is somewhat reminiscent of the position taken by those inflatedly masculine medieval theologians who granted a soul to men but denied it to women.
>
> *Egon Brunswik*

Steve Bauer won the first day of the 1991 Tour de France, but placed 97th out of 200 at the end of the three-week race (Abt, 1991). Every day of this grueling bicycle tour covers a different type of terrain, and winning on one day does not guarantee good performance on the others. Likewise, Take The Best's success in Gigerenzer and Goldstein's (chapter 4) competition inferring German city populations does not guarantee that it will do well in other competitions. So before selling your sophisticated multiple regression software and converting to fast and frugal ways, heed this chapter. The strategies will complete two tours of 20 environments and predict everything from fish fertility to fuel consumption. The first tour will be data fitting: Strategies will train on the same course on which the race will be held. The second tour will be harder because the strategies will not be allowed to see the actual course until they race on it; they must train off-location. The results of these tours will pave the way for deciding when it pays to be fast and frugal and when it is better to use a more complex strategy such as multiple linear regression.

Meet the Environments

The glamour of the Tour de France is that it covers a wide variety of terrains, from flat to hilly to mountainous. Steve Bauer won the first day because he excelled on the plains, but lost the tour because he could not keep up on the mountains. The winner must be an all-rounder. Our tour is no different, consisting of 20 diverse environments. An environment consists of objects, each associated with a criterion to be predicted and a number of cues that may be helpful in predicting it. The task in the competition is to infer which of two objects scores higher on the criterion, for example, inferring which of two high schools has a higher dropout rate. Cues useful for making this inference could include the percentage of low-income students at the high school, the average SAT score, and the degree of parental involvement in their children's schooling. Our environments cover disparate domains from the objective number of car accidents on a stretch of highway to the subjective ratings of the attractiveness of public figures (table 5-1). The environments vary in size from 11 objects (ozone levels in San Francisco measured on 11 occasions) to 395 objects (fertility of 395 fish), and from 3 cues (the minimum needed to distinguish among the strategies) to 18 cues. To win this tour, an inference strategy will have to perform well in a variety of environments. Most of the environments come from statistics textbooks and are used to teach statistics, usually as examples of good applications of multiple regression. This should make it less than easy for Take The Best to compete with regression.

Meet the Competitors

Gigerenzer and Goldstein's competition (chapter 4) pitted a wide range of inference strategies against each other. Below we briefly describe four of the strategies; for more details see the previous chapter.

Take The Best

Imagine a bicycle built from the favorite parts of several racers, one contributing a frame, another a brake, a third a crankshaft. Instead of bicycle parts, Take The Best is assembled from cognitive building blocks: simple heuristics for search, stopping, and decision (see chapters 1 and 4 for definitions of these terms).

The first step of Take The Best is the recognition heuristic. In both tours, we will test the competitors in prediction tasks where all objects are recognized and all cue values are known; thus Take The Best will not be able to take advantage of the recognition heuristic, as it could in the previous chapter. Recall that Take The Best tries cues in order, one at a time, searching for a cue that discriminates between the two objects in question. For example, when inferring two professors' salaries, the rank cue might be tried first. If both professors are of the same rank (say both associate professors), then the gender cue might be tried. If one of the professors is a woman and the other is a man, then we say that the gender cue "discriminates." Once a discriminating cue is found, it serves as the

Table 5-1: A Description of the 20 Environments Used in the Competition

Psychology

Attractiveness of men: Predict average attractiveness ratings of 32 famous men based on the subjects' average likeability ratings of each man, the percentage of subjects who recognized the man's name (subjects saw only the name, no photos), and whether the man was American. (Based on data from a study by Henss, 1996, using 115 male and 131 female Germans, aged 17–66 years.)

Attractiveness of women: Predict average attractiveness ratings of 30 famous women based on the subjects' average likeability ratings of each woman, the percentage of subjects who recognized the woman's name (subjects saw only the name, no photos), and whether the woman was American. (Based on data from a study by Henss, 1996, using 115 male and 131 female Germans, aged 17–66 years.)

Sociology

High school dropout rates: Predict dropout rate of the 57 Chicago public high schools, given the percentage of low-income students, percentage of nonwhite students, average SAT scores, etc. (Based on Morton, 1995, and Rodkin, 1995.)

Homelessness: Predict the rate of homelessness in 50 U.S. cities given the average temperature, unemployment rate, percentage of inhabitants with incomes below the poverty line, the vacancy rate, whether the city has rent control, and the percentage of public housing. (From Tucker, 1987.)

Demography

Mortality: Predict the mortality rate in 20 U.S. cities given the average January temperature, pollution level, the percentage of nonwhites, etc. (Based on McDonald & Schwing, 1973; reported in StatLib.)

City population: Predict populations of the 83 German cities with at least 100,000 inhabitants based on whether each city has a soccer team, university, intercity train line, exposition site, etc. (From *Fischer Welt Almanach*, 1993.)

Economics

House price: Predict the selling price of 22 houses in Erie, PA, based on current property taxes, number of bathrooms, number of bedrooms, lot size, total living space, garage space, age of house, etc. (Based on Narula & Wellington, 1977; reported in Weisberg, 1985.)

Land rent: Predict the rent per acre paid in 58 counties in Minnesota (in 1977 for agricultural land planted in alfalfa) based on the average rent for all tillable land, density of dairy cows, proportion of pasture land, and whether liming is required to grow alfalfa. (Alfalfa is often fed to dairy cows.) (Data provided by Douglas Tiffany; reported in Weisberg, 1985.)

Professors' salaries: Predict the salaries of 51 professors at a midwestern college given gender, rank, number of years in current rank, the highest degree earned, and number of years since highest degree earned. (Reported in Weisberg, 1985.)

(continued)

Table 5-1: Continued

Transportation

Car accidents: Predict the accident rate per million vehicle miles for 37 segments of highway, using the segment's length, average traffic count, percentage of truck volume, speed limit, number of lanes, lane width, shoulder width, number of intersections, etc. for Minnesota in 1973. (Based on an unpublished master's thesis in civil engineering by Carl Hoffstedt; reported in Weisberg, 1985.)

Fuel consumption: Predict the average motor fuel consumption per person for each of the 48 contiguous United States using the population of the state, number of licensed drivers, fuel tax, per capita income, miles of primary highways, etc. (Based on data collected by Christopher Bingham for the *American Almanac for 1974*, except fuel consumption, which was given in the 1974 *World Almanac*; reported in Weisberg, 1985.)

Health

Obesity at age 18: Predict fatness at age 18 of 46 children based on body measurements from age 2 to age 18. The body measurements included height, weight, leg circumference, and strength. (Based on the longitudinal monitoring of the Berkeley Guidance Study, Tuddenham & Snyder, 1954; reported in Weisberg, 1985.)

Body fat: Predict percentage of body fat determined by underwater weighing (a more accurate measure of body fat) using various body circumference measurements (which are more convenient measures than underwater weighing) for 218 men. (Data supplied by A. Garth Fisher from the study of Penrose et al., 1985; reported in StatLib.)

Biology

Fish fertility: Predict the number of eggs in 395 female Arctic charr based on each fish's weight, its age, and the average weight of its eggs. (Data courtesy of Christian Gillet, 1996.)

Mammals' sleep: Predict the average amount of time 35 species of mammals sleep, based on brain weight, body weight, life span, gestation time, and predation and danger indices. (From Allison & Cicchetti, 1976; reported in StatLib.)

Cow manure: Predict the amount of oxygen absorbed by dairy wastes given the biological oxygen demand, chemical oxygen demand, total Kjedahl nitrogen, total solids, and total volatile solids for 14 trials. (Moore, 1975; reported in Weisberg, 1985.)

Environmental Science

Biodiversity: Predict the number of species on 26 Galapagos islands, given their area, elevation, distance to the nearest island, area of the nearest island, distance from the coast, etc. (Based on Johnson & Raven, 1973; reported in Weisberg, 1985.)

Rainfall from cloud seeding: Predict the amount of rainfall on 24 days in Coral Gables, FL, given the types of clouds, the percentage of cloud cover, whether the clouds were seeded, number of days since the first day of the experiment, etc. (From Woodley et al., 1977; reported in Weisberg, 1985.)

Oxidant in Los Angeles: Predict the amount of oxidant in Los Angeles for 17 days given each day's wind speed, temperature, humidity, and insolation (a measure of the amount of sunlight). (Data provided by the Los Angeles Pollution Control District; reported in Rice, 1995.)

Table 5-1: Continued

Ozone in San Francisco: Predict the amount of ozone in San Francisco on 11 occasions based on the year, average winter precipitation for the past two years, and ozone level in San Jose, at the southern end of the Bay. (From Sandberg et al., 1978; reported in Weisberg, 1985.)

Note. For each environment we specify the criterion, a sample of the cues for predicting the criterion, and the source of the data. Recall that the cues are either binary or were dichotomized by a median split, and that the task is always to predict which of two objects scores higher at the criterion.

basis for an inference, and all other cues are ignored. For instance, if gender discriminates between two professors, the inference is made that the male earns a higher salary, and no other information about years of experience or highest degree earned is considered. Could such one-reason decision making be accurate? This chapter will answer this question.

Since Take The Best does not integrate information or require extensive computations, it is fast. Since it has a stopping rule to effect limited search for cues, it is frugal. In this competition, Take The Best looks up cues in the order of their validities, which it has to estimate from a training set. Recall that the validity of a cue is defined as the number of correct inferences divided by the number of correct and incorrect inferences made using the cue alone (chapter 4).

The Minimalist

The fast and frugal Minimalist looks up cues in a random order, stopping when it finds a cue that discriminates between the two objects. Otherwise, it is exactly the same as Take The Best. In the simulation, it will not be able to take advantage of the recognition heuristic, for the same reason as for Take The Best.

Multiple Regression

Multiple linear regression is the most thoroughly trained and well-equipped rider in the pack. It rides on sophisticated computations rather than on fast and frugal building blocks. Regression assumes the data can be approximated by a hyperplane plus independent, identically distributed errors with zero mean. It then finds the hyperplane that minimizes the squared vertical distance between the hyperplane and the data points. Finding an optimal fitting surface is not the kind of calculation that can be easily carried out with pencil and paper or a standard pocket calculator—a computer is called for. When regression is used to make a prediction, all of the available cues must be gathered and plugged into the model, so it is not frugal. Furthermore, since multiple regression requires extensive computations, it is not fast.

Dawes's Rule

Dawes's rule is a simplification of regression. The model is still linear, but instead of optimal weights, only unit weights (+1 or −1) are used (Dawes, 1979). That is, it adds up the number of pieces of positive evidence and subtracts the number of pieces of negative evidence. We operationalize the assignment of the unit weights by giving a cue a weight of +1 if a cue's validity is above chance (.5) and −1 if it is below chance. Since using Dawes's rule to make a prediction still requires all of the cues, it is not frugal. But unlike regression it is fast, since the weighting scheme is trivial.

We have defined athletes with four differing strategies. Who will win the 20-environment tour?

The First Tour: Fitting Known Environments

In our first tour, the riders were allowed to examine every detail of the race course before the competition. There were no missing cue values or unrecognized objects, unlike the scenarios in chapters 2, 3, and 4. All of the cue and criterion values were available for calculating cue validities or linear weights. The strategies then predicted the criterion values (which they had already seen). This type of contest is called data fitting. The first two environments that were fit, high school dropout rates and professorial salaries, will be described in detail to give a sense of how the strategies compete against one another. Then we will jump to the end of the race to see who won the overall tour and by how much.

Dropping Out

The first stage of the tour is important for American society: predicting dropout rates at Chicago public high schools. The 1995 rates were published in *Chicago* magazine (Morton, 1995; Rodkin, 1995), along with possible cues such as the socioeconomic and ethnic compositions of the student bodies, the sizes of the classes, the attendance rates of the students, the parent participation rates, and the scores of the students on various standardized tests.

We prepared the raw data from the magazine to suit the four inference strategies. We converted all cue values that were real numbers into ones and zeroes using the median as a cutoff. These ones and zeroes were assigned such that the ones corresponded to higher values on the criterion.

After the data were transformed into the appropriate format, their characteristics could be measured. Overall, the dropout environment looked fairly challenging. The average cue validity was only .63, compared to .76 for the German city population data. The maximum cue validity was also rather low, .72. These characteristics should create considerable difficulties for Take The Best, which relies on only the best cue that discriminates between two high schools. Furthermore, the environment comprised a total of 18 cues, double the number in the city population environment. Since Dawes's rule improves in accuracy with the addition of more cues (see chapter 6), this environment was a particularly tough test for Take The Best and the Minimalist.

Before revealing the accuracy of the strategies in predicting dropout rates, let us review the results on German city populations (figure 4-2; the values on the right of the graph where all objects are recognized). Recall that multiple regression made 74% correct inferences. Dawes's rule did very well in comparison, also earning 74% correct. The surprising finding was that Take The Best matched the 74% performance of these linear strategies. Finally, the exceedingly simple Minimalist scored a respectable 70% correct.

What happened on the more difficult high school dropout environment? Despite the lower cue validities, regression was still able to get 72% of the inferences correct (table 5-2, "fitting"). Perhaps the large number of cues made up for the low validities. Dawes's rule did not seem to be able to take as much advantage of the many cues, getting only 64% correct. Take The Best made 65% of the inferences correctly—slightly better than Dawes's rule but still seven percentage points behind the performance of linear regression. The Minimalist was again the weakest strategy, but not too far behind Take The Best with 61% correct. Take The Best and the Minimalist looked up on average only a few cues (table 5-2). Speed and frugality paid the price of seven percentage points in lost accuracy on the difficult high school dropout data.

Policy Implications Discovering which strategy best fits the data can have important consequences for public policy. For example, Take The Best regarded attendance rate, writing test score, and social science test score

Table 5-2: Predicting High School Dropout Rates

		Accuracy (% Correct)	
Strategy	Frugality	Fitting	Generalization
Minimalist	2.7	61	58
Take The Best	3.4	65	60
Dawes's rule	18	64	62
Multiple regression	18	72	54

Note. Performance of two fast and frugal heuristics (Minimalist, Take The Best) and two linear strategies (Dawes's rule, multiple regression) in predicting which of two Chicago high schools has a higher dropout rate. There were 57 public high schools and 18 predictors (table 5-1). Performance is measured in terms of frugality (average number of cues looked up) and accuracy (% correct). Accuracy is measured both for fitting data (test set = training set), and for generalization (test set ≠ training set). The average number of cues looked up was about the same for both kinds of competition.

as the most valid cues for dropout rate, in that order. In contrast, linear regression's top three predictors were percentages of Hispanic students, students with limited English, and black students. Thus, each strategy led to different implications for how we can help schools lower dropout rates. While a user of Take The Best would recommend getting students to attend class and teaching them the basics more thoroughly, a regression user would recommend helping minorities assimilate and supporting English as a second language (ESL) programs. Because regression resulted in the best fit, it looked like the regression user would be able to give better advice for lowering dropout rates.

Professors' Income

Let us now consider how well the strategies predict individual professors' salaries from the following five cues: gender, rank (assistant, associate, full professor), number of years in current rank, highest degree earned, and number of years since degree earned. The data is from a midwestern college, which shall remain anonymous. Clearly, this environment already had one binary variable (gender); the rest were dichotomized at the median.

This environment had a maximum cue validity of .98 and its average cue validity was .79, similar to the city population environment. It had only five cues, about half as many as for predicting populations. How would this affect the accuracy of Take The Best and the Minimalist? One intuitive answer would be that high cue validities and few cues allow the two heuristics to keep up with the algorithms that integrate information across cues; let us see if this was true.

Crossing the finish line first was the rider on the fanciest and most expensive bicycle, multiple regression, with a stunning 83% correct (table 5-3, "fitting"). This was surprising since the environment seemed to be

Table 5-3: Predicting Professors' Salaries

		Accuracy (% Correct)	
Strategy	Frugality	Fitting	Generalization
Minimalist	2.1	73	72
Take The Best	2.3	80	80
Dawes's rule	5	75	75
Multiple regression	5	83	80

Note. Performance of two fast and frugal heuristics (Minimalist, Take The Best) and two linear strategies (Dawes's rule, multiple regression) in predicting which of two professors at a midwestern college has a higher salary. There were 51 professors and five predictors (table 5-1). Performance is measured in terms of frugality (average number of cues looked up) and accuracy (% correct). Accuracy is measured both for fitting data (test set = training set), and for generalization (test set ≠ training set). The average number of cues looked up was about the same for both kinds of competition.

about the same as the cities except with half as many cues. Taking one cue at a time, Take The Best somehow managed second place by scoring 80% correct. Dawes's rule, the leaner linear model, got 75% correct, not as far behind linear regression as it was with school dropout rates. The Minimalist finally pulled in at 73%, almost as good as Dawes's rule.

It turned out that the best cue for predicting professor salary was rank, with a cue validity of .98. It may not come entirely as a surprise that the second best cue was gender, with a validity of .88. In this environment, regression was mostly in agreement, giving rank the greatest weight, followed by highest degree earned and gender.

The Overall Winner of the First Tour

We now have a sense of how the competition works and how the characteristics of the environments might affect the strategies. Let us finally find out which strategy won on the complete range of environments, that is, fitting both the mountain roads and the plains closely enough to win the overall tour.

How frugal were the heuristics? The Minimalist searched for only 2.2 cues on average to make an inference. Take The Best needed slightly more cues, 2.4, whereas the two linear strategies always used all the available information, 7.7 cues on average (the linear strategies have no heuristics for search and stopping). Thus, the two heuristics looked up fewer than a third of the cues. If they are so frugal, how accurate can they be?

Perhaps it is no surprise that the first-place finisher was multiple linear regression, which used all information and subjected it to complex computation (table 5-4, "fitting"). Across the 20 environments, regression scored 77% correct. However, the second-place finisher may be a surprise. The

Table 5-4: Performance Across 20 Data Sets

		Accuracy (% Correct)	
Strategy	Frugality	Fitting	Generalization
Minimalist	2.2	69	65
Take The Best	2.4	75	71
Dawes's rule	7.7	73	69
Multiple regression	7.7	77	68

Note. Performance of two fast and frugal heuristics (Minimalist, Take The Best) and two linear strategies (Dawes's rule, multiple regression) across all 20 data sets. The average number of predictors was 7.7. Performance is measured in terms of frugality (average number of cues looked up) and accuracy (% correct). Accuracy is measured both for fitting data (test set = training set), and for generalization (test set ≠ training set). The average number of cues looked up was about the same for both kinds of competition. For a similar result with slightly different data sets, see Gigerenzer et al. (1999), and for the performance of various strategies on the 20 individual data sets, see table 8-1.

fast and frugal Take The Best finished the tour only two percentage points behind regression, with 75% correct. This is close to what Gigerenzer and Goldstein (chapter 4) found, suggesting that our set had more cases similar to the city population data than to the high-school dropout data. The fast but not frugal Dawes's rule scored two percentage points behind Take The Best with 73% correct. It was quite a surprise that Dawes's rule scored worse than Take The Best, given that Take The Best was even more frugal and did not integrate what little cue information it did gather. Finally, the Minimalist pulled in last with 69% accuracy, a respectable score considering its extreme simplicity. The price of using a fast and frugal heuristic was small, about two percentage points for Take The Best and about eight for the Minimalist. Furthermore, more cue information did not guarantee more accuracy, since Take The Best was slightly more accurate than Dawes's rule despite using fewer cues.

We Knew It All Along

One reaction to a novel claim is to say that it is impossible. Gigerenzer and Goldstein (1996a) showed that their claim—that fast and frugal heuristics can also be accurate—was possible; and this chapter has further shown that it is not only sometimes possible but is, in fact, often the case. Environments in which the price of simplicity is high, such as when predicting dropout rates in high schools, seem to be the exceptions and not the rule.

Another reaction when an "impossible" novel claim has finally been proven is to say one "knew it all along" (see chapter 9 on hindsight where this memory distortion is modeled by Take The Best). In this section, we review the psychological literature to find what actually was known all along about how well fast and frugal heuristics can perform relative to more complex strategies.

The comparison is not entirely straightforward because earlier research differs from ours in a number of ways. First, the range of strategies compared in earlier studies was mostly restricted to different weighting schemes for linear models. Second, the range of environments was typically restricted to artificially generated data sets with multivariate normal distributions for the cues and criteria. Finally, the type of competition differed, usually involving not just fitting given data, but generalizing to new data, that is, training an algorithm on one part of the data and then making predictions on another part. If both parts are of equal size, this is usually called cross-validation. In the next section we will rerun the whole competition using cross-validation. First let us consider the previous literature.

Research on simple strategies began in earnest in the mid-1970s. Through computer simulations and mathematical analysis, researchers such as Schmidt (1971), Dawes and Corrigan (1974), and Einhorn and Hogarth (1975) found that a unit-weighted linear model (which we call Dawes's rule) was on average almost as accurate as multiple linear regression—and far more robust to boot. (A "robust" strategy or model is one that remains accurate when generalizing to new data, such as in cross-validation.) For example, in predicting grade point averages, a unit-weighted linear model made predictions that correlated .60 with the actual values, while a cross-validated regression model scored .57. Note that because regression was cross-validated—making predictions on data different from that on which it was trained—its performance can be lower than the unit-weighted model (which was not cross-validated). In the three other tasks considered, unit weights had a higher accuracy than cross-validated regression in two (Dawes & Corrigan, 1974). As Paul Meehl put it, "in most practical situations an unweighted sum of a small number of 'big' variables will, on the average, be preferable to regression equations" (quoted in Dawes & Corrigan, 1974, p. 105).

In a related but more recent line of research, Ehrenberg (1982) analytically compared regression weights to other weights. He showed that for typical values of a one-cue prediction problem (e.g., with a correlation of .7 between the criterion and the cue), using a slope differing from the optimal by as much as plus or minus 30% results in only a 4% increase in unexplained error. Dawes and Corrigan (1974, citing an unpublished manuscript by Winterfeldt & Edwards, 1973) called this the phenomenon of the flat maximum: Weights even vaguely near the optimal lead to almost the same output as do optimal weights.

These studies seem to say that Dawes's rule is often almost as accurate as multiple regression. But life is not quite that simple. First of all, in those cases in which real environments were used, only a few such environments were checked. Second, this research cross-validated only for regression but not for Dawes's rule, with the argument that "it is the human judge who knows the directional relationship between the predictor variables and the criterion of interest" (Dawes, 1979, p. 573). But even experts must have some method by which they estimate the direction of cues, and so the cross-validated simulations in the next section test how well Dawes's rule performs when it must estimate the direction of the cue, too. We will wait until a later section to operationalize Meehl's suggestion of using only "a small number" of variables; we will continue to use all the cues for now. Our work goes beyond previous research by operationalizing all aspects of Dawes's rule—testing cross-validated Dawes's rule against cross-validated regression—and seeing if the old findings still hold up.

We also go beyond previous research in pursuing the trade-off question more intensively: Just how much simpler can inference strategies be without losing too much accuracy? Our simulations test not just Dawes's rule against regression but also against Take The Best and the Minimalist. There have been some scattered experiments also trying very simple heuristics (e.g., Hogarth & Makridakis, 1981; Kleinmuntz & Kleinmuntz, 1981), but only Payne, Bettman, and Johnson (1988, 1990) have launched a consistent program of study. Their program focuses on preferences (e.g., between gambles) rather than on cue-based inferences, and they measure performance by a correlation of choices with a weighted additive model (the expected payoff) rather than with an external criterion (since for subjective choice there is none). For example, in their competitions the simpler heuristics typically achieved from 60 to 70 percent of the performance of the weighted additive model benchmark, but by this measurement method, the simple heuristics cannot be more accurate than the "rational" answer of the weighted additive model. Only with an external standard for the number of correct inferences is it possible to show that simple heuristics can be more accurate than more complex strategies. Thus, previous research has focused on preferences rather than inferences, and on artificial rather than real-world environments. As a consequence, it has not shed much light on the accuracy of simple heuristics in making inferences about the real world.

In this chapter, we test heuristics on a wider range of empirical data environments than has been used before. We run the Tour de France of heuristic decision makers.

he Second Tour: Generalizing to New Objects

Imagine a bicycle rider who spent all his time training on the plains of the Midwest and then tried to race in the varied landscapes of the Tour de France. What would probably happen? He might fail completely on the mountains. This is not to say he would have to go to France to train; as long as he could find a mixture of Colorado mountains, midwestern plains, and winding New England streets, he could adequately prepare for the Tour. Training on a course and racing (testing) on another is generalization, as opposed to fitting.

More precisely, generalization means that the strategies build their models (i.e., calculate regression weights, determine cue orders or cue directions, etc.) on some subset of all objects, the training set. The strategies then make predictions about the remaining objects, the test set. Generalization is a more difficult and realistic test of the strategies than training and testing on the same objects. In our simulations, we tested generalization by breaking the environment into halves, with a random assignment of objects to one half or the other. This is called cross-validation. The performance is then the proportion correct in the test set. Each environment was split 1000 times into training and test sets, in order to average out any particularly helpful or harmful ways of dividing the data.

Dawes's rule and the Minimalist might not seem to be doing any estimation, but in fact they use the first half of the environment to estimate the direction in which the cues point. In our simulations, they did this by calculating whether the cue validity was above or below the chance level. (This is equivalent to testing whether the cue has a positive or negative Goodman-Kruskal rank correlation with the criterion, as shown in chapter

6.) Take The Best estimates the direction of the cues and then orders them from best to worst predictor. Multiple regression estimates the optimal beta weights, taking into account the relationships between the variables.

We will now race through each of the environments and consider how the strategies perform in generalization. In chapter 4, we saw one case of generalization, predicting city populations. When the algorithms were trained on half of the cities, as in our second tour, Take The Best was slightly more accurate (72%) than multiple regression and Dawes's rule (71% each). Does this result generalize? Could it be that one-reason decision making can be more accurate across the 20 environments?

Dropping Out Again

We first tested generalization when predicting high-school dropout rates. The simplest strategy, the Minimalist, fell from 61% to 58%; similarly, Dawes's rule dropped from 64% to 62%. Take The Best, which estimates both cue direction and cue order, took a slightly larger loss and dropped to 60% (table 5-2, "generalization"). Finally, multiple regression dropped a whopping 18 percentage points, from 72% to 54%, which also made it by far the least predictive strategy of the bunch. It seems the simpler strategies are the more robust ones in generalization. What explains regression's huge drop?

We believe the answer is overfitting. Imagine that a bicycle rider trains on a course beginning with a steep ascent, continuing with a long, flat plain, and ending with a final descent, having exactly the same proportion of uphill, flat, and downhill regions as the test course will have. Every day, the rider's body gets used to pumping hard and heavy at first, then cruising quickly, then relaxing on the way down. The danger is that the rider may get so used to this pattern that he can no longer deal well with other combinations of hills and plains: If the test course is a drop, then a flat plain, ending with a steep ascent, the rider might have difficulties adjusting. Such overfitting can happen to inference strategies, too: They can learn the particular quirks of their training data, such as details of cue orders and intercorrelations, too well. The more closely a strategy tries to fit the training landscape, the greater is the danger of overfitting.

In the case of the dropout environment, there were 18 cues. Such an abundance of cues offered ample opportunity for accidental correlations. If regression built these accidents into its model, then its predictions on the second half of the data, which need not have the same accidental correlations as the first half, would be inaccurate.

Public Policy Again, consider possible policy implications. In the fitting tour, the regression user would have confidently recommended expanding ESL classes to help the dropout rate. Because regression had the best predictions, this would seem the best policy. However, in the generalization tour, Take The Best was more accurate than regression, and it appears that regression overfitted the training data. While regression put a heavier weight on the influence of ESL classes on dropout rates based on the training data, this may have been a fluke that will not generalize. On the other hand, Take The Best's recommendation, based on the training data, to encourage attendance and teach the basics may be more generalizable. Dawes's rule and the Minimalist do not suggest specific recommendations—just to improve on all fronts—because they weight all predictors equally.

Later in this chapter, we will argue that regression carries a lower risk of overfitting in larger environments with more objects (or fewer cues). There are, though, only 57 public high schools in Chicago, and if this number of objects does not suffice for regression, then regression simply should not be used. There is no more data to collect for it. One might try to train regression on dropout rates from other cities or from previous time periods in Chicago, but then one risks again overfitting, finding factors relevant to other places and times than to today's Chicago public high schools.

Professors' Income

Let us also briefly consider generalizing predictions of professors' salaries based on five cues. Regression's score dropped only slightly, from 83% to 80%. Dawes's rule and Take The Best held their ground at 75% and 80%, respectively (table 5-3, "generalization"). The Minimalist dropped one point to 72%. Compared with the task of predicting high school dropout rates, predicting income was based on a smaller number of cues and on cues with higher validity. It seemed that with these characteristics, the drop in accuracy also was less than was the case in predicting city populations.

We now have some idea of how generalization affects the strategies. Which of the strategies will make robust generalizations across the 20 environments?

The Winner of the Second Tour

On average, regression dropped a stunning nine percentage points in accuracy, from an average of 77% for the fitting task to 68% for generalization (table 5-4, "generalization"). Meanwhile, Dawes's rule fell four percentage points, from an average of 73% to 69%, as did the Minimalist, from 69% to 65%. The small size of these drops was probably due in part to the fact that these two strategies estimated very little—only the directions of the cues.

The overall winner was Take The Best at 71% accuracy, down five percentage points. Take The Best earned the highest accuracy in generalization among the four strategies, despite its fast and frugal nature. Across the 20 environments, regression and Dawes's rule used an average of 7.7 cues per inference, whereas Take The Best only used 2.4 cues, the same small number as for the fitting task. The fact that a heuristic can disobey the rational maxim of collecting all available information and yet be the most accurate is certainly food for thought.

Take The Best outperformed multiple regression by an average of three percentage points when making generalizations. Startling at this result is, it is not entirely inconsistent with the previous literature, which showed that in several types of environments regression generalized less well than the simpler Dawes's rule (without cross-validation). Our second tour, however, has shown that Dawes's rule is also in danger of overfitting. Moreover, simplicity and frugality, pushed to the extreme, can eventually have a price: The Minimalist placed last. But this price was not very high, for the Minimalist's average performance was a mere three percentage points behind multiple regression.

Tinkering With the Rules of the Tour

Some colleagues were skeptical about the possibility that one-reason decision making could be fast, frugal, and accurate at the same time. They suggested modified versions of the competition, predicting that the counterintuitive accuracy of Take The Best would quickly vanish. One early conjecture voiced against the results reported in Gigerenzer and Goldstein (1996a; see also chapter 4), was that the recognition heuristic, with its high empirical validity (.8) for population size, would be the main cause for the accuracy of Take The Best. We have taken care of this conjecture in this chapter: In both tours, all objects were recognized, so that the recognition heuristic could not operate. We will consider four further modifications and conjectures.

Use exact rather than dichotomized numbers. In the simulations reported, we have dichotomized all quantitative cues at the median (except for the binary cues, such as gender) rather than using the exact values. This procedure was assumed to mimic the limited knowledge about cue values that people typically have, and the potential unreliability of precise values. Each competitor, the linear strategies and the heuristics, based their predictions on these binary or dichotomized values. A reasonable conjecture is that part of the power of multiple regression is lost when it is applied to dichotomized data. Some colleagues suggested rerunning the tour and letting every strategy have the exact quantitative cue values. There are two major ways lexicographic strategies such as Take The Best can be extended to make inferences from quantitative values (Payne et al., 1988, 1990). In the first version, search continues until the first cue is found in which the two values are different; in the second, search continues until the difference between cue values exceeds a threshold or "just noticeable difference." To avoid the arbitrariness in defining how large the threshold should be, we went with the first version.

We reran Take The Best and multiple regression under the conditions of the second tour. Take The Best, when adapted to quantitative values, was even more frugal than its standard version. Search often stopped after the first or second cue, because even small quantitative differences were sufficient to halt search and prompt a decision. But how accurate were the inferences based on quantitative predictors? Our colleagues were right: Multiple regression did improve when given real numbers—but so did Take The Best. Across the 20 environments, Take The Best made 76% correct predictions, compared to regression, which also earned 76% correct. Thus, one-reason decision making in the form of Take The Best could still match multiple regression in accuracy, even with exact quantitative values. This counterintuitive result came as a surprise to us, but, by then, we were getting used to surprises.

Give Dawes's rule another chance by using only the "big" cues. Recall Paul Meehl's conjecture: "an unweighted sum of a small number of 'big' variables will, on the average, be preferable to regression equations." In the two tours, Dawes's rule had access to all of the cues. Meehl, however, suggested using only the "big" cues, that is, the most valid ones. When Dawes's rule estimates the direction of a cue in the training set, erroneous estimates of the direction occur most often with low-validity cues. We developed a version of Dawes's rule that ignored cues with an estimated validity of .7 or less. We reran this truncated version of Dawes's rule under the conditions of the second tour. The accuracy indeed increases from 69% with Dawes's rule (see table 5-4) to 71% with the truncated Dawes's rule, three percentage points above the accuracy of regression. Meehl's intuition turned out to be correct across the 20 environments. Using only a small number of "big" cues, without weighting them, is, on average, more accurate in *generalization* than a regression that weights all cues. But Meehl's intuition can be pushed even further. Take The Best, which uses only the "best" cue that discriminates between two objects, turns out to be as accurate.

But what if Take The Best does not have the order of cues (as in Tour 1), and needs to estimate it from a very small sample? Recall first that Take The Best does not try to estimate an optimal order of cues (as, for instance, classification trees attempt to do; see chapters 6 and 8). Instead, it uses a simple and frugal method to create an order (for binary cues, the cue order can be calculated with one simple pass through the objects; see Czerlinski, 1998 for details). Ordering cues by their cue validity, as Take The Best does, is not called "optimal" because this procedure ignores all dependencies between cues. We saw that, for estimating population size, Take The Best lost on accuracy when the training set was very small, but multiple regression lost even more (figure 4-3). These results suggest that Take The Best is relatively robust when making predictions from a small number of observations. But does this result generalize to other data sets? We tested the accuracy of Take The Best when it had to estimate the cue order from just 10 randomly chosen objects rather than from the full first half of the objects, as in the second tour. It then made predictions on the other half of the environment just as in the regular second tour.

This tour tested the degree to which Take The Best depends on copious information for assessing the order in which to try cues. The result across 20 environments was that Take The Best scored about 66% correct predictions, losing five percentage points from when it had access to half the environment for training. These numbers match the result in estimating city populations in chapter 4. This Sample-10 tour is similar to allowing bicycle riders only very limited training, say by announcing the layout of the race course only days ahead of the race. It seemed that even with very few observations, Take The Best could still make reasonably accurate predictions.

But Take The Best cannot estimate quantities, whereas multiple regression can. This conjecture addresses the generality of Take The Best, Take The Last, and the Minimalist. These three heuristics can make predictions about which object has a higher value on a criterion, such as which of two highways is more dangerous, but they cannot make quantitative predictions, such as how high the car accident rate on one of those two highways is. The heuristics are specialized for particular classes of tasks, whereas multiple regression is more general. In chapter 10, we will study a heuristic that can make quantitative predictions, and employs one-reason decision making like Take The Best. What we call the adaptive toolbox is a collection of different heuristics designed from the same kinds of building blocks. The building blocks, not the specific heuristics, have generality. The specificity of the individual heuristics enables them to be fast, frugal, and accurate—with little trade-off.

How Does Take The Best Do So Well?

What is the difference between the environments in which Take The Best performed poorly and those in which it did well? This question concerns the ecological rationality of Take The Best, that is, the fit between the structure of the heuristic and that of an environment (chapter 1). This question is the focus of the next chapter—here we will raise the question, review previous research, and test three of its predictions empirically.

Characterizing Environments: A Review

The literature suggests variables that matter in predicting whether cross-validated regression or Dawes's rule without cross-validation would win our competition. Our goal, however, was to understand the performance of fast and frugal heuristics, so it is unclear whether these earlier findings are relevant. Furthermore, previous research has used either simulation studies of hundreds of randomly generated environments, or mathematical analysis with numerous simplifying assumptions about the form of the data. There is cause to doubt whether such findings would generalize to our empirical environments. Let us nevertheless consider what has been discovered, but with these caveats in mind.

Schmidt's (1971) simulations on random data (multivariate normal distributions) showed that in cross-validation regression outperformed Dawes's rule on average only for large numbers of objects. For example, with four cues, one needs a sample size of at least 50 objects for regression to beat Dawes's rule. For six cues, one needs at least 75 objects. For 10 cues, 100 objects are required. As a rule of thumb, it seems one should not use regression with fewer than 10 cues per object; otherwise, unit weights will outperform regression weights on average. Regression is likely to overfit the data when there are too few objects for the number of cues. That is, regression takes account of numerous intercorrelations that may be artifacts of the current sample. Similarly, the fewer kinds of training course a bicycle rider is exposed to, the more likely she is to overfit the ones she has seen.

Einhorn and Hogarth (1975; see also Hogarth, 1981) confirmed Schmidt's findings and added two other factors. Dawes's rule can be expected to perform about as well as multiple regression when (a) the coefficient of determination (R^2), from the regression model is in the moderate or low range (.5 or smaller) and (b) the cues are intercorrelated. The coefficient of determination measures the linear fit between the criterion and the cues.

If we take these three factors together, the literature indicates that regression models are slightly more accurate than Dawes's rule if there are many objects per cue, a high linear predictability of the criterion, and cues are not highly correlated. Under the opposite conditions, Dawes's rule is slightly better.

We shall now see if number of objects per cue, ease of linear predictability, or degree of cue intercorrelation can explain why Take The Best was so successful in the real-world environments of the two tours.

Is It the Number of Objects per Cue?

Figure 5-1 shows that the advantage of Take The Best over multiple regression (the difference in accuracy) depends on the ratio between the number of objects and cues. Take The Best won by more when there were fewer objects per cue. However, a large number of objects per cue, even more than 10, did not guarantee that regression would tie or outperform Take The Best. For instance, the largest ratio in figure 5-1 was obtained for predicting body fat, with 218 men measured on 14 cues, resulting in about 16 objects per cue. Even with this high ratio of objects per cue, Take The Best made a higher proportion of accurate inferences than multiple linear regression.

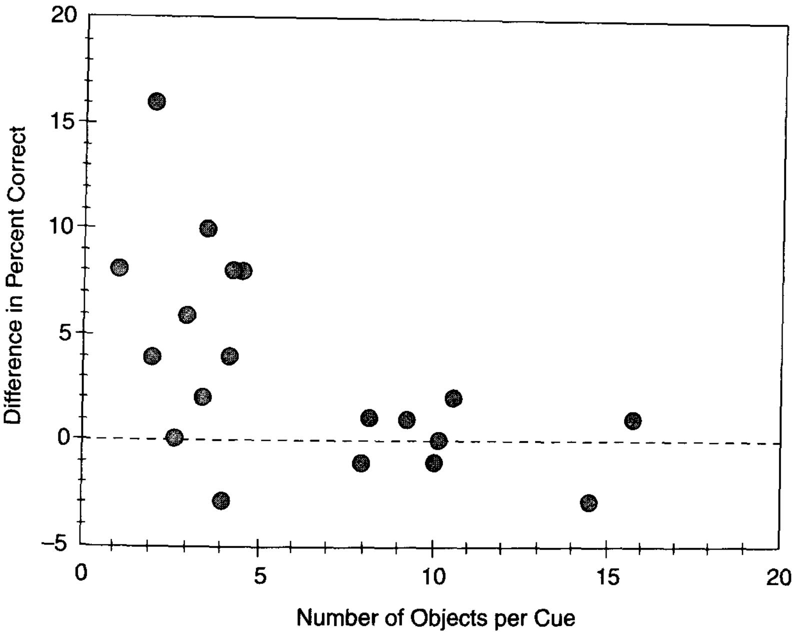

Figure 5-1: Take The Best's advantage over multiple regression (in percent correct) plotted against the number of objects per cue for 19 of the 20 environments (Tour 2). The missing environment is fish fertility, which is well off the scale with 395 objects and 3 cues. Take The Best scored 1.8 percentage points behind regression on the fish fertility data.

The finding that the number of objects per cue was a good predictor of Take The Best's advantage provides support to the hypothesis that regression was overfitting when there were few objects per cue. There is a more direct test of this hypothesis: Compare cross-validated regression with regression that merely fits the data, as the number of objects per cue is varied (figure 5-2). The result shows a trend similar to that in figure 5-1: The smaller the number of objects per cue, the larger the difference between the performance of regression in the fitting and generalization tasks. In both figures, the plots have nonconstant variance and appear curved.

Is It the Ease of Linear Predictability?

The second characteristic of envi^ronments is the coefficient of determination (R^2). The idea is that in environments with a high coefficient of determination, multiple regression results in better predictions than Dawes's rule, while the exact weighting scheme does not matter much for data that is not very linear anyway. We measured R^2 by running regression on the full environment. However, there appeared to be no relation between R^2

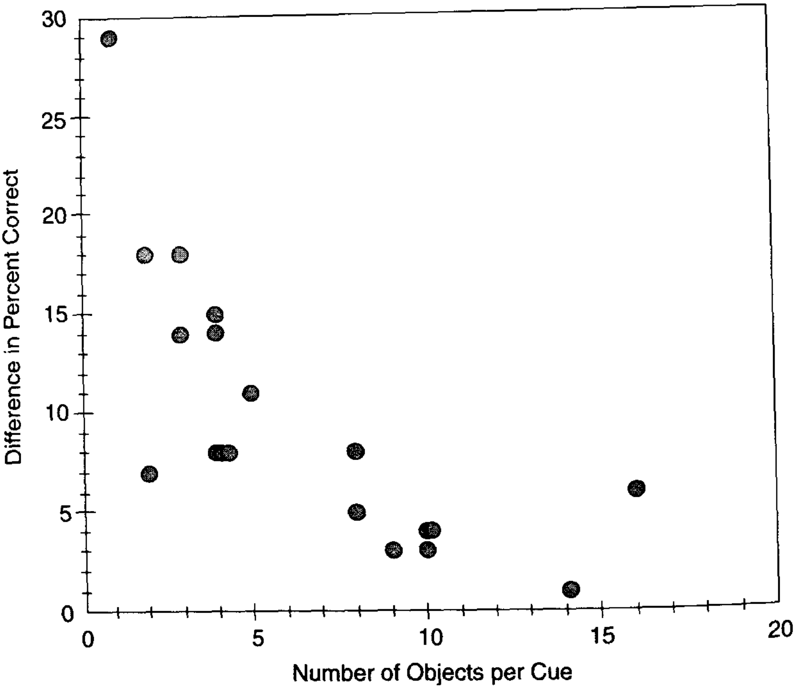

Figure 5-2: The accuracy of multiple regression in the fitting task (Tour 1) minus its accuracy in the generalization task (Tour 2) plotted against the number of objects per cue for 19 of the 20 environments. The fish fertility environment is again omitted. In data fitting, regression scored 0.4 percentage points higher than in cross-validation on this data set.

and the difference in accuracy between Take The Best and multiple regression.

Is It the Cue Intercorrelation?

If all cues were perfectly correlated ($r = 1.0$), then one-reason decision making would be as accurate as a linear combination of all cues. Therefore, several of our colleagues have suggested that the higher the intercorrelation between cues, the greater the advantage Take The Best has over regression. (Some others proposed the opposite relationship.) To test this hypothesis, we measured the correlations between each pair of cues in each environment, took the absolute values, and then averaged them. Will environments with high average absolute correlations give Take The Best more of an advantage over regression? We found no trend, so this variable does not seem to explain Take The Best's success. Nor did one of the following variables: the maximum cue intercorrelation, the minimum cue intercorrelation, and the variance of the correlations.

What structures of real-world environments does Take The Best exploit in order to perform so well? From the three characteristics reported in studies that compared Dawes's rule with regression, only one, the ratio between the number of objects and the number of cues, was related to the advantage Take The Best had over regression. Thus, the hypotheses derived from previous work comparing Dawes's rule with regression only partially shed light on the question of which environment structures can be exploited by fast and frugal heuristics. The following chapter offers more insight into this question, based on mathematical intuition and proof.

What We Have Learned

To control for the natural reaction to all new results, the "I knew it all along" reflex, we had asked several prominent researchers in judgment and decision making to predict how close Take The Best would come to multiple regression in accuracy. These researchers were expert on noncompensatory strategies and multiple regression. Their predictions were consistent: They bet on between 5 and 10 percentage points more accuracy for multiple regression. Our results surprised them as much as us.

In this chapter, we have considered those surprises:

1. The original results obtained by Gigerenzer and Goldstein (1996a) and summarized in chapter 4 generalized to 20 environments and to situations where the recognition heuristic played no role. This undermines the conjecture that there is something peculiar or wrong with the original domain of population sizes of German cities.
2. When one replaces the fitting task used by Gigerenzer and Goldstein with a generalization task, the fast and frugal Take The Best was even more accurate across 20 real-world environments than multiple regression. Take The Best achieved this accuracy despite using less than one third of all cues. Also, the myopic Minimalist came close to multiple regression in accuracy. Extending earlier findings, Dawes's rule slightly outperformed regression even when both were cross-validated.
3. Several variants of the competition did not change these results much. For instance, even with quantitative rather than binary predictors, Take The Best still matched and slightly outperformed the accuracy of multiple regression.

An important issue that we could not resolve in this chapter is the *how* question. How can one-reason decision making be as accurate, and sometimes even more accurate, than linear strategies—despite the latter's use of all cues and, in some cases, complex matrix computations? The result in figure 5-1 indicates that the relation between the number of cues and objects plays a role in the answer. But this does not explain the cause nor provide a proof. The next chapter will give several analytical answers and proofs. What structures of information in environments can Take The Best exploit, that is, what structures make a heuristic ecologically rational?

We began this chapter with a Darwinian message from Egon Brunswik: To understand the mind one needs to analyze the texture of its environment, past and present. Brunswik, however, also tentatively suggested that multiple regression could provide a model for how the mind infers its environment, and many neo-Brunswikians since have relied exclusively on these linear models. Our results indicate instead that mental strategies need not be like multiple regression to make accurate inferences about their environments. In the situations we studied in this chapter, simple heuristics can achieve the same goal. One-reason decision making can win a whole Tour.