Understanding and Improving Real-world Quantitative Estimation

Norman R. Brown & Robert S. Siegler

Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213

nbrown@psy.cmu.edu
siegler@psy.cmu.edu

Abstract

One possible method for improving real-world quantitative estimation is to "seed the knowledge-base" with explicit quantitative facts. This method was employed in two population estimation experiments. In Experiment 1, subjects estimated the populations of 99 countries. They then studied the populations of 24 of these countries. Finally, they estimated the populations of all 99 countries a second time. As predicted, the post-learning estimates for the 75 "transfer" countries were much more accurate (48%) than the pre-learning estimates. However, the rank-order correlations between estimated population and true populations showed almost no improvement. These results suggested that there may be two analytically distinct components to estimation, a range component and a ranking component, and that an arbitrary set of quantitative facts is likely to affect the former but not the latter. The aim of Experiment 2 was to demonstrate that one can affect the ranking component by presenting subjects with a consistent set of population facts. In this experiment, one group of subjects was presented with facts that consistently confirmed their prior belief that European countries are quite large and Asian countries are quite small. Another group was presented with a set that consistently disconfirmed this view. As predicted, rank-order correlations between estimated and true populations were negatively affected by the bias-confirming facts and positively affected by the bias disconfirming facts.

The world is composed of innumerable discrete entities; each of these entities has many properties; some of these properties can be specified in quantitative terms. It is common for a person to known that a certain entity exists, to recognize that this entity must have a certain quantitative property, and still to have no knowledge of the value of that property. For example, the typical U.S. college student is likely to have heard of Malaysia, to be absolutely certain that some number of people live there, and to be very uncertain as to what that number might be. Nonetheless, when pressed, this student will produce an estimate that is at least more accurate than a random guess and often will express some confidence in the rough accuracy of the estimate.

How do people generate *real-world quantitative estimates* of this sort? The cognitive literature provides two distinct perspectives on the matter. One focuses on domain-specific knowledge. According to this view, the estimation process is driven by a cycle of retrieval and inference. This process first recovers a fact from long-term memory that is related to the "target" entity (e.g., Malaysia) and/or to the "target" property (e.g., population). If the fact is relevant, it triggers an inference that narrows the response range in some way. For example, a person attempting to estimate the population of Malaysia might recall that there are about 120 million Japanese. If this person also believes that Malaysia has fewer people than Japan, he or she can confidently infer that the population of Malaysia is less than 120 million. Generally, a single fact will not allow the response range to be narrowed enough to yield a precise estimate. As a result, the cycle of retrieval and inference may continue until a precise estimate has been reached, until all immediately relevant knowledge has been exhausted, or until some time or effort limit is exceeded (Collins, 1978; Collins & Michalski, 1989).

The second perspective on real-world estimation holds that people depend on a small number of general purpose heuristics (e.g., availability, representativeness) to estimate the quantitative properties of objects and events (e.g., Brown, Rips, & Shevell, 1985; Lichtenstein, Slovic, Fischhoff, Layman, & Combs, 1978). According to this view, a person attempting to estimate the population of Malaysia might realize that he or she knows relatively little about this country. This would imply, via the application of the availability heuristic (Tversky & Kahneman, 1973, 1974), that the population of Malaysia is relatively small. Alternatively, a person attempting to estimate the population of Norway might depend on its representativeness. This person might decide that Norway is typical of Scandinavian countries and then infer that the population of Norway is likely to be similar to the populations of other typical Scandinavian countries, such as Sweden.

Interestingly, the same estimation task often produces results consistent with both the knowledge-based and heuristics perspectives. For example, a number of studies

This research was supported in parted by a National Institute of Mental Health postdoctoral fellowship to Norman R. Brown, and in part by A.W. Mellon Foundation grant awarded to Robert S. Siegler.

have demonstrated that people consider well-known public and autobiographical events to have happened more recently than less well-known events of the same objective age (e.g., Brown et al., 1985; Wagenaar, 1986). These results indicate that availability plays a role in date estimation. There is also clear evidence that domain-specific knowledge plays a central role in this task. For example, Brown (1990) reported an experiment in which subjects were asked to think aloud as they estimated dates for a set of well-known public events. Analysis of these protocols revealed that 78% of the estimates were justified with reference to one or more domain-specific fact.

The existence of evidence within a single task that is consistent with both perspectives suggests that any satisfactory approach to estimation must account for the influences of both general heuristics and domain-specific knowledge. One way to reconcile these two approaches is to assume a knowledge-based architecture and to view a given estimate as a weighted blend of available relevant information. Thus, when estimating the population of Malaysia, one might recall the population of Japan and use it to truncate the response at 120 million. One might also use availability to classify Malaysia as a not-very-large country. Taken together these two sources of information suggest not only that the target population is smaller than 120 million, but that it probably is a good deal smaller.

Viewing estimates as a weighted blend of relevant information provides a plausible mechanism for reconciling knowledge-based and heuristic perspectives. However, when estimates are conceived in this way, one is confronted with the issue of how the estimation process weighs competing sources of information. Drawing on recent work carried out within the cue validity framework (e.g., Gigerenzer, Kleinbolting, & Hoffrage, in press; MacWhinney, 1987), we hypothesized that competing facts are weighed according to (a) their *predictive strength* or *validity*, and (b) their *specific relevance* to the task. In the current context, predictive strength reflects the ability of a fact or inference to correctly predict the value of the to-be-estimated quantity. In other words, we assumed that people believe some inferences to be more credible or more reliable than others and that they weigh the more credible inferences more heavily than the less credible inferences. We also believed that inferences based on specific quantitative facts (e.g., the population of Japan is 120 million) would be weighed more heavily than those that were not. The two experiments described below bear directly on the accuracy on this particular instantiation of the specificity assumption.

Experiment 1

In an earlier experiment we found that (a) estimates of national populations tended to be very inaccurate in an absolute sense; the average population estimate was 3/4 of an order of magnitude away from the actual population. We also found that (b) better-known countries were considered to have larger populations than less well-known countries, other things being equal (Brown & Siegler, in preparation). We took the first finding to indicate that accurate knowledge of national populations is extremely uncommon, and the second to indicate that availability played a major role in this task.

The aim of the current experiment was to see if we could improve estimation performance and decrease the dependency on availability by introducing a set of relevant quantitative facts. In this experiment, subjects first estimated the populations of 99 countries. They then learned the actual populations of 24 of these countries. Finally, they were provided another set of estimates for all 99 countries. We expected that subjects would use the population facts they learned in two ways. First, these facts might serve as the basis for accurate generalizations about the geographical, social, economic and historical factors that result in populations of various sizes. Second, it seemed likely that the populations might be retrieved and used as quantitative reference points when subjects attempted to estimate the populations of the "transfer countries." These reference points would allow subjects either to truncate the range of possible responses or to select a numerically anchored region within the range that was likely to contain the target country's population.

Given these expectations, we predicted that "seeding the knowledge-base" would have the following effects on estimation performance. First, subjects should display a decreased dependence on availability. This follows from the specificity assumption made above; both the population facts and the newly derived demographic generalizations are more specifically relevant than availability judgments. Therefore, inferences based on this new information should be weighed more heavily than inferences based on availability. Second, we predicted that seeding the knowledge-based would lead to an improved rank-order correlation between true populations and estimated populations for the transfer countries. This would come about if subjects were able to use the seed set to correctly induce the factors that predict population size, or if they were able to select the most appropriate seed country or countries to serve as quantitative reference points. These same considerations led us to predict that mean estimates across the the transfer countries should be much more accurate in an absolute sense after the knowledge-base had been seeded.

Method

Twenty-four Carnegie Mellon undergraduates participated in this experiment. Each of these subjects performed four tasks during a 1 hr experimental session: a knowledge rating task, an initial estimation task, a learning task, and a final estimation task. In all but the learning task, subjects were exposed once to each of 99 countries. These 99 countries represented all but one of the countries that had populations of at least 4 million in 1989 (*Information Please Almanac*, 1989). The one exception was the United States, whose population was given to subjects as an example before the first estimation task.

During the knowledge rating task, subjects were presented with the names of the 99 test countries, one at a time, on a computer controlled video display. They were

instructed to evaluate their knowledge of each country on a 0-to-9 scale, with 0 indicating no knowledge of the country in question, 9 indicating a great deal of knowledge, and intermediate values indicating intermediate levels of knowledge. In this task, as in the others, subjects responded by typing their answers at the computer keyboard.

Following the knowledge rating task, subjects performed the initial estimation task. Again, the 99 test countries were presented, one at a time in a random order. Subjects were instructed to respond to each country with their best estimate of that country's current population.

Subjects were then presented with the four study-test blocks of the learning task. During each block, subjects were given the opportunity to study the actual population of each of 24 seed countries and were then tested on their knowledge of these populations. These countries were a subset of the full set of 99 countries. They were selected so that there were 6 countries in each cell of a 2 (country knowledge: High and Low) X 2 (estimation accuracy: High and Low) factorial design. Half of of the seed countries had received high knowledge ratings in a prior experiment and half received low ratings; half had received accurate estimates in the prior experiment and half had received inaccurate estimates. The high-knowledge, high-accuracy seed countries were: South Africa, Spain, Egypt, Italy, Great Britain, and West Germany. The the high-knowledge, low-accuracy seed countries were: Israel, Switzerland, Greece, Australia, Canada, and Vietnam. The low-knowledge, high-accuracy countries were: the Netherlands, Venezuela, Kenya, Romania, the Sudan, and Argentina. The low-knowledge, low-accuracy countries were: Bolivia, Zimbabwe, the Ivory Coast, Chile, Zaire, and Thailand.

The study-test blocks were divided into a study phase and a test phase. During the study phase, each seed country was presented with its population for 6 seconds. After subjects had studied all 24 study-test countries, they began the test phase. The task here was to respond to each country's name by typing its true population. When subjects could not recall a country's exact population, they were to respond with their closest approximation.

After completing the learning task, subjects began the second estimation task.The procedure followed during this task was identical to the one followed during the first estimation task.

Results and Discussion

The results provided clear support for only one of the three predictions made above. As predicted, "seeding the knowledge-base" resulted in a large decrease in absolute error across 75 transfer countries (i.e., the 75 countries that were not presented during the learning task). Specifically, for each subject, we first computed the absolute difference between the estimated population and the true population for each transfer country. Then, we computed the median absolute difference (MAD), for each subject, over all 75 transfer countries. Average MAD for the transfer countries decreased 48%, from 20.9 million in the first estimation task, to 10.9 million in the second ($t(23)= -2.45$, $p < .05$).

In contrast to the large decrease in MAD, the seeding procedure had little effect on the correlation between estimated and true population. For the 75 transfer countries, the average (taken over subjects) rank-correlation between these two measures was .40 before the learning task, and .43 after the learning task ($t(23)=1.57$, $p > 1$).

Finally, the data completely failed to support the prediction that seeding the knowledge-base would decrease reliance on availability. The results relevant to this prediction come from a pair of regression analyses, one on the estimates before the learning task and one on the estimates after the learning task. In both cases, the dependent measure was the median estimated population for each of the transfer countries. The predictor variables were mean knowledge rating, true population, and true land area. The medians were computed over subjects. The three largest countries, China, India, and the Soviet Union, were excluded from these analyses because they unduly influenced the outcome of the regressions. The R^2 computed for the pre-learning estimates was .66, as was the R^2 for the post-learning estimates.

If subjects had depended less on availability after the learning task than before, then the knowledge variable should play a larger role in the analysis conducted on the the pre-learning estimates than in the analysis conducted on the post-learning estimates. Contrary to this prediction, the knowledge variable played a smaller role in the former than in the latter. In the pre-learning analysis, country knowledge accounted for 35% of the unique variance; in the post-learning analysis, it accounted for 42% ($p < .001$. in both cases). Neither actual population nor actual land area accounted for more than 4% of the unique variance in either analysis.

These results suggest that there may be two analytically distinct components to real-world quantitative estimation. One component is an absolute or *range* component, and the other is a relative or *ranking* component. In the population estimation task, the absolute component can be equated with the assumptions concerning the plausible range of national populations, and the relative component with the knowledge used to locate a country's population within the assumed range.

The seeding procedure primarily affected the absolute component. It makes sense that studying the populations of 24 countries would lead subjects to develop a better understanding of the response range. There is support for this claim. The subjects who decreased the size of their estimates to the greatest extent across the two estimation tasks were those who provided the largest overestimates in the pre-learning task. Similarly, the subjects who increased their estimates the most were those who provided the most extreme underestimates in the pre-learning task (r=.98).

At the outset of this study, we expected that the seeding procedure would affect both the relative and the absolute components. However, the two failed predictions described above indicate that this procedure had little or no effect on the relative component. We believe that there are two reasons for this outcome. First, it appears that high task specificity is not necessarily equivalent to high predictive strength. That is, quantitative reference points will not

necessarily dominate performance when other credible source of information are available. Recalling the population of Japan may provide an upper bound for an estimate of Malaysia's population, but it may have no effect on a strongly held, availability-based belief that Malaysia has a fairly small population.

In addition to overestimating the potential usefulness of explicit reference points, we also seem to have overestimated the ability of subjects to act as "naive demographers." It appears that subjects were not able to use the population facts presented during the learning task to correctly update their beliefs about relative populations of different countries. In retrospect, this is not too surprising. Information provided by populations of the high-accuracy countries was consistent with subjects' prior beliefs, while the information provided by populations of the low-accuracy was inconsistent It seems unlikely that an ambiguous situation of this sort would compel subjects to reevaluate their beliefs.

Experiment 2

The aim of Experiment 2 was to demonstrate that seeding the knowledge-base with sets of population facts that are consistent and informative can influance relative population estimates. In the current context, a consistent set of facts is one in which all of the seed countries from an identifiable geographical region have similar populations (e.g., all of the European populations are small), and an informative set of facts is one in which the generalizations implicit in the set always confirm or always disconfirm prior beliefs about populations.

We knew from prior experiments (Brown & Siegler, in preparation) that subjects tended to believe that small European countries have relatively large populations and that large Asian countries have relatively small populations. In the current experiment, we attempted to influence these beliefs, and hence the ranking of European and Asian countries, by seeding the knowledge-base with sets of facts that consistently confirmed or consistently disconfirmed them. The expectation was that exposure to disconfirming population facts (i.e., facts about small European countries and large Asian countries) would improve the rank-order correlation between estimated and true population, and that exposure to confirming population facts (i.e., facts about large European countries and small Asian countries) would decrease this correlation.

Method

Three factors were varied: seed set, transfer region, and trial block. Seed set was a between subjects factor. It involvled the particular population facts that subjects learned. The 20 subjects in one group, the *bias-disconfirming* group, saw the populations of three small European countries (Switzerland, Sweden, the Netherlands) and three large Asian countries (Thailand, the Philippines, Vietnam). The 20 subjects in a second group, the *bias-confirming* group, saw the populations of three large European countries (Great Britain, Italy, West Germany) and three small Asian countries (Cambodia, Sri Lanka, Malaysia). Finally, 20 subjects were run in a control condition; these subjects were never exposed to the actual populations of any countries.

Trial block and transfer region were both within-subjects factors. Subjects in all three groups were required to estimate the populations of 36 countries once during each of 4 trial blocks. This set of countries consisted of the 12 seed countries and 24 transfer countries. Six transfer countries were selected from Asia (Burma, South Korea, Pakistan, Bangladesh, Japan, Indonesia), six from Europe (Norway, Denmark, Austria, Belgium, Greece, Portugal), six from Africa (Chad, Zimbabwe, the Ivory Coast, South Africa, Ethiopia, Nigeria) and six from Latin America (Honduras, Bolivia, Ecuador, Argentina, Mexico, Brazil). We refer to Asia and Europe as the *seeded* regions because both the confirming and disconfirming countries were drawn from them, and to Africa and Latin America as the *unseeded* regions, because subjects were not exposed to the populations of any of the countries from these regions.

The procedure followed on the first of the four trial blocks was identical to the procedure used during the Experiment 1 estimation tasks, except that subjects were only tested on the populations of 36 countries. As in the earlier experiment, subjects were instructed to enter a response that reflected their best estimate of each country's current population.

At the beginning of the second trial block, subjects in the bias-confirming and bias-disconfirming conditions. For the bias-disconfirming subjects, the computer display listed the names and populations of the smallest European seed country and the largest Asian seed country. Similarly, subjects in the bias-confirming condition saw the names and populations of the smallest Asian country and the largest European country. At the beginning of the third estimation block, two more country names and populations were added to the display. During this block, subjects in the experimental conditions were presented with information about the second largest of the small seed countries and the second smallest of the large seed countries. Finally, during the fourth block, the experimental subjects saw all six of the relevant seed set countries.

Subjects in the control condition simply provided four sets of estimates for the 36 stimulus countries and never saw the actual populations of any of these countries. Presentation order of the stimulus countries was randomized for each subject and each block.

Predictions

The central prediction of this experiment concerned possible changes in rank-order correlations across blocks. Specifically, we predicted a Seed Set X Transfer Region X Trial Block interaction. We expected that subjects who received the disconfirming seeds might realize that some European countries have very small populations and that some Asian countries have quite large populations. This understanding should lead subjects to decrease estimates for the European transfer countries and increase them for the

Asian transfer countries. This in turn should lead to an increased rank-order correlation between estimated and true population for the countries in the seed regions.

In the bias-confirming condition, we expected that the seed facts would increase subjects' confidence in their prior biased beliefs. To the extent that these beliefs are strengthen by exposure to the confirming seed set, subjects should increase population estimates for the European countries and decrease them for the Asian counties, leading to an overall decrease in the rank-order correlation between estimated and true population.

In brief, for countries in the seeded regions, rank-order correlation between estimated and true population should increase over blocks in the bias-disconfirming condition and decrease over block in the bias-confirming condition. In contrast, we predicted no change in rank-order correlation for the countries in the unseeded regions, in either the bias-confirming or bias-disconfirming conditions. This is because neither seed set had clear implications for the accuracy of the beliefs that determine the relative size of African and Latin American countries. Finally, since the control subjects were given no information about national populations, we expected no change in the rank-orderings of the countries in either the seeded or unseeded regions.

We expected a large decrease in MAD for the countries from the seeded regions and for countries from the unseeded regions, in both the bias-confirming and bias-disconfirming conditions. This is because the seed facts should provide information that would allow subjects to evaluate and update their range assumptions. Since range assumptions play an important role in determining the absolute accuracy of all responses (at least when range assumptions can be off by several orders magnitude), the improvement in MAD should not be restricted to the countries drawn from the seeded regions. The MAD scores for the control subjects should not change over blocks.

Results and Discussion

In order to test predictions concerning changes in the rank-ordering of countries, we first obtained eight rank-order correlations per subject. For each subject and each trial block, we computed one correlation between estimated and actual population for the 12 transfer countries drawn from the seeded regions, and one for the 12 transfer countries drawn from the unseeded regions. These correlations were then submitted to an ANOVA, which indicated that the interaction between seed set, transfer region and trial block was significant ($F(6,171)=5.21$, $p < .0001$).

As predicted, the rank-order correlation for the seeded regions increased when subjects were presented with bias-disconfirming seed populations and decreased when they were presented with bias-confirming seed populations (see Figure 1). Also as predicted, the rank-order correlations for the unseeded regions were unaffected by the presence of either confirming or disconfirming information. Finally, the control condition indicates subjects did not alter their ranking-ordering of countries in the absence of externally provided cues.

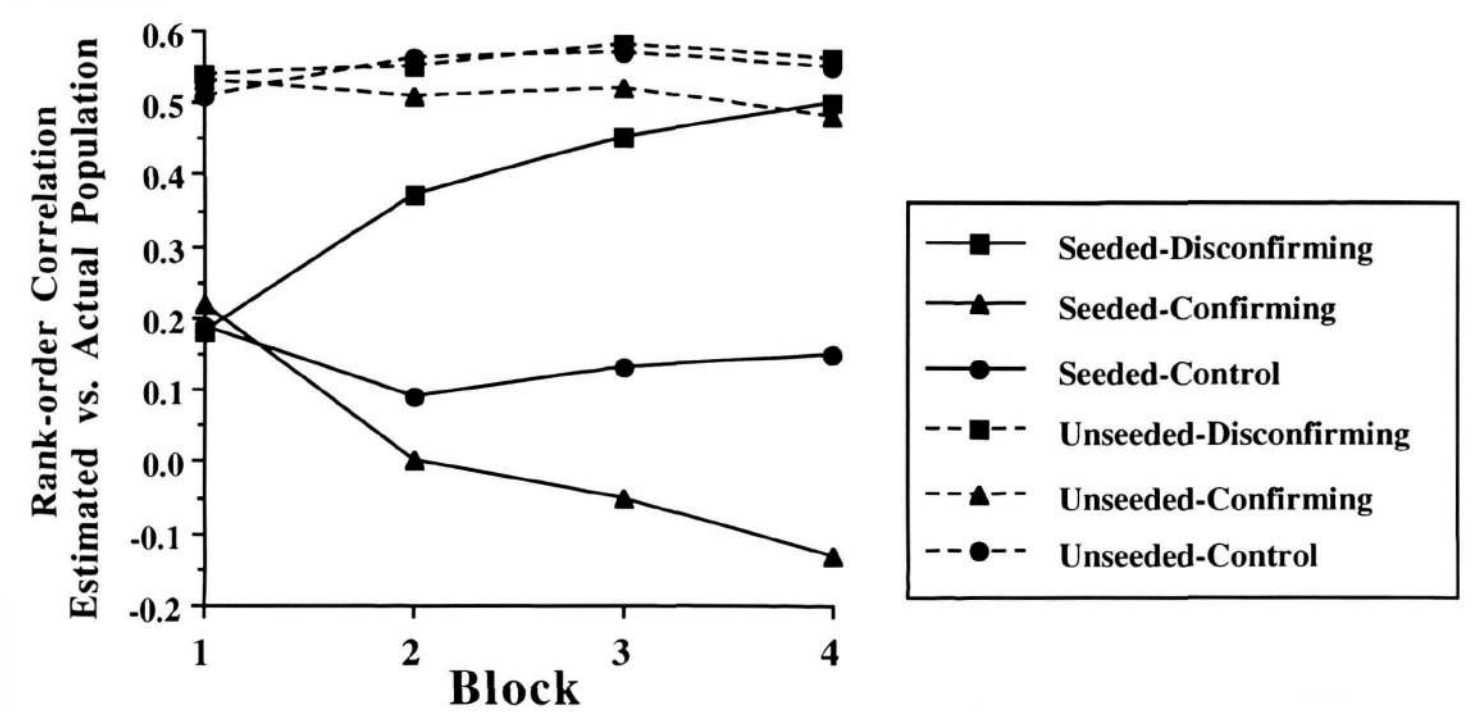

Figure 1. Mean rank-order correlations computed across countries in the seeded regions (Asia and Europe: solid lines) and the unseeded regions (Africa and Latin America: dashed lines), for subjects in the bias-disconfirming (squares), bias-confirming (triangles), and control conditions (circles). Experiment 2 data.

An ANOVA was also performed on the the accuracy data. In order to do this, we first computed the absolute difference between the estimated and actual population for each response. Next, for each subject and each block, we obtained medians of these absolute differences over the 12 seeded transfer countries and over the 12 unseeded transfer countries.

The most interesting result in this ANOVA was a significant Seed Set X Trial Block interaction ($F(6,171)=3.27$, $p < .01$). Across the four trial blocks, estimates in the control condition became slightly less accurate (MAD = 27.9 million in Block 1, MAD = 34.5 million in Block 4); estimates in the bias-confirming condition became somewhat more accurate (MAD = 35.6 million in Block 1, MAD = 24.0 million in Block 4); and estimates in the bias-disconfirming condition improved a great deal. (MAD = 34.5 million in Block 1, MAD = 18.5 million in Block 4). In contrast to the analysis performed on the rank-order correlations, the Seed Set X Transfer Region X Trial Block interaction was not signficant for the MAD measure ($F(6,171)=1.19$, $p > .1$). This is consitent with the prediction that the seeding proceedure would improve preformance in an absolute sense for both the seed and unseeded regions.

Conclusions

There are three main points to take away from the research just described. First the process that generates real-world quantitative estimates often blends qualitative and quantitative information. It appears that this process does not necessarily grant a special status to information that is explicitly quantitative. Second, estimation can be seen as having two distinct components: an absolute or range component and a relative or ranking component. When the knowledge-base is seeded with an arbitrary or neutral set of quantitative facts, one can expect to improve performance in an absolute sense but not in a relative sense. Finally, it appears that seed facts must be carefully selected if one hopes to improve performance in both relative and absolute senses. Specifically, the set of facts presented to subjects must be consistent, informative, and valid. A consistent set of facts is one that allows for an obvious mapping between the items' qualitative and quantitative category; an informative set of facts is one that provides subjects with evidence that suggests changes in their beliefs; and a valid set of facts is one that leads to an increase of the predictive validity of the modified beliefs.

References

Brown, N. R. (1990). The organization of public events in long-term memory. *Journal of Experimental Psychology: General, 119,* 297-314.

Brown, N. R., & Siegler, R. S. (in preparation). Quantitative estimation in knowledge-rich domains.

Brown, N. R., Rips, L. J., & Shevell, S. K. (1985). The subjective dates of natural events in very long-term emory. *Cognitive Psychology,* 17, 139-177.

Collins, A. M. (1978). Fragments of a theory of human plausible reasoning. In D. Waltz (Ed.), *Theoretical issues in natural language processing--2.* (pp. 194-201). Urbana, IL: University of Ilinois.

Collins, A. M., & Michalski, R. (1989). The logic of plausible reasoning: A core theory. *Cognitive Science, 13,* 1-49.

Gigerenzer, G., Kleinbolting, H., & Hoffrage, U. (in press). Confidence in one's knowledge: A Brunswikean view. *Psychological Review.*

Information please almanac. (1989). Boston: Houghton Mifflin.

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events.*Journal of Experimental Psychology: Human Learning and Memory, 4,* 551-578.

MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 249-308). Hillsdale, NJ: Erlbaum.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5,* 207-232.

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

Wagenaar, W. A. (1986). My memory: A study of autobiographical memory over six years. *Cognitive Psychology, 18,* 225-252.