Strategies for Revising Judgment:

How, and How Well, Do People Use Others' Opinions?

Jack B. Soll
INSEAD

Richard P. Larrick
Fuqua School of Business, Duke University

January 2004

Preliminary Draft
Please do not quote or cite without permission of the authors

A basic tenet of social psychology is that people look to others to resolve uncertainty in their judgments (Festinger, 1954). Research in social psychology has painted two pictures of the consequences of this process. The dominant image has been one of people yielding too readily to the influence of others (Asch, 1952). This image comes from the large body of work on normative influence in which judgments were given publicly and the judgments of others were deliberately designed to be inaccurate. But a second image—frequently proposed but less explored—is that judgments can become more accurate under the influence of others (Deutsch & Gerard, 1955). Improvements in judgmental accuracy, however, received little theoretical or empirical attention in the classic investigations (see reviews by Allen, 1965; Tajfel, 1969). In this research, we address three questions: What strategies do people use when revising their own judgments in light of the judgments of others? Under what conditions are different strategies effective? Finally, do people adopt the most effective strategies?

Research on groups illustrates the potential benefits of using the judgments of others. Starting in the 1920s, a number of studies in psychology demonstrated that an average of uncertain quantity estimates from multiple individuals tends to be more accurate than most of the individuals whose estimates were being averaged (for a review of the early literature, see Lorge, Fox, Davitz, & Brenner, 1958). This result was initially viewed as surprising, and researchers debated its cause. Eventually, the success of averaging was appropriately explained not as an emergent property of groups but as a statistical necessity. Today, combining uncertain estimates to reduce error is used so routinely in academics and in practice that its application is rarely questioned (Borsboom, Mellenbergh, & van Heerden, 2003). Thus it is surprising that the idea of averaging to improve accuracy, as with many statistical insights, was the result of a cultural development that occurred relatively recently across a number of scientific fields. Prior to the insight, scientists faced a “combining opinions” dilemma: Should different, conflicting judgments be averaged? Or should one judgment be chosen as the “correct” answer? As late as the nineteenth century, it was not uncommon for astronomers and chemists to choose one or several individual measurements that they believed to be of the best quality, and discard the rest (Gigerenzer et al., 1989; Stigler, 1986).

An interesting case study is that of the great mathematician Leonhard Euler, who failed to solve a problem in astronomy because he was unwilling to average observations. Euler entertained averaging as a possibility, but rejected it because he feared that “the errors of the observations and of the calculations can multiply themselves.” (Euler, 1749, as quoted by Stigler, 1986). Euler was dealing with observations from many different sources, reported over centuries under unknown circumstances. In a laboratory study, Soll (1999) presented participants with an analogous situation, in which each information source had an unknown bias and was more or less susceptible to measurement error. Many participants endorsed the view that measurement errors will add up when averaging across sources. The basic statistical result that averaging cancels out errors is apparently not intuitively obvious.

The present article focuses on a particular form of combining judgments across sources. We are interested in how people update their beliefs in light of advice from another person. In our experiments, participants report a set of initial estimates, then receive advice in the form of estimates from another participant. Models based on information integration theory (Anderson, 1971; Birnbaum, 1976) or anchoring (Tversky & Kahneman, 1974; Hogarth & Einhorn, 1992) would suggest that people arrive at an intermediate solution, somewhere between the two initial guesses, perhaps closer to their own than to the advisor’s. In contrast, if people are no more sophisticated than pre-modern mathematicians and scientists (and why should they be?), we would expect them to often choose from among estimates, even when averaging leads to greater accuracy.

The last point deserves some elaboration. Averaging does not always lead to better results than choosing, but it often does. We develop a model that delineates the environmental conditions under which averaging or choosing is more accurate. A key contribution of the model is the observation that averaging is very robust, performing well even when there is a large difference in the skills of the two judges. The model also highlights the possibility that a single strategy is not well suited to all environments. If people always take the same approach, whether it be averaging or choosing, they will perform well in some circumstances and poorly in others.

The interplay between judgmental strategy and environment has received substantial attention in judgment and decision research. Simplifying heuristics such as the lexicographic choice rule or equal weighing of information work remarkably well in some environments but not in others (Payne, Bettman, & Johnson, 1993; Einhorn & Hogarth, 1975). Additionally, there is some evidence that people tend to apply effort-saving heuristics when it is most advantageous to do so (Payne, Bettman, & Johnson, 1988). Along the same lines, Gerd Gigerenzer and his colleagues (e.g., Gigerenzer & Goldstein, 1996) have proposed that people use “fast and frugal algorithms” in inference problems. Such algorithms are quick to implement and require few processing operations. One such algorithm that has been widely studied is “take-the-best”, which is remarkable because it gives an answer based on only one piece of information. Consider a series of questions of the type “Which city is more populous, Cleveland or El Paso”. Take-the-best entails starting with the most valid cue to which one has access (e.g., whether a city has a major league baseball team). If the cue discriminates, the city with the cue is chosen. Otherwise, the algorithm proceeds to the next most valid cue. It has been shown that for comparative judgment tasks with binary cues, take-the-best outperforms an equal weighting of cues, and performs as well as regression in some circumstances. The success of take-the-best depends on a match with the environment, performing relatively well when cues are noncompensatory and when the judge learns which cues are best (Hogarth & Karelaia, 2003).

In the combining opinions problem, the initial quantity estimates can be thought of as cues that the judge combines. The judge may either choose one estimate (take-the-best) or average them (equal weighting). Other weighting schemes are possible, but rarely provide a significant advantage (Dawes, 1979; Einhorn & Hogarth, 1975). It is difficult to say which strategy is faster or more frugal. Choosing requires that the judge invest processing power in deciding who is most expert. In contrast, averaging requires that the judge identify an estimate that is roughly in the middle. Although literally computing the average might be difficult, it turns out that moderate deviations from averaging have minimal effect on performance, a point that we return to later. Thus, we construe the averaging strategy loosely as reporting a value somewhere in the middle. We would argue that finding an intermediate value is not too difficult for most people, and so choosing and averaging are both relatively fast and frugal. In comparative judgment with binary cues, take-the-best outperforms equal weighting across a wide range of environments (Hogarth & Karelaia, 2003). The task we have in mind is different, because the criterion is a continuous quantitative variable, and the cues – initial estimates from oneself and an advisor – are continuous and on the same scale. For this type of problem, a combination of estimates is often more accurate than just one (Hogarth, 1978).

In the next section, we provide a brief discussion of the psychological literature on combining quantity estimates. We then describe why averaging is such a robust strategy, and offer a model that describes the conditions under which averaging or choosing will perform better. We then present results from four empirical studies that vary environmental factors relevant to the performance of different strategies. In the first three of the studies, averaging outperforms intuitively revised estimates, while in the last study intuition performs slightly better. As we will show, intuitive revision tends to resemble choosing, even when environmental conditions favor averaging. Thus,we can use our model to predict the conditions under which intuitive revision exceeds or falls short of averaging. In the general discussion, we will argue that environments in which averaging works well are common. A consequence is that people would do well to average more often.

Literature on Revising Quantity Estimates

While the history of science suggests that people are choosers, much work in cognitive psychology has concluded that when faced with multiple pieces of information, people do some form of averaging. Birnbaum and his colleagues (Birnbaum & Stegner, 1979; Birnbaum, Wong, & Wong, 1976) applied information integration theory to the problem of combining expert opinions that vary in credibility, and found that the data were best described by a relative weighted averaging model. Budescu (2004) found that people tend to average the opinions of multiple experts. Recently, theorists have considered how people revise their own quantitative beliefs. Several recent studies have supported the view that people anchor on their initial estimate and then adjust toward the advice. In a typical task, participants make a series of initial estimates, learn the estimates of an anonymous "advisor", and then provide final estimates. Several researchers have reported that final estimates on average move 30% of the way toward the advice (Harvey & Fischer, 1997; Lim & O'Conner, 1995; Yaniv & Kleinberger, 2000; Yaniv, 2003). Participants in Harvey and Fischer's cue-learning studies adjusted at least 20% even when the advisor had fewer learning trials than the participant. The authors postulated that this “token” adjustment reflects a social norm that free advice should not be ignored. Yaniv (2003; Yaniv & Kleinberger, 2000) has attributed the discounting of advice to the fact that people have privileged access to their own reasons for judgment but not to other’s reasons. He further stresses that participants “…resolved the discrepancy between their own and the other opinion by adhering to their own opinion and making a token shift to the other opinion.” (Yaniv, in press, p. 5).

The evidence in favor of weighted averaging or anchoring appears overwhelming. However, we would draw attention to a particular aspect of the analysis that allows for the possibility that people, like Euler, are often choosers. Essentially, the issue is this: Most studies that have concluded that people are averaging use averages to arrive at this conclusion. Suppose that mean adjustment in a study is reported to be 20%. This can happen in many ways. One possibility is that a judge consistently adjusts 20%, perhaps with some normally distributed random error. Alternatively, a judge could alternate between 0% and 100% adjustment at a ratio of 4:1. The latter judge is choosing, but if only the mean is analyzed it is easy to draw the conclusion that the judge is making a token adjustment. Averaging and choosing have vastly different implications for accuracy, as we shall soon see.

Few researchers have reported data analysis that is fine grained enough to show whether people are truly averaging on individual questions. There are, however, some hints that people may be choosing more than has been suspected. Yaniv and Kleinberger (2000, study 1) reported a mean adjustment of 29%, but a median adjustment of 2%. The low median implies that participants hardly budged from their initial estimates on at least half the items. Similarly, Yaniv (2003, study 1) reported that on 58% of items, adjustments were between 0 and 30%, inclusive.

It is possible that most of the adjustments were 0, in which case the "token shift" would be a very rare event.

Additional evidence against averaging comes from an attempt to validate the weighted-averaging model advocated by Birnbaum. Lees and Triggs (1997) asked participants to predict a letter of the alphabet based on one or two probabilistic cues of varying validity, also letters of the alphabet. The cues were unbiased but noisy. For example, when the criterion was the letter "M", the distribution for each cue was alphabetically symmetric around "M", but with varying amounts of dispersion. Learning trials were provided to give participants a sense of the relative validity of the cues. The pattern of means supported weighted averaging, and in fact looked very similar to results reported by Birnbaum (1976) for a similar task. However, an item-by-item analysis yielded a very different conclusion. Histograms of responses revealed a bimodal distribution, with judgments in between the cue values on only 36% of the trials. The aggregate results completely missed this switching behavior, thus giving a misleading impression of the psychological process.

In summary, while many studies have reported results consistent with averaging, few have conducted the fine-grained analysis necessary to confirm averaging at the item level. When an item level analysis is carried out, the data often reveal strategies that go beyond taking a weighted average of the available information. We do not claim that people never average – in many cases they probably do. However, it is also the case that when responses are aggregated over many items, alternative strategies such as choosing do not have a chance of being detected.

The Benefits of Averaging

The effectiveness of averaging has been demonstrated in fields as diverse as psychiatry (e.g., Goldberg, 1965, 1970), meteorology (Staël Von Holstein, 1971), and economics (Clemen, 1986; Graham, 1996). Two major reviews in the forecasting literature have concluded that averaging performs well compared to other statistical methods of aggregation (Armstrong, 2001; Clemen, 1989). This result applies not only to combining intuitive judgments, but also to combining the forecasts of different statistical models (Makridakis & Hibon, 2000; Makridakis & Winkler, 1983). In a compelling demonstration, Armstrong (2001) reanalyzed thirty studies conducted between 1960 and 2000, which varied substantially in topic, forecasting methods, time horizon, number of forecasts combined, and method for evaluating accuracy (e.g., mean absolute deviation, root mean squared error, Brier score, etc.). Across the studies, averaging led to a reduction in error from 3.4% to 23.5% relative to the mean accuracy of the individual components, with a mean reduction of 12.5%.

Why does averaging work so well? Consider what happens with estimates from two judges. There are two possible configurations for the estimates relative to the truth. One possibility is that both estimates fall on the same side of the truth. Suppose that two faculty are estimating the number of ph.d. applications they will receive this year. Their estimates are 70 and 80, and the correct answer is 90. The judges miss the truth by 20 and 10, respectively, so their mean discrepancy is 15. Averaging gives an answer of 75, which also misses by 15. In general, when estimates are all on the same side of the truth, the discrepancy between the average estimate and the truth equals the mean discrepancy of the individual estimates. Now consider estimates that *bracket* the truth, such as 70 and 100. The mean individual miss is still 15, but the average estimate of 85 performs much better, missing the truth by just 5. In general, when estimates bracket, the discrepancy between the average and the truth must be less than the mean discrepancy of the individual estimates. At a minimum, averaging performs at the mean performance level of the individual estimates, and in the case of bracketing can perform much better. While we demonstrated this using absolute deviation as the measure of performance, the result holds for any convex penalty function (Winkler, 1971; Hogarth, 1978), such as squared error, and for correlation (Dawes, 1970). In a straightforward extension, the result also holds when accuracy scores are aggregated across items (Larrick & Soll, 2003). That is, the mean absolute deviation (MAD) of averaging is at least as low as the average MAD of the judges.

To get a sense of the degree to which averaging might help, consider a simple example where the forecast errors of two judges follow a standard bivariate normal distribution with zero correlation, which implies that the judges' estimates bracket the truth 50% of the time. Figure 1 shows MAD as a function of the weight on one's own judgment (which we will term weight-on-self and abbreviate *WS*). MAD is about 0.8 if either judge is used alone (*WS* = 0 or 1), and improves as *WS* approaches 0.5. Averaging leads to a MAD of 0.56, or a 30% improvement over either judge alone. Figure 1 yields several interesting insights. First, even small adjustments away from extreme weighting lead to substantial improvement. Just moving 10% from one judge's estimate toward the other yields approximately one-third of the total benefits of averaging. Second, MAD is relatively insensitive to moderate deviations from the optimal weight. A *WS* of 0.6 or even 0.7, applied consistently, is nearly as good as 0.5 (see von Winterfeldt & Edwards, 1986). Finally, choosing will be ineffective in this situation. Alternating between judges results in a mean weight of 0.5, but yields none of the benefits of averaging.

Although previous research has shown that the mean *WS* is approximately 0.7, the above analysis suggests that this self-bias is not too serious if people apply this weight consistently. In the problem above, a consistent weight of 0.7 leads to 24% improvement over using either judge alone. Some inconsistency is also not a major problem: If the mean weight is normally distributed around 0.7 with σ = 0.2, the improvement drops to 21%. Of course, if a mean weight of 0.7 is achieved by alternating between weights of 1 and 0 (the choosing strategy), then accuracy will not improve.

Averaging is the optimal strategy in the above example, but moderate deviations from averaging perform nearly as well. The reverse is also true; averaging often performs well when it is not optimal (Dawes & Corrigan, 1974; Einhorn & Hogarth, 1975). There are two basic reasons for this. First, due to the flat optimum result, averaging will come close to optimal performance as long as the optimal weights are not too extreme. Second, while averaging by definition will underperform optimal weights (except, of course, when averaging is optimal), it can outperform *estimated* optimal weights, even when those weights are derived using formal statistical models (Dawes, 1979; Einhorn & Hogarth, 1975). For example, Winkler (1984) found that averaging the forecasts of two analytical models outperformed regression weights estimated from past performance. Averaging will not always outperform regression weights, but it will typically be close, and under well-defined conditions can do better on out-of-sample forecasts (Einhorn & Hogarth, 1975; Camerer, 1981). For these reasons, many decision theorists agree that, prescriptively, averaging is a sound strategy that provides a viable alternative to intuitive aggregation (Armstrong, 2001; Clemen, 1989; Ferrell, 1985; Goodwin & Wright, 1998).

Aside from bracketing, two additional factors figure in the relative success of averaging. First, if the judges differ substantially in expertise (Libby, Trotman, & Zimmer, 1987), then the optimal weights will be relatively extreme, and averaging might not perform as well as choosing or other intuitive strategies. Second, if intuitive weights deviate sufficiently from optimal weights, then averaging might outperform intuition even when optimal weights are extreme. For example, suppose that (a) the optimal weights are 0.8 on the more accurate judge and 0.2 on the less accurate judge; (b) the judge assigns weights of 0.8 and 0.2; and (c) the judge detects the more accurate judge with probability *p*. At what *p* does averaging beat the noisily applied optimal weight of .8? The precise answer depends on the joint distribution of errors. For purposes of illustration, assume that forecast errors follow a bivariate normal distribution, with means and correlation equal to zero. In this case, the intuitive strategy outperforms averaging if $p \geq 0.72$. Several studies of expert identification suggest that people are only somewhat better than chance at selecting the most accurate member of a group (Henry, Strickland, Yorges, & Ladd, 1996; Yetton & Bottger, 1982). If it turns out that the expert detection rate is typically less than 0.7, averaging may perform well even when there are large differences in expertise.

In summary, averaging always performs at least as well as the average judge (assuming a convex penalty function such as *AD*), and is likely to outperform intuitive combination strategies in many situations. At least two existing studies provide evidence that people adopt a choosing strategy at least some of the time (Lees & Triggs, 1997; Yaniv & Kleinberger, 2000). Building from the existing literature on group judgment and linear models, we have identified three variables that determine the relative success of averaging. These are the bracketing rate, the difference in accuracy, and the probability of identifying the better judge. The variables are compensatory. For example, a high bracketing rate can offset a large difference in accuracy, such that averaging will still be the better strategy. In the next section, we provide a quantitative model using these variables to describe the conditions under which averaging or choosing performs better.

Model

This section develops a methodology for comparing the averaging and choosing strategies when combining the estimates of two judges. While other theorists have emphasized the importance of the expert detection probability (Einhorn, Hogarth, & Klempner, 1977) and variation in expertise (Libby et al., 1987), our model is the first to combine these variables, along with the bracketing rate, into a unified theoretical framework. The baseline strategies provide benchmarks against which to compare intuitive revision. Because people know how to average (or at least to compromise), we might claim that people are revising suboptimally if they consistently perform worse than averaging. In addition, in some cases the baseline strategies might also provide descriptive models of behavior. For example, Lees and Triggs' (1997) data are suggestive of a choosing strategy.

Before introducing the model, we wish to clarify several points. First, averaging and choosing are generic strategies that can be applied to multiple contexts, such as group judgment (Einhorn et al., 1977; Hastie, 1986) and the panel of experts problem (Fischer, 1981; Hogarth, 1978; Libby & Blashfield, 1978). Although our model is general and can be directly applied to these other contexts, we develop it for an advice taking setting, where a judge has an initial estimate, and then revises that estimate after learning the estimate of a second judge. Second, averaging and choosing are pure strategies that, in practice, are probably not applied exclusively or perfectly. Rather, a person might apply a baseline strategy noisily, or use a hybrid of the two strategies, neither of which we try to capture in the model. Finally, a judge who stays with his or her initial estimate has by default identified the self as the expert, and in this sense is applying the choosing strategy. That is, we interpret a failure to change one's mind as a reflection of the belief that one knows better than others.

We conceptualize the judgment of Judge *j* on question *i* as the sum of the true answer T_i, a judge-specific bias B_j, and a question-level error term:

$$X_{ij} = T_i + B_j + e_{ij}, \tag{1}$$

where $i = 1, ..., n$ and $j = 1, 2$, for which n is the total number of questions in the set. The bias B_j is defined as the difference between the judge's mean estimate and the mean correct answer among the questions in the set. Bias represents the tendency for the judge to over- or underestimate the correct answers, and is easiest to interpret when the questions all come from the same domain of knowledge (e.g., a judge who tends to overestimate age by five years). Finally, the error term e_{ij} includes all other sources of deviation from the truth, including randomness in the environment, randomness in the response, and systematic idiosyncrasies in how the judge evaluates a specific question. We assume that e_{ij} is normally distributed with mean zero and variance σ_j^2.

Judge *j*'s accuracy on question *i* is defined as the absolute difference from the truth, $AD = |X_{ij} - T_i|$. From Equation 1, this can also be expressed as the absolute sum of bias and error, $|B_j + e_{ij}|$. Einhorn et al. (1977, Equation 5) derived the formula for *MAD* (i.e., the mean or expected *AD*) for the case of normally distributed errors. Applying their formula using our framework and notation, we obtain

$$MAD = \sigma\left[\frac{B}{\sigma}\left(2F\left(\frac{B}{\sigma}\right)-1\right)+2f\left(\frac{B}{\sigma}\right)\right], \qquad (2)$$

where *F* and *f* are the *cdf* and *pdf* for the standard normal distribution, respectively. Notice in Equation 2 that accurate judgment requires that both bias *and* error are small. A judge who has small errors but large bias may be highly correlated with the truth, but still inaccurate by our measure of accuracy. When bias is zero Equation 3 reduces to the well-known formula for mean absolute deviation from the mean for the Normal distribution, $MAD = \sigma\sqrt{2/\pi}$.

Equation 2 can be used to compute the MAD achieved by averaging. The bias for averaging, B_{avg}, is simply the average bias of the individuals. The variance of the error term is given by $\sigma_{avg}^2 = .25\sigma_1^2 + .25\sigma_2^2 + .5\sigma_1\sigma_2\rho$, where ρ is the correlation between e_1 and e_2. Plugging B_{avg} and σ_{avg} into Equation 3 gives MAD_{avg}. The *MAD* acheived by choosing will equal the MAD of one of the two judges, depending on who is chosen. Without loss of generality, let $MAD_1 \le MAD_2$ for judges 1 and 2. The expected accuracy from choosing is $E\left(MAD_{chs}\right) = pMAD_1 + (1-p)MAD_2$, where *p* is the probability that Judge 1 (the more accurate judge) is identified as the expert.

MAD ratio, Bracketing, and Expert Detection

Three parameters determine the conditions under which averaging or choosing is more accurate. The first parameter, *R*, measures the relative expertise of the two judges, and is defined as the ratio of the individual *MAD*s, highest over lowest. Assuming that Judge 1 is more accurate, *R* = 1.5 implies that Judge 2 is 50% less accurate than Judge 1. The second parameter is the bracketing rate *Br*, defined as the percentage of questions for which the true answer lies between the estimates of Judge 1 and Judge 2. More bracketing implies greater benefits from averaging. The third model parameter is the expert detection probability *p*.

The three parameter model provides a straightforward way to describe the conditions under which averaging or choosing performs better. In the absence of bias the model precisely specifies when averaging will outperform choosing and vice versa. With bias, there is a range of *R* and *Br* combinations for which the model cannot predict perfectly which strategy is more accurate; nevertheless, its performance in this range is still highly accurate. We use different methods to analyze the no-bias and bias cases, so they are considered separately below.

The No-Bias Case

In order to have as general a treatment as possible, we first define *k* as the factor required to equate averaging and choosing in terms of accuracy. Therefore,

$$MAD_{avg} = kE\left(MAD_{chs}\right)$$

$$MAD_{avg} = k\left(pMAD_1 + (1-p)MAD_2\right). \tag{3}$$

The right-hand side represents the expected performance of choosing, multiplied by the factor k. Averaging outperforms choosing when $k < 1$. In the absence of bias,

$$MAD_i = \sqrt{2/\pi}\sigma_i \text{ , and}$$

$$MAD_{avg} = \sqrt{2/\pi}\sigma_{avg} = \sqrt{\left(2/\pi\right)\left(.25\sigma_1^2 + .25\sigma_2^2 + .5\sigma_1\sigma_2\rho\right)}\,.$$

Substituting these identities into Equation 3 and rearranging terms gives:

$$\rho = \frac{4k^2\left(p + R - Rp\right)^2 - R^2 - 1}{2R}\,. \tag{4}$$

Equation 4 specifies the condition that must hold for averaging to perform at a factor k relative to choosing. For example, if $k = 1$, then averaging and choosing are equally accurate, and Equation 5 specifies the relationships among ρ, R, and p that must hold for that to be true. Ultimately, we are interested in the trade-offs among the bracketing rate, R, and p. Equation 4 will be useful as an input in our next set of equations.

We now turn to the bracketing rate, which is defined as the proportion of questions for which the true answer lies between the two estimates:

$$Br = P\left(X_1 > T, X_2 < T\right) + P\left(X_1 < T, X_2 > T\right)$$

$$= P(B_1 + e_1 > 0, B_2 + e_2 < 0) + P(B_1 + e_1 < 0, B_2 + e_2 > 0).$$

As the sum of a normal random variable and a constant, the term $B_j + e_j$ is normally distributed with mean B_j and variance σ_j^2. The bracketing rate can be expressed as

$$Br = 1 - \left[F_\rho\left(\frac{-B_1}{\sigma_1}, \frac{-B_2}{\sigma_2}\right) + F_\rho\left(\frac{B_1}{\sigma_1}, \frac{B_2}{\sigma_2}\right)\right], \tag{5}$$

where again F is the *cdf* of the standard bivariate normal with correlation ρ. In the no-bias case, Equation 5 simplifies to

$$Br = 1 - 2F_{\rho}\left(0,0\right). \qquad (6)$$

In the absence of bias, the bracketing rate depends only on the correlation in errors. The equations are used to plot the relative performance of averaging and choosing as shown in Figure 2. The first step is to fix k and p. In the figure, expert detection is perfect (p = 1), and curves for several values of k are plotted. For each value of R, we apply Equation 4 to derive the correlation that must hold when R, k, and p are fixed. Equation 6 gives the bracketing rate implied by the derived correlation. The curves in Figure 2 are obtained by iteratively applying this process. The curves plot the (R, Br) combinations for which the accuracies of averaging and choosing have a constant ratio. For example, the curve k = 0.8 plots those points for which averaging is 20% more accurate than choosing. The curve k = 1.2 plots those points for which averaging is 20% less accurate than choosing.

A special curve is the one for which k = 1; This *iso-accuracy* curve plots the points for which averaging and choosing are equally accurate. Averaging outperforms choosing at points above the curve, and underperforms choosing at points below it. For example, if R = 1.2, then a bracketing rate of 0.275 is sufficient for averaging to be as accurate as choosing. If R = 1.5, then a bracketing rate of 0.420 is sufficient.

Figure 3 plots the iso-accuracy curve for different levels of p. Not surprisingly, the region for which averaging outperforms choosing grows larger as the probability of detecting the expert declines. For example, if one has a 70% chance of identifying the better judge (p = 0.7), averaging would be the better strategy even if R is as high as 3, provided a high level of bracketing (40-50%).

The Bias Case

The simplicity of the no-bias model is very appealing. The *MAD* ratio, the bracketing rate, and the probability of detecting the expert can be estimated from data. Relative to choosing, averaging is more accurate the lower the *MAD* ratio, the greater the bracketing rate, and the greater the probability of detecting the expert. The iso-accuracy curve makes a clear prediction about whether averaging or choosing will work better in a specific situation. The model explains not only why averaging often succeeds, but also why it sometimes fails. In addition, the model is potentially useful in applied settings. Practitioners may find it easier to understand and estimate the bracketing rate as opposed to standard statistical measures such as correlation and standard deviation.

Unfortunately, when judges are biased it is no longer possible to derive a unique iso-accuracy curve for a given combination of *R* and *Br*. One could plot different iso-accuracy curves for different sets of underlying parameters. The precise location of the iso-accuracy curve will depend on the correlation in errors, the bias-to-sigma ratio of each judge, and the ratio of the sigmas. Such a model would make accurate predictions, but it is too complicated to yield substantial insight. For example, increasing the bias-to-sigma ratio of one judge may help or hinder averaging, depending on the precise values of the other parameters. Moreover, in practice it may be very difficult to estimate bias-to-sigma ratios. Our approach, therefore, is to use the results from the no-bias case as an approximation to what happens when there is bias. Of course, the model will occasionally predict averaging is more accurate than choosing when it is not and vice versa. However, we will show that such errors are infrequent, and tend to be relatively minor when they do occur. It turns out that, under a wide range of realistic parameter values, the no-bias model provides excellent predictions even when bias is present. The remainder of this section quantifies the performance of the no-bias model as an approximation to the bias case.

To test the approximation, we systematically varied the values of the parameters for two judges. The parameters were varied as follows: *bias*, -4 to 4 for one judge, 0 to 4 for the other judge, both in increments of 0.2 (due to the symmetry of the bivariate normal, it is unnecessary for both biases to range from –4 to 4); s*igma*: 1 to 4 in increments of 0.2 for both judges; and *correlation*, -0.7 to 0.7 in increments of 0.1. Altogether, the resulting dataset has over three million observations. Using the formulas presented earlier, the *MAD* ratio and bracketing rate were computed for each combination of underlying parameters, as were the expected MADs for averaging and choosing.

Consider first what happens in the absence of bias. We selected 300 observations at random from the dataset, with the condition that both biases were zero (see Figure 4, Panel A). As shown in Figure 4 (Panel A), the iso-accuracy curve perfectly separates the +'s (averaging is more accurate) and o's (choosing more accurate). We then randomly selected an additional 300 observations, with the condition that the absolute bias-to-sigma ratios were less than or equal to 1.5. This cutoff was chosen because in our experiments we rarely observe higher ratios. Here the no-bias iso-accuracy curve no longer predicts perfectly; there are several *intrusions* of +'s and o's to the wrong side of the curve (see Panel B). However, it is clear that the curve is highly accurate. The intrusion rate is 1.94% for the entire restricted dataset (|bias/sigma| ≤ 1.5), and 3.00% for the unrestricted dataset.

To get a sense of the extent of the intrusion problem, we plotted all intrusions for the restricted dataset for *p* = 1 and *p* = 0.7, and then traced the outer contour of the intrusions to produce lower and upper bounds to the iso-accuracy curve (see Figure 5). These bounds form a *zone of indeterminacy* within which either averaging or choosing may be more accurate. However, even within this narrow zone, the model still makes accurate, if not perfect, predictions when $p = 1$: a hit rate of 87% when the model favors choosing (between the iso-accuracy curve and the lower bound), and 84% when it favors averaging (between the curve and the upper bound). The zone of indeterminacy is similar for both levels of probability (Figure 5).

Thus far we have shown that intrusions are infrequent and that they occur within a well-defined, narrow range of MAD ratio and bracketing combinations. It is also interesting to know the magnitude of the errors, which we now analyze for $p = 1$. The value k indicates, for each observation, the accuracy of averaging relative to choosing. For a given combination of *MAD* ratio and bracketing, it is possible to estimate k using the no-bias assumption. We call this estimate k_{nb}. The correlation between k and k_{nb} over the entire restricted dataset is 0.994. The median and 90th percentile of the distribution of the absolute difference between k and k_{nb} are 0.011 and 0.063, respectively. The corresponding percentile values within the zone of indeterminacy are 0.010 and 0.048. Finally, to address the seriousness of the mistakes we looked specifically at the intrusions. When the model incorrectly predicts in favor of choosing, the median absolute difference between k and k_{nb} is 0.019. When the model incorrectly predicts in favor of averaging, the median absolute difference is 0.064. These results show that the no-bias assumption yields very accurate estimates of the relative performance of averaging and choosing in cases of bias.

Summary

In the case of no bias, the iso-accuracy curve perfectly predicts when averaging will be more accurate than choosing. In the presence of bias, we simplify the model by approximating the iso-accuracy curve with the curve derived from the no-bias case. This approximated curve is still highly accurate. There is a narrow zone of MAD ratio and bracketing rate combinations for which the approximated curve does not predict perfectly. However, even within this zone the approximation is highly accurate, in terms of both frequency and magnitude of mistaken predictions.

Data

Past research has shown that people tend to weight their own opinions more highly than those of others. However, the literature has not produced an adequate description of item-by-item weighting strategies. Mean *WS* (weight on self) across a set of items is useful as a global measure of the extent to which an advisor's judgment is taken into account, but cannot be used to measure the extremity of weights. For that, we introduce a new statistic. On a given question, *WX* (weight on perceived expert) is defined as the greater of *WS* and 1 - *WS*. Naturally, *WX* is constrained to the interval .5 – 1. Consider a person who switches between weights of 0 and 1. While this person's mean *WS* equals the proportion of 1s, mean *WX* must be 1.

We conducted four studies to describe item-by-item intuitive weighting strategies, and compare their accuracy to simple baselines, such as averaging. In each study, participants estimated a series of quantities, and then revised their guesses based on the estimates of another participant. The studies included manipulations intended to alter perceptions of relative expertise. Because people attend to credibility (Birnbaum, 19xx), we expected that mean *WS* would vary substantially across conditions. Just because mean *WS* is shifting, however, does not imply that mean *WX* will shift as well. In a pure choosing model, changes in perceived credibility would merely change the number of occasions that WS equals zero or one. This would affect mean *WS* but not mean *WX*. In contrast, we would assert that a unimodal distribution of *WS* would be implied by an anchoring or weighted-averaging model, with the mode somewhere between 0 and 1. In such a model, changes in perceived credibility should shift the mode, making weights more or less extreme, and thus having an effect on mean *WX*. In short, if manipulations of perceived expertise affect mean *WS* but not mean *WX*, that would provide strong support for choosing.

While we expect people to choose rather than average, we also suspect that simple revision strategies will tend to outperform intuition. To test this, we compared intuitive revision with four baseline strategies: Staying with initial estimates, averaging, perfect choosing (i.e., setting the expert detection probability equal to one), and consistent *WS*. For this last baseline, we compute the mean *WS* for a given judge, and then apply that weight consistently to all questions in the set. While consistent weighting can improve judgment, in many environments equal-weighting does even better (Dawes, 1979; Einhorn & Hogarth, 1975). We anticipate that averaging will often outperform consistent *WS*, but that whether averaging outperforms choosing will depend on the characteristics of the environment, as specified by the model.

The studies are sufficiently similar that they can be presented together. In all four studies, participants made a set of initial quantitative estimates, copied their answers onto an "advice sheet," received the advice sheet of another participant in the study, and gave final estimates on which they were paid for accuracy. Each study included a manipulation intended to affect perceptions of relative accuracy. If people are choosers, the manipulation should affect mean *WS* but not mean *WX*. In addition, the studies were designed to ensure some variation in the environmental parameters described in the model. We first present the methods of all four studies, and then proceed to the results.

Method

Study 1 (Salaries-feedback) : As part of a classroom exercise, 76 University of Chicago masters of business administration (MBA) students predicted starting salaries for alumni of 25 American MBA programs three years after graduation. The schools each were randomly generated from a list of top MBA programs appearing in the *Financial Times*. Participants first made initial estimates and copied them onto an "advice sheet," which was then randomly and anonymously distributed to another participant in the classroom. Some participants also received performance feedback for own and advisor's estimates on a 5-item pre-test that everyone took one week earlier. Participants received a base payment of $0.60 for each question, less $0.02 for every $1,000 by which they missed the correct answer. Negative payments were allowed on individual questions, but not for the study as a whole. Total payments ranged from $5 to $10.75, with a median of $8.70.

Study 2 (Salaries-familiarity): Fifty-three University of Chicago MBA students performed a similar task as in Study 1 as part of a classroom exercise. Twenty-five US business schools were randomly selected from the *Financial Times* list. All participants rated their familiarity on each school on a 1 (never heard of it) to 7 (extremely familiar) scale and estimated salary. In one group of 32 participants, familiarity ratings were passed along with the advice, whereas in the other group they were not. The payment scheme was the same as in Study 1, with payments ranging from $3 to $10, and a median payment of $7.40.

Study 3 (Country facts): Participants were mid-career executives, most in their mid to late 30s, participating in an executive MBA program and representing 24 nationalities. They were paired such that each dyad member was from a different country. For each country, they estimated (1) the percentage of the population living in an urban area, (2) the percentage of the population under age 15, (3) the percentage of adult males that smoke, (4) the percentage of married women of childbearing age using a modern method of birth control, and (5) the percentage of members of parliament or congress who are women. Correct answers were obtained from the web sites of the World Health Organization and the United Nations. Participants recorded estimates first for their own country and then for their partner's. They then exchanged all ten answers without discussing them. Finally, for each question participants assessed the probability that their own initial estimate was closer to the truth than their partner's, and also recorded a final private estimate. The base payment for each question was $1, and 5 cents were deducted for each percentage point by which the final estimate deviated from the correct answer, up to a maximum deduction of $1. Payments ranged from $2.70 to $7, with a median of $5.

Study 4 (Five topics): Seventy University of Chicago students, mostly undergraduates, answered questions on five topics: Air distances between U.S. cities, dates of events in U.S. history, ages of recent Oscar nominees, popularity ranks of pop singles, and number of wins of NBA basketball teams. For each topic, twenty questions were randomly drawn from a standard list[1]. Two versions of a booklet were created, 10 questions per page, each page devoted to a different topic. There was no overlap in questions between the two versions. While recording their initial estimates for the 50 items, participants also reported their confidence for each item on a scale from 1 (not at all confident) to 7 (extremely confident). After making all 50 initial estimates, they assessed expertise on each topic on a scale from 1 (completely unknowledgeable) to 7 (very knowledgeable), and then copied all 50 answers onto an advice sheet, which went to

[1] Description of sources goes here.

another participant in the study. Some participants (the Exchange group) also recorded the confidence and expertise assessments onto the advice sheet as well. In the final segment of the procedure, participants received the advice sheet from the same person who received their advice – thus participants were arranged in dyads. Dyads in the Exchange group met with each other for 5 minutes to exchange information about why they might or might not be expert about different topics or questions. Finally, all participants recorded final answers, knowing that the bonus payment for each question was 30 cents, less 2 cents per unit deviation from the correct answer, except for air distances, in which case it was less 2 cents per unit of percentage error. This was necessary because the distances between cities varied substantially across questions. Negative payments were allowed for individual questions, but the minimum payment for each topic was zero. All participants received $5 just for showing up. Bonus payments, which were added to the $5 base, ranged from $0.90 to $9.80, with a median of $6.40.

Results

Distribution of weights. The overall distribution of *WS* was W-shaped in each study, as shown in Figure 6. There are several aspects of the histograms that stand out. In each study participants completely ignored advice (*WS* = 1) a fair amount of the time, approaching 40% in all but one study. Fully accepting advice (*WS* = 0) was much less common, but still represents a mode in each distribution. Combining cases of complete ignorance and complete acceptance, the rate of pure choosing (*WX* = 1) ranged from 35% in the country facts study to 56% in the five topics study. There is also a mode at simple averaging (weights in the .4-.6 category), which people do between 13% and 26% of the time across studies.

Together, the weight categories *WS* = 0, *WS* = 1, and *WS* = .4-.6 account for the majority of judgments in each study: 64% in salaries-feedback, 66% in salaries-familiarity, 69% in five-topics, and 61% in country facts. Token adjustments from one's initial opinion (*WS* = .8-.99) were especially rare, with a maximum of 11% of weights in the salaries-feedback study. Thus, on a scale that can be conceived of as having 101 points, judges used just 22 points on more than 60% of the judgments. These data strongly support the idea that people primarily use averaging and choosing as their basic revision strategies. As discussed in the introduction, applying the better of these two strategies often outperforms attempts to optimize weights. People appear to have the right portfolio of simple strategies, although it remains to be seen whether they apply the right strategy at the right time.

How means of weights vary with cues to expertise. The means of *WS* and *WX* are displayed in Table 1. In the salaries-feedback study we computed means separately for participants who learned that they were more or less accurate than their partner in the pre-test, in addition to those who received no feedback. In the salaries-familiarity study, we computed the mean familiarity rating for each participant. For those who exchanged ratings, we performed a median-split based on the difference in familiarity between participant and advisor. This produced one group that was on average more familiar than advisors, and another group that was of similar familiarity or less familiar. In the country facts study we computed means separately for own and partner's country. Lastly, in the five-topics study we categorized each participant as having low confidence on a given topic if their mean confidence rating for that topic was less than 3.0, and as having high confidence otherwise. Using this cutoff roughly split the dataset in half, allowing for the four pairings of own and advisor's confidence displayed in Table 1.

Varying perceptions of expertise had a greater impact on $\overline{\mathrm{WS}}$ than $\overline{\mathrm{WX}}$. In the salaries-feedback study, one-way analysis of variance (ANOVA) revealed that the three feedback groups differed on $\overline{\mathrm{WS}}$, $F(1, 73) = 10.97$, $p < .001$, but not on $\overline{\mathrm{WX}}$, $F < 1$. In the salaries-familiarity study, there were again differences in $\overline{\mathrm{WS}}$, *F*(2, 50) = 3.82, p = .03, but not $\overline{\mathrm{WX}}$, *F* < 1.1. In the country facts study, $\overline{\mathrm{WS}}$ was significantly greater for own country (paired *t* = 11.36, p < .001), but this time there was a difference in $\overline{\mathrm{WX}}$ as well (paired *t* = 3.05, p = .003).

Finally, in the five topics study we analyzed each condition as a 2 (high or low own confidence on a topic) x 2 (high or low advisor's confidence) factorial, using either $\overline{\mathrm{WS}}$ or $\overline{\mathrm{WX}}$ as the dependent variable in separate ANOVAs.[2] When participants exchanged confidence ratings, the analysis of $\overline{\mathrm{WS}}$ revealed a main effect of own confidence, *F*(1, 176) = 10.43, *p* = .001, and advisor's confidence, *F*(1, 176) = 25.75, *p* < .001, but no interaction (*F* < 1). In contrast, confidence did not affect $\overline{\mathrm{WX}}$, all Fs < 1. When participants did not exchange confidence, there was only a main effect of own confidence on $\overline{\mathrm{WS}}$, *F*(1, 164) = 6.31, *p* = .01, but no effect of advisor's confidence, *F* = 2.49(1, 164), *p* = .12. Again, there was no interaction (*F* < 1). $\overline{\mathrm{WX}}$ was not sensitive to either confidence variable, all *F*s < 1.1. Because participants contributed at most five observations to each mean in this study, we note that dividing the degrees of freedom by five leads to identical conclusions at $\alpha = .05$.

As we had anticipated, $\overline{\mathrm{WS}}$ is much lower than $\overline{\mathrm{WX}}$. Looking only at $\overline{\mathrm{WS}}$, as was done in past research, it appears that people are often close to achieving the full benefit of averaging, because of the flat optimum result discussed earlier. The fact that $\overline{\mathrm{WX}}$ is typically above .8 discounts that possibility. It would of course be possible to achieve $\overline{\mathrm{WX}}$ = .8 by consistently

[2] Since there were five topics, there were five times as many observations as participants. However, in the no exchange condition, one participant did not assess confidence on one topic, reducing the number of observations by two (one for the participant, and one for the advisee).

adjusting a token 20% toward advice, a mechanism first proposed by Harvey & Fischer (1997). The distribution of weights in our data is inconsistent with this explanation.

Moreover, our analyses created groups of participants that perceived themselves to be more or less expert than their advisor. The subgroups varied greatly on $\overline{\mathrm{WS}}$, but not on $\overline{\mathrm{WX}}$. In the salaries-feedback study, for example, $\overline{\mathrm{WS}}$ ranges from .54-.77, compared to just .82-.84 for $\overline{\mathrm{WX}}$. This difference in range is too large to be explained by the change in scale. Additional insight into the stability of $\overline{\mathrm{WX}}$ can be gained by examining the percentage of intuitive revisions that reflect averaging (WX ≤ .6) or choosing (WX = 1). This data is shown in the last two columns of Table 1. In Study 1, the relative frequency of choosing and averaging is similar across the groups, and so $\overline{\mathrm{WX}}$ is highly stable. In Studies 2-4, there is some variability in the choosing rate, leading to some modest differences in $\overline{\mathrm{WX}}$. People are more likely to average in some conditions than in others.

When do people average? Thus far, we have seen that people often employ extreme weights, but occasionally they average. There are several reasons why people might average more in one situation rather than another. One reason is that people might perceive cues to expertise to be more diagnostic when they favor the self rather than the advisor. In Study 3, for instance, an Estonian might have believed that it is very important to be from Estonia to answer questions about Estonia correctly, while a person not from Estonia might not think so. Alternatively, holding beliefs about diagnosticity constant, people might respond to those beliefs differently depending on whom the cues favor. This might be true, for instance, if people found it aversive to completely ignore their initial opinions.

The results of Studies 2 and 3, as reported in Table 1, are consistent with either interpretation. The results of Study 4 support the conjecture that people are sensitive to diagnosticity. $\overline{WX}$ was smaller in the no exchange condition, where participants did not meet with advisors to discuss expertise and did not receive advisors' confidence ratings. Without this extra information, participants may have been less sure about who was expert, and thus more inclined to average. However, the Study 4 results do not address the question of whether people are biased in how they interpret or respond to cues to expertise.

Additional relevant data comes from the country facts study. Recall that before giving their final estimates, for each question participants assigned the probability that their estimate was closer than their partner's. If averaging is a function of perceived expertise, then participants should have averaged most when the assigned probability was near .5. As shown in Table 2, we found exactly this relationship, again demonstrating that participants were sensitive to the perceived diagnosticity of the available information.

Did participants treat the country cue differently depending on whether it was their country? This question is difficult to answer with the existing data, because we do not know what other private information is being combined with the country cue to arrive at a probability. Nevertheless, the table does show that participants tended to use more extreme probabilities on their own country than on the partner's. Did participants respond to beliefs about relative expertise differently depending on who was favored by the country cue? Here the answer is apparently yes. For instance, when participants assigned a probability in the range .61-.99 that they were closer (mean of .808 for own country and .776 for partner's), they ignored one of the judges 49% of the time for their own country, compared to 28.3% of the time for the partner's country. Apparently, 80% probability has different implications for opinion revision depending on the country.

It is also interesting to note that even when the probability judgment was near .5, participants still chose about a quarter of the time. If we constrain the analysis to cases where the assessed probability was exactly .5, $WX = 1$ on 27.3% of questions. The fact that people still sometimes choose when they are unsure about who is more expert suggests that choosing has strong intuitive appeal.

People use extreme weights less, and average more frequently, when they are less sure about who is more expert. However, there is also evidence that other considerations come into play in opinion revision, aside from beliefs about expertise. People may feel compelled to place at least some weight on their own opinion, and they may feel obliged to place some weight on an advisor who has some claim to expertise (e.g., an Estonian answering questions about Estonia), even if they feel that they are more expert. Teasing apart these influences on weight is important, but beyond the scope of this article. Rather, we emphasize here that people often use extreme weights, and that people are more likely to average when they are unsure about who is more expert.

Accuracy. We compared the performance of the initial estimates, intuitively revised estimates, and the three revision strategies across the three studies. Each row in Table 3 gives results for a different set of questions. The accuracy scores given in the last five columns are all MADs, with the exception of city distances. For that topic, participants were paid based on mean absolute percentage error, so that is what we report here. In each row, accuracy scores that do not share a subscript letter differ significantly from each other. The comparison between averaging and intuitively revised estimates was planned and tested at $\alpha = .05$. The remaining nine pairwise comparisons were analyzed with paired t-tests, setting $\alpha = .006$ (.05/9), as prescribed by the Bonferroni procedure.

All four revision strategies provided a substantial advantage over the initial estimates. Which one was best? That depends on the environment. In Studies 1 and 2, averaging led to more accurate salary estimates than any of the other strategies, including both intuitive revision and perfect choosing. For both studies, The MAD ratio is relatively low at 1.3 and the bracketing rate is around 40%. The model (Figure 2) predicts that averaging will outperform perfect choosing in this case. In contrast, in Study 3 averaging continues to beat most strategies, including intuitive revision, but it does not beat perfect choosing. In terms of the model, this is because the difference in expertise is much greater in Study 3 – the MAD ratio is 1.8. Given a bracketing rate near 40%, the model predicts that perfect choosing would perform better in this case. The success of perfect choosing in Study 3 suggests that participants could have potentially outperformed averaging, if only they had a better cue to relative expertise. In the absence of such a cue, perfect choosing represents an unattainable ideal.

Finally, in Study 4, MAD ratios were high, and averaging was statistically indistinguishable from intuitive revision on all topics but city distances, where it performed significantly worse. Given the high MAD ratios, perfect choosing performed best as it did in Study 3. This time, intuitive revision managed to come closer to this ideal, directionally surpassing the performance of averaging in three of the five topics.

Study 3 and Study 4 have similar parameter values and, in both cases, perfect choosing surpassed averaging as would be expected. Surprisingly, however, intuitive revision performed worse than averaging in Study 3 and better than averaging in Study 4, despite the similar parameter values. The answer to this inconsistency is that the expert detection rate, which appears similar—and modest—at the aggregate level in the two studies, varies in a subtle way. For some dyads in Study 4, the MAD ratios were very high, exceeding two. For these participants, averaging fared especially poorly when compared with simply adopting the estimates of the better judge, assuming that the better judge can be identified. It turns out that for MAD ratios of at least two, the expert detection rate is 83.3%. The rate drops to 55.3% for MAD ratios less than two. Of 350 observations (5 topics for each of 70 participants), the MAD ratio exceeded two on 78 of them (21.7%). Looking only at these 78 cases intuitive revision outperformed averaging on every topic. Conversely, averaging performed better on every topic on the remaining observations. The near-tie between averaging and intuitive revision in Study 4 is driven by a small number of observations in which there is a large difference in expertise. Participants often correctly detected the better judge in these cases, avoiding the large errors that would have resulted from averaging. At the same time, the remaining majority of participants missed out on the gains from averaging.

There are two additional patterns in Table 3 that deserve more elaboration. First, consider the fact that revised estimates and perfect choosing perform very similarly in the two salary studies. This seems surprising given the low levels of expert detection. If participants in Study 1 always adopted the judgments of the better judge, they would have achieved a MAD of 12.9. In fact, they identified the better judge at about chance level, and still managed a MAD of 13.1. This result partly reflects the fact that participants shift between choosing much of the time and averaging some of the time (a pattern that should not be interpreted as making token adjustments). They likely achieved gains in accuracy when they did average, in line with the lower MAD score for averaging. If, for example, participants in Study 1 chose randomly on half the trials and averaged on the other half, they would have achieved a MAD of about 13.6 (averaging the 15.0 and 12.2 from the first line of Table 3), which is not too different from the actual score for revised estimates.

Another noteworthy result in Table 3 is the performance of the consistent *WS* strategy. One can think of this strategy as an attempt to optimize the weights—a strategy that would produce a unimodal pattern of WS that would hold if people were making token adjustments. It is analogous to behavioral bootstrapping, a technique in which a judge's estimates are regressed on a set of cues, and the resulting regression equation is used to make predictions. Typically, bootstrapping outperforms intuitive judgment on out-of-sample predictions, because it removes the random error component of human judgment (cite). We have conceptually replicated this effect, because consistent *WS* beats the initial estimates in each row of Table 3. Another classic result is that equal weighting of cues, after standardizing to equate for differences in scale, tends to outperform bootstrapping (Dawes; Einhorn & Hogarth). This holds in our data as well. Averaging is statistically superior to consistent WS in studies 1 and 2, and directionally superior in all but the city distances topic of Study 4.

It is important to note that although averaging works in the aggregate, individual judges may be harmed when their estimates are averaged with an inferior advisor. This point is addressed in Table 4. To facilitate comparisons across studies, the MAD scores were standardized for each topic using the mean and variance of the MADs of the initial estimates on that topic. Lower scores indicate greater accuracy. As one might expect, averaging performs extremely well for the less accurate judge, beating initial estimates and intuitive revision in every study. For the more accurate judge, averaging outperformed both initial estimates and intuitive revision in Studies 1 and 2. In Study 3, averaging again outperformed intuitive revision, but actually scored slightly worse than the initial estimates. In this study, MAD ratios were relatively high, so the performance of the better judge tended to suffer slightly from revision. Of course, participants were not much better than chance at detecting the better judge, so by averaging participants could have locked in a solid performance. Finally, Study 4 is the only one in which averaging underperformed the intuitive revision of the more accurate judge. Note, though, that the differences in Study 4 are not great. Overall, across all studies averaging tends to perform about as well as intuitive revision for the more accurate judge, and in most cases performs much better for the less accurate judge. Given that people cannot perfectly predict which judge is more accurate, averaging appears to be the more robust strategy, as it performs well whether or not one is more accurate than one's advisor.

General Discussion

Much like early scientists, people often choose among estimates rather than average them. When people do average, it appears to be because they are unsure which source is more expert. In four empirical studies of quantitative estimates with a single advisor, the distribution of weights was W-shaped; people often retain their initial estimate, occasionally average, and occasionally fully adopt the advisor's estimate. This contradicts past research that concluded that people make minor adjustments toward advice. Participants in our studies adjusted about 30% toward advice on average, but this is an aggregate result that arises from a combination of choosing and occasionally averaging. We also found that participants would have been more accurate had they averaged more often. While people appear to favor a choosing strategy, based on our model the conditions in our studies often favored averaging. Even when the difference in expertise might have justified choosing, the rate at which participants identified the better judge was insufficiently high. The only exception was the five-topics study, and even here averaging approximately matched the performance of intuitive revision.

Strategies for Revising

It is interesting that people have settled on choosing and averaging to revise opinions, because these are exactly the two weighting policies that have proved most accurate in many empirical studies. Under certain well-defined conditions, it pays to 'take-the-best,' that is, use a single piece of information to the exclusion of all else (Gigerenzer & Goldstein, 1996). In many other circumstances it is better to use multiple cues. Furthermore, unless optimal weights can be estimated with sufficient precision, equal-weighting of cues will lead to greater accuracy (Dawes, 1979; Einhorn & Hogarth, 1975). While people have the right two strategies, they do not employ them in proper proportions. People appear to choose far too often. Why is that? The answer, we believe, is that people do not know why averaging is such a good strategy. People appear to average as a last resort, when they are unconfident about who is more expert. Some evidence for this conclusion comes from the country facts study. Participants tended to average when they thought the chances were roughly equal that their own estimate was closer than the advisor's. The more sure they were that one judge was better, the more likely they were to choose rather than average. Additional evidence comes from a series of experiments by Larrick and Soll (2003). Participants were asked to guess what happens to accuracy when estimates are averaged. For example, one study described two yen-to-dollar currency forecasters whose historical MADs were 4.7 and 5.3. The joint distribution of over- and underestimating the truth was provided in a 2 x 2, from which the bracketing rate can easily be deduced. When the bracketing rate was 24%, about three quarters of participants guessed that the MAD from averaging would be 5.0 – the same as the average level of performance. When the bracketing rate was increased to 90%, still half of the participants thought that averaging would lead to average performance. Two additional studies in Larrick and Soll (2003) showed people are more likely to appreciate averaging when bracketing is frequent or highly salient. For example, most people understand that averaging works when judges have opposing biases. However, when no special attention is drawn to bracketing, most people assume that averaging will merely lock in average results. A reasonable interpretation of the results of the present article, as well as Larrick and Soll (2003), is that when people average it is because they are risk averse. According to this explanation, people would prefer to accept an average level of performance rather than risk guessing wrong and getting the performance of the worse judge. If this is true, then when people average they are using a good aggregation strategy but for the wrong reason.

Given that people often choose when there is one advisor, what would they do when there is more than one? Consider the case of two advisors, so there are three estimates altogether. If people consistently choose the middle estimate, they would probably do very well, as this strategy will approximately mimic averaging. On the other hand, people will not do as well if they consistently choose a favorite judge. There is some evidence to suggest that in combining the opinions of others, people exclude outliers and weight the most confident judges most heavily (Yaniv, 1997). When combining the opinions of equally capable experts people appear to average (Budescu, 2004). More research is needed to determine whether people attempt to guess which judge is best and ignore the others. Future work should also investigate the extent to which it matters that the combiner of opinions also has an initial opinion, as in the present article.

Is Averaging Good Advice?

Based on the model, we expect people to do well when there is a large difference in expertise and people can detect it. The bracketing rate also figures into the equation, but based on our empirical studies bracketing is typically in the range of 30-40%. Moreover, an important result from the five-topics study is that a judge’s chances of detecting the expert are increasing with the MAD ratio. Because participants often chose the best judge when there was a large difference in expertise, averaging would have made them substantially worse off. Does this result imply that a general prescription to average is misguided?

To address this question, consider a combiner who must combine the quantitative opinions of two judges (it does not matter for now whether the combiner is one of the judges). Suppose that the combiner is very confident that one of the judges is more accurate than the other. Based on this information alone, it is not possible to say whether the combiner should average, as that depends on beliefs about the MAD ratio and to a lesser extent bracketing. What if the combiner believes that one analyst is 50% less accurate than the other (a MAD ratio of 1.5) and estimates bracketing to be 30%? According to Figure 3, the probability of correctly identifying the better judge would have to be about .75 to justify choosing in this situation. Based on present results, this level of expert detection seems unlikely given the difference in expertise. But suppose that the combiner believes the MAD ratio to be 2.5. If the combiner is right, averaging would lead to substantially lower accuracy. Since people are good at detecting the expert when there is such a large difference in accuracy, wouldn't it be a mistake to average in this case? The answer to this depends partly on the validity of cues to expertise possessed by the combiner, and partly on the dispersion of accuracy in the pool from which the judges were sampled. This is an interesting modeling problem outside the scope of the present article. We would point out, though, that if the dispersion of accuracy in the pool is small, then when the combiner perceives a ratio of 2.5, the actual ratio is likely to be much smaller.

We would argue that many real-world environments are such that the dispersion of accuracy is small, MAD ratios are less than two, expert detection is weak, and hence averaging is a good strategy. Consider the following two cases. First, imagine that you are unknowledgeable on a topic. Our recommendation would be to consult with at least two people who are knowledgeable and to average their opinions. Given that these people were selected for their expertise from a larger pool, it is unlikely that one is at least twice as inaccurate as another. Second, if you are knowledgeable our advice would be to consult with at least one other person and average. Again, since the advisor is not chosen at random but selected for expertise, it is unlikely that the MAD ratio is very large, and averaging is the best strategy. Of course, when more than two judges are identified are in the selected pool, averaging over all of them can yield substantial gains in accuracy (Hogarth, 1978), far exceeding the performance of the single judge identified as best.

Suppose there is only one advisor. Does it make sense to average? Based on our results, expert detection may not be far above chance unless the MAD ratio is at least two. Moreover, averaging can do very well when the MAD ratio is below two, especially if the bracketing rate is in the 30-50% range and the probability of detecting the expert is not perfect. For lower MAD ratios, even when averaging does not perform as well as choosing, it will come close to it. Thus, we would argue that a MAD ratio of two is a reasonable cut-off for strategy selection. We might then ask the combiner to estimate the MAD ratio – "Do you believe that, on average, one judge is twice as far from the truth as the other?" We think that people would rarely predict such a large difference in accuracy, in which case our advice is to average. Choosing may or may not be appropriate when people estimate the MAD ratio to be greater than two. Although our results indicate that people can accurately detect high MAD ratios, that does not imply that the MAD ratio is high when people say it is. While we believe that there are situations in which people should choose, more research is necessary to understand the conditions under which predictions about relative expertise are likely to be accurate.

References

Allen, V. L. (1965). Situational factors in conformity. In L. Berkowitz (Ed.), Advances in Experimental Social Psychology (Vol. 2, pp. 133-175).

Anderson, N. H. (1971). Integration theory and attitude change. Psychological Review, 78(3), 171-206.

Armstrong, J. S. (2001). Evaluating forecasting methods. In J. S. Armstrong (Ed.), Principles of Forecasting: A Handbook for Researchers and Practitioners. New York: Kluwer.

Asch, S. E. (1952). Social Psychology. Oxford, England: Prentice-Hall.

Birnbaum, M. H. (1976). Intuitive numerical prediction. American Journal of Psychology, 89(3), 417-429.

Birnbaum, M. H., & Stegner, S. E. (1979). Source credibility in social judgment: Bias, expertise, and the judge's point of view. Journal of Personality and Social Psychology, 37, 48-74.

Birnbaum, M. H., Wong, R., & Wong, L. K. (1976). Combining information from sources that vary in credibility. Memory & Cognition, 4(3), 330-336.

Camerer, C. (1981). General conditions for the success of bootstrapping models. Organizational Behavior and Human Performance, 27, 411-422.

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. International Journal of Forecasting, 5, 559-609.

Dawes, R., & Corrigan, B. (1974). Linear models in decision making. Psychological Bulletin, 81(2), 95-106.

Dawes, R. M. (1970). An inequality concerning correlation of composites vs. composites of correlations. Oregon Research Institute Methodological Note, 1(1).

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. American Psychologist, 34(7), 571-582.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. Journal of Abnormal & Social Psychology, 51, 629-636.

Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighing schemes for decision making. Organizational Behavior and Human Performance, 13, 171-192.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. Psychological Bulletin, 84(1), 158-172.

Festinger, L. (1954). A theory of social comparison processes. Human Relations, 7, 117-140.

Fischer, G. W. (1981). When oracles fail---a comparison of four procedures for aggregating subjective probability forecasts. Organizational Behavior and Human Performance, 28, 96-110.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. Psychological Review, 103(4), 650-669.

Gigerenzer, G., Switjtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). The Empire of Chance: How Probability Changed Science and Everyday Life. Cambridge, England: Cambridge University Press.

Goldberg, L. R. (1965). Diagnosticians versus diagnostic signs: The diagnosis of psychosis vs. neurosis from MMPI. Psychological Monographs, 79.

Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inference. Psychological Bulletin, 73, 422-432.

Goodwin, P., & Wright, G. (1998). Decision Analysis for Management Judgment. Chichester, England: John Wiley & Sons.

Graham, J. R. (1996). Is a group of economists better than one? Than none? Journal of Business, 69(2), 193-232.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. Organizational Behavior and Human Decision Processes, 70(2), 117-133.

Hastie, R. (1986). Review essay: Experimental evidence on group accuracy. In B. Grofman & G. Guillermo (Eds.), Information pooling and group decision making (Vol. 2, pp. 129-157). Greenwich, CT: JAI Press.

Henry, R. A., Strickland, O. J., Yorges, S. L., & Ladd, D. (1996). Helping groups determine their most accurate member: The role of outcome feedback. Journal of Applied Social Psychology, 26, 1153-1170.

Hogarth, R. M. (1978). A note on aggregating opinions. Organizational Behavior and Human Performance, 21, 40-46.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. Cognitive Psychology, 24, 1-55.

Larrick, R. P., & Soll, J. B. (2003). Intuitions about combining opinions: Misappreciation of the averaging principle.

Lees, C. D., & Triggs, T. J. (1997). Intuitive prediction: Response strategies underlying cue weights in the relative-weight averaging model. American Journal of Psychology, 110(3), 317-356.

Libby, R., & Blashfield, R. K. (1978). Performance of a composite as a function of the number of judges. Organizational Behavior and Human Decision Processes, 21, 121-129.

Libby, R., Trotman, K. T., & Zimmer, I. (1987). Member variation, recognition of expertise, and group performance. JAP, 72(1), 81-87.

Lim, J. S., & O'Connor, M. (1995). Judgmental adjustment of initial forecasts: Its effectiveness and biases. JBDM, 8, 149-168.

Lorge, I., Fox, D., Davitz, J., & Brenner, M. (1958). A survey of studies contrasting the quality of group performance and individual performance, 1920-1957. Psychological Bulletin, 55(6), 337-372.

Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions, and implications. International Journal of Forecasting, 16(14), 451-476.

Makridakis, S., & Winkler, R. L. (1983). The Combination of Forecasts. Journal of the Royal Statistical Society A, 146, 150-157.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). The Adaptive Decision Maker. New York: Cambridge University Press.

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. Cognitive Psychology, 38, 317-346.

Staël Von Holstein, C.-A. S. (1971). An experiment of probabilistic weather forecasting. Journal of Applied Meteorology, 10, 635-645.

Stigler, S. M. (1986). The History of Statistics: The Measurement of Uncertainty Before 1900. Cambridge, Mass: Belknap Press.

Tajfel, H. (1969). Social and cultural factors in perception. In G. Lindzey & E. Aronson (Eds.), The Handbook of Social Psychology (2 ed., Vol. 3). Reading, MA: Addison-Wesley.

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

Winkler, R. L. (1971). Probabilistic prediction: Some Experimental Results. Journal of the American Statistical Association, 66, 675-685.

Winkler, R. L. (1984). Combining Forecasts. In S. Makridakis & A. Andersen & R. Carbone & R. Fildes & M. Hibon & R. Lewandowski & J. Newton & E. Parzen & R. Winkler (Eds.), The Forecasting Accuracy of Major Time Series Methods. Chichester, England: John Wiley & Sons.

Yaniv, I., & Kleinberger, E. (2000). Advice Taking in Decision Making: Egocentric Discounting and Reputation Formation. Organizational Behavior and Human Decision Processes, 83(2), 260-281.

Yetton, P. W., & Bottger, P. C. (1982). Individual versus group problem solving: An empirical test of a best-member strategy. Organizational Behavior and Human Performance, 29, 307-321.

Table 1

Weighting Statistics

			% WX	
	$\overline{WS}$	$\overline{WX}$	≤ .6	= 1
			averaging	choosing
Study 1 (salaries feedback)				
No feedback	.70	.82	19.4	46.3
pre-test better than advisor	.77	.84	14.8	43.5
pre-test worse than advisor	.54	.84	17.5	48.9
Study 2 (salaries familiarity)				
ratings not exchanged	.71	.82	17.3	45.8
more familiar	.74	.85	18.1	55.4
similar or less familiar	.61	.80	23.1	39.8
Study 3 (country facts)				
Own country	.75	.81	23.9	42.5
Partner's country	.41	.76	29.1	27.2
Study 4 (five topics)				
Exchange				
High/low	.80	.89	9.6	60.2
High/High	.62	.88	10.1	58.8
Low/high	.48	.88	11.8	60.0
Low/low	.70	.89	12.6	62.7
No Exchange				
High/low	.68	.84	12.4	50.0
High/High	.61	.84	17.7	52.7
Low/high	.52	.85	16.8	48.3
Low/low	.57	.86	13.8	52.7

Table 2

Weighting and calibration statistics by judged probability in Experiment 3

	own country					partner's country						
judged				% WX					% WX		avg. judg-	proportion
probability	n	$\overline{WS}$	$\overline{WX}$	$\leq .6$	$= 1$	n	$\overline{WS}$	$\overline{WX}$	$\leq .6$	$= 1$	ed prob.	self closer
0	4	.42	.83	--	--	17	.17	.83	15.4	30.8	0	.38
.01-.39	15	.29	.75	33.3	33.3	103	.29	.78	17.9	28.4	.23	.46
.40-.60	126	.68	.74	35.7	29.6	138	.48	.73	40.2	25.2	.50	.50
.61-.99	154	.84	.85	15.9	49.0	55	.54	.76	28.3	28.3	.80	.52
1	16	.99	.99	0	93.3	1	--	--	--	--	1	.68

Table 3

Accuracy Results and Domain Characteristics

				MAD				
	MAD ratio	Bracket-ing rate	Expert detect'n	initial	revised	avg'ing	perfect choosing	consist-ent WS
Study 1								
Salaries	1.3	38%	51%	15.0_{a}	13.1_{b} *(12.4%)*	12.2_{c} *(18.1%)*	12.9_{b}	14.8_{b}
Study 2								
Salaries	1.3	41	43	17.9_{a}	15.8_{b} *(11.9)*	14.2_{c} *(20.8)*	15.7_{b}	15.1_{b}
Study 3								
Own country	1.8	39	58	13.6_{a}	12.4_{b} *(9.0)*	11.1_{c} *(18.0)*	10.8_{c}	13.4_{bc}
Partner's country	1.8	40	54	14.4_{a}	11.8_{b} *(18.0)*	11.2_{b} *(22.3)*	10.7_{b}	13.7_{b}
Study 4								
City distances	1.7	43	54	45.4_{a}	34.7_{b} *(23.6)*	37.5_{c} *(17.4)*	33.0_{c}	35.6_{c}
Dates of inventions	1.8	36	73	26.5_{a}	21.2_{bc} *(20.0)*	22.8_{b} *(14.0)*	20.1_{c}	23.7_{c}
Celebrity ages	1.7	37	60	8.8_{a}	7.4_{b} *(15.9)*	7.1_{b} *(19.3)*	6.7_{b}	8.4_{b}
Song rankings	1.3	20	54	15.6_{a}	15.0_{ab} *(3.8)*	14.7_{b} *(5.8)*	14.3_{b}	15.5_{b}
Basketball	1.7	34	64	9.8_{a}	7.8_{b} *(20.4)*	8.1_{b} *(17.3)*	7.6_{b}	9.2_{b}

Note: Within each row, MAD values that do not share a subscript differ significantly.

Table 4

Performance of Opinion Revision Strategies When the Judge is Less or More Accurate than Advisor

		Intuitive	Aver-	Perfect	Consistent
	Initial	Revision	aging	Choosing	WS
Study 1 (salaries feedback)					
Less accurate	$.56_{a}$	$-.29_{c}$	$-.73_{b}$	$-.68_{b}$	$-.35_{c}$
More accurate	$-.48_{a}$	$-.70_{b}$	$-.77_{b}$	$-.48_{a}$	$-.74_{b}$
Study 2 (salaries familiarity)					
Less accurate	$.54_{a}$	$-.28_{b}$	$-.91_{c}$	$-.46_{b}$	$-.48_{b}$
More accurate	$-.61_{a}$	$-.77_{ab}$	$-.88_{ab}$	$-.61_{a}$	$-.89_{b}$
Study 3 (country facts)					
Less accurate	$.56_{a}$	$-.26_{b}$	$-.52_{c}$	$-.58_{cd}$	$-.29_{bd}$
More accurate	$-.56_{a}$	$-.40_{a}$	$-.47_{a}$	$-.56_{a}$	$-.49_{a}$
Study 4 (five topics)					
Less accurate	$.48_{a}$	$-.25_{b}$	$-.34_{c}$	$-.48_{d}$	$-.29_{bc}$
More accurate	$-.46_{a}$	$-.46_{a}$	$-.32_{b}$	$-.46_{a}$	$-.46_{a}$

Note. Accuracy scores are standardized as described in the text. Lower scores reflect greater accuracy. Means in the same row not sharing a subscript differ significantly. The comparison between averaging and intuitive revision was planned and tested at $\alpha = .05$. The remaining nine pairwise comparisons were analyzed with multiple dependent *t* tests, setting $\alpha = .006$.

Figure 1

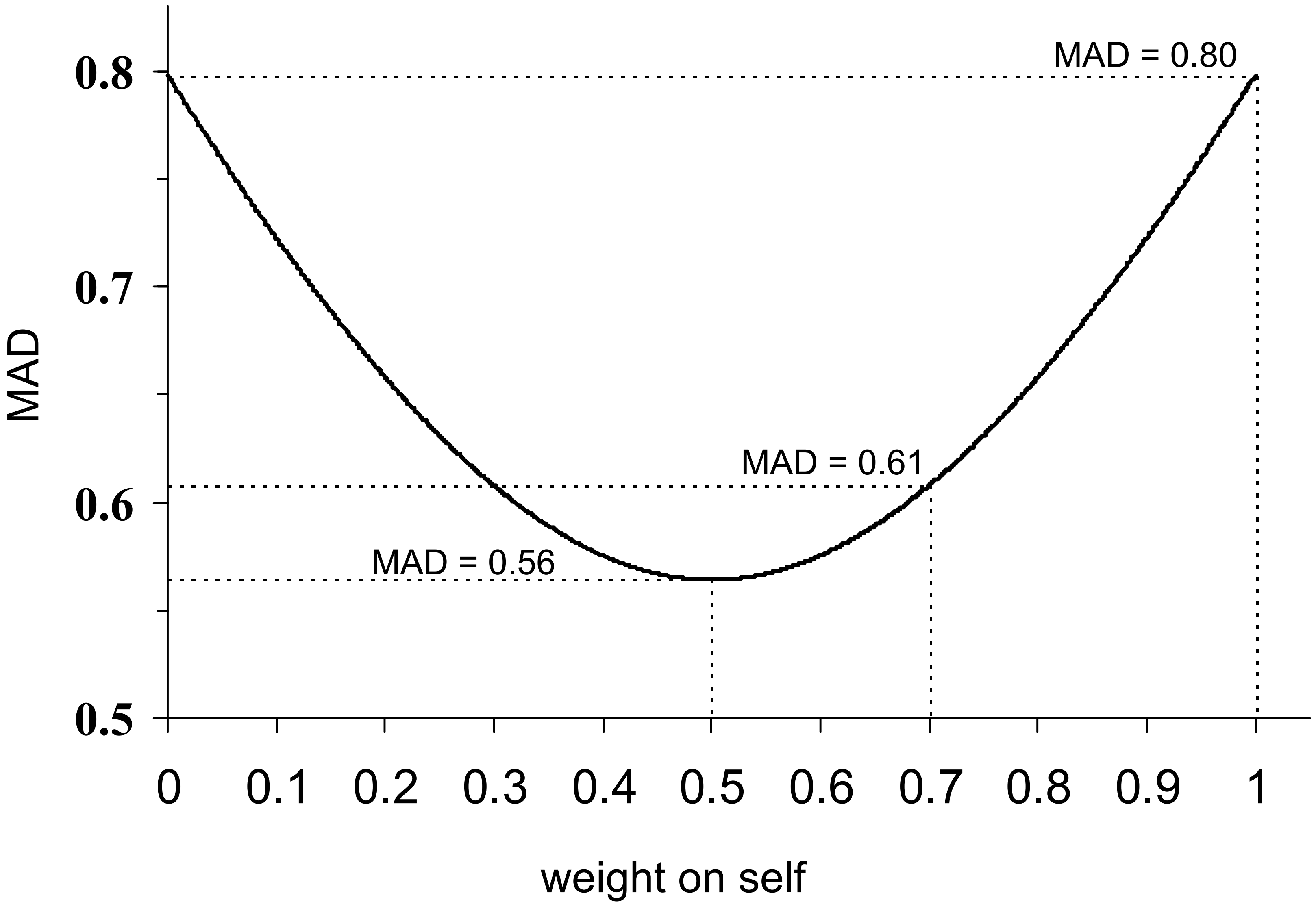

MAD = 0.80
MAD = 0.61
MAD = 0.56
0.8
0.7
0.6
0.5
MAD
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
weight on self

Figure 2

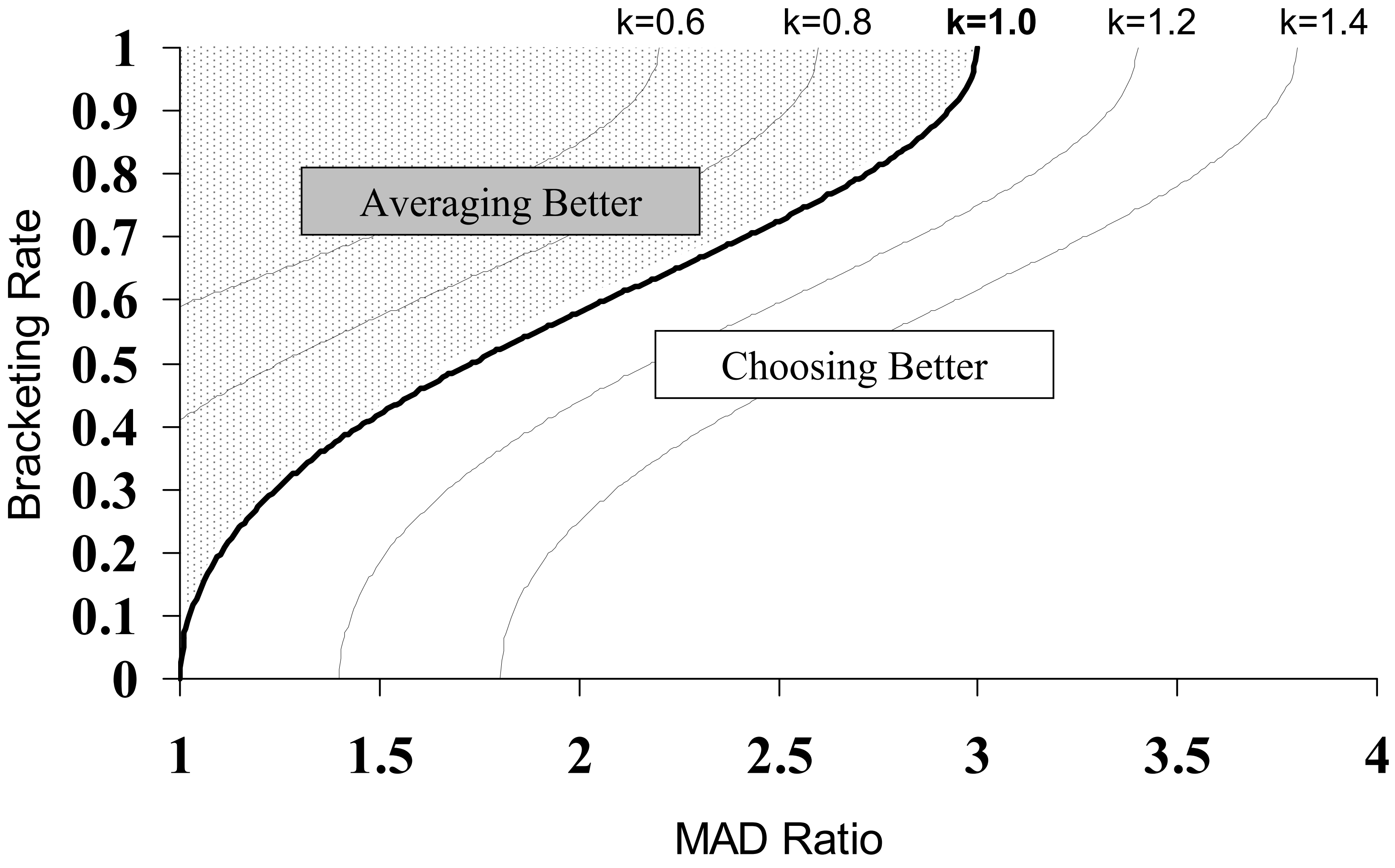
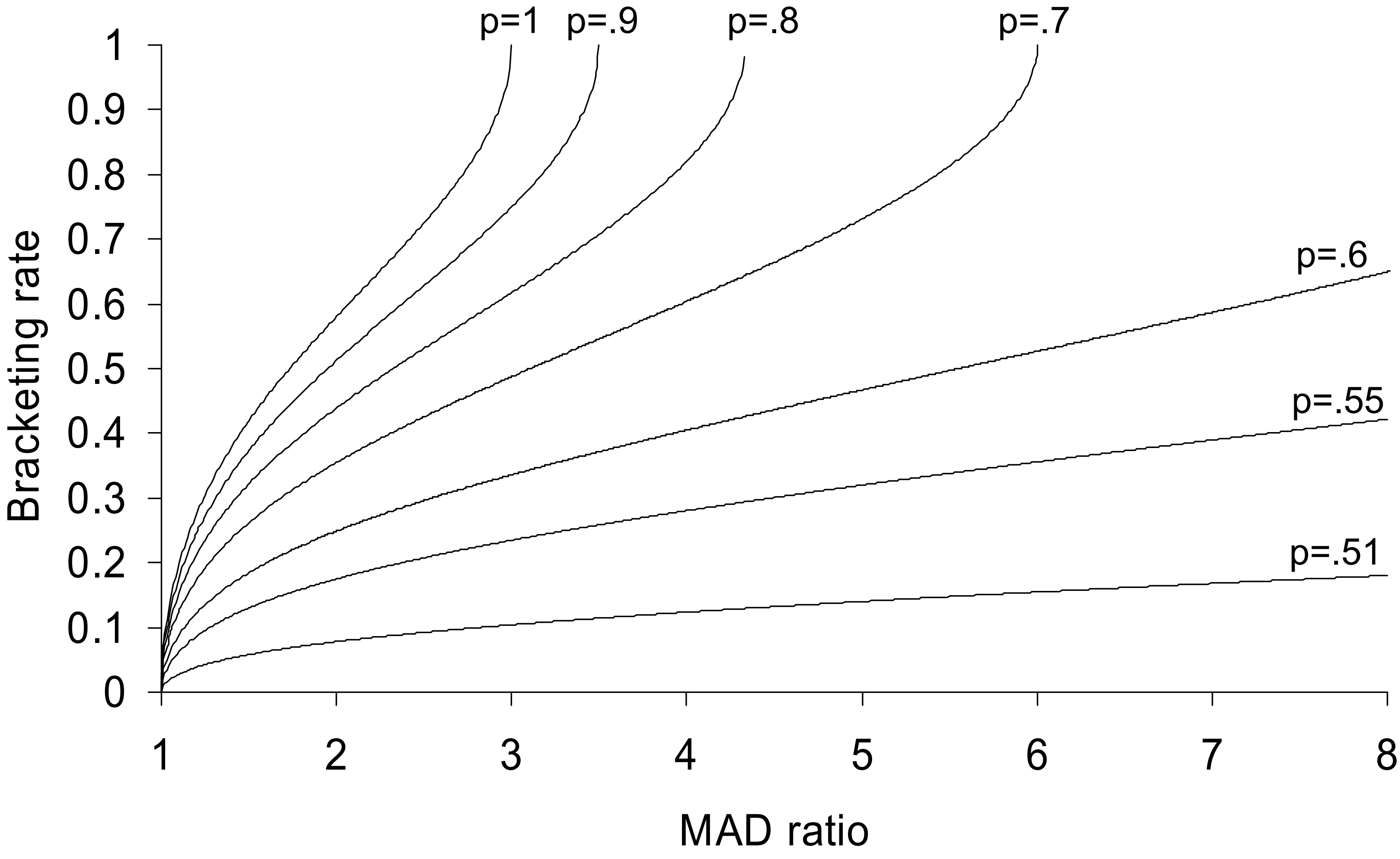

k=0.6
k=0.8
k=1.0
k=1.2
k=1.4
Averaging Better
Choosing Better
Bracketing Rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
1.5
2
2.5
3
3.5
4
MAD Ratio

Figure 3



p=1
p=.9
p=.8
p=.7
p=.6
p=.55
p=.51
Bracketing rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
7
8
MAD ratio

Figure 4

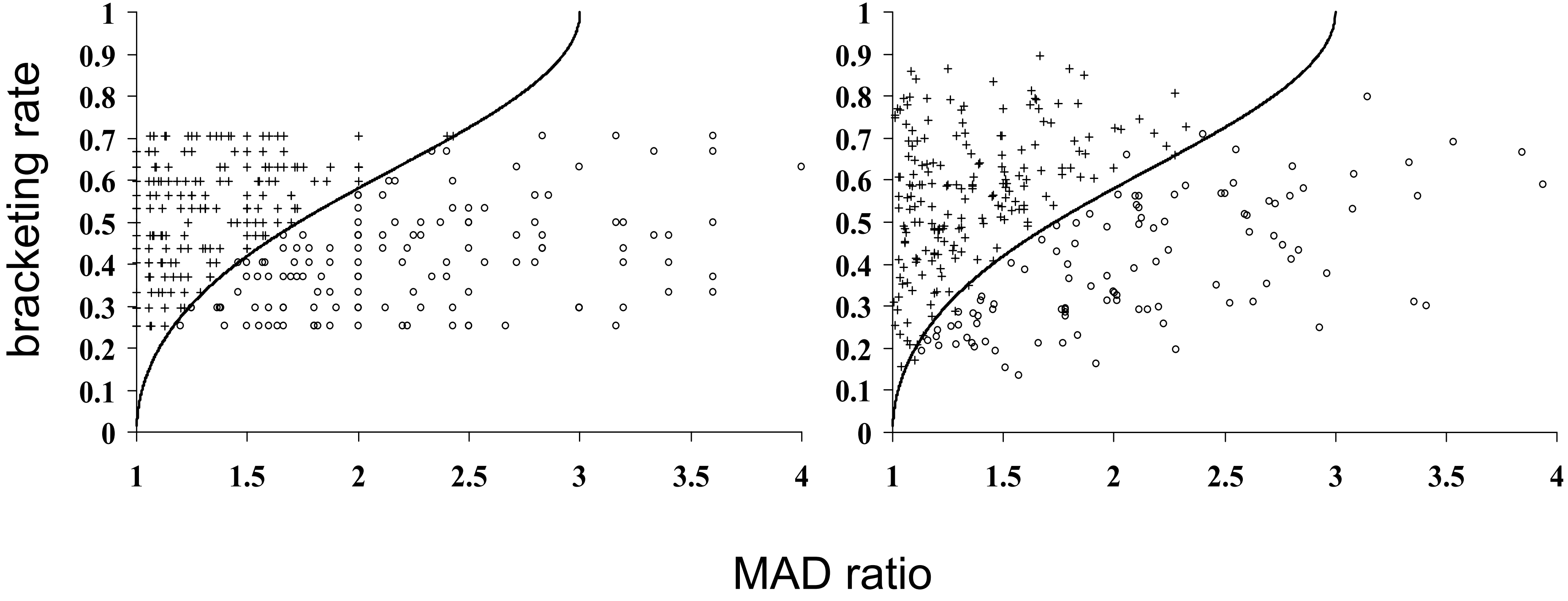
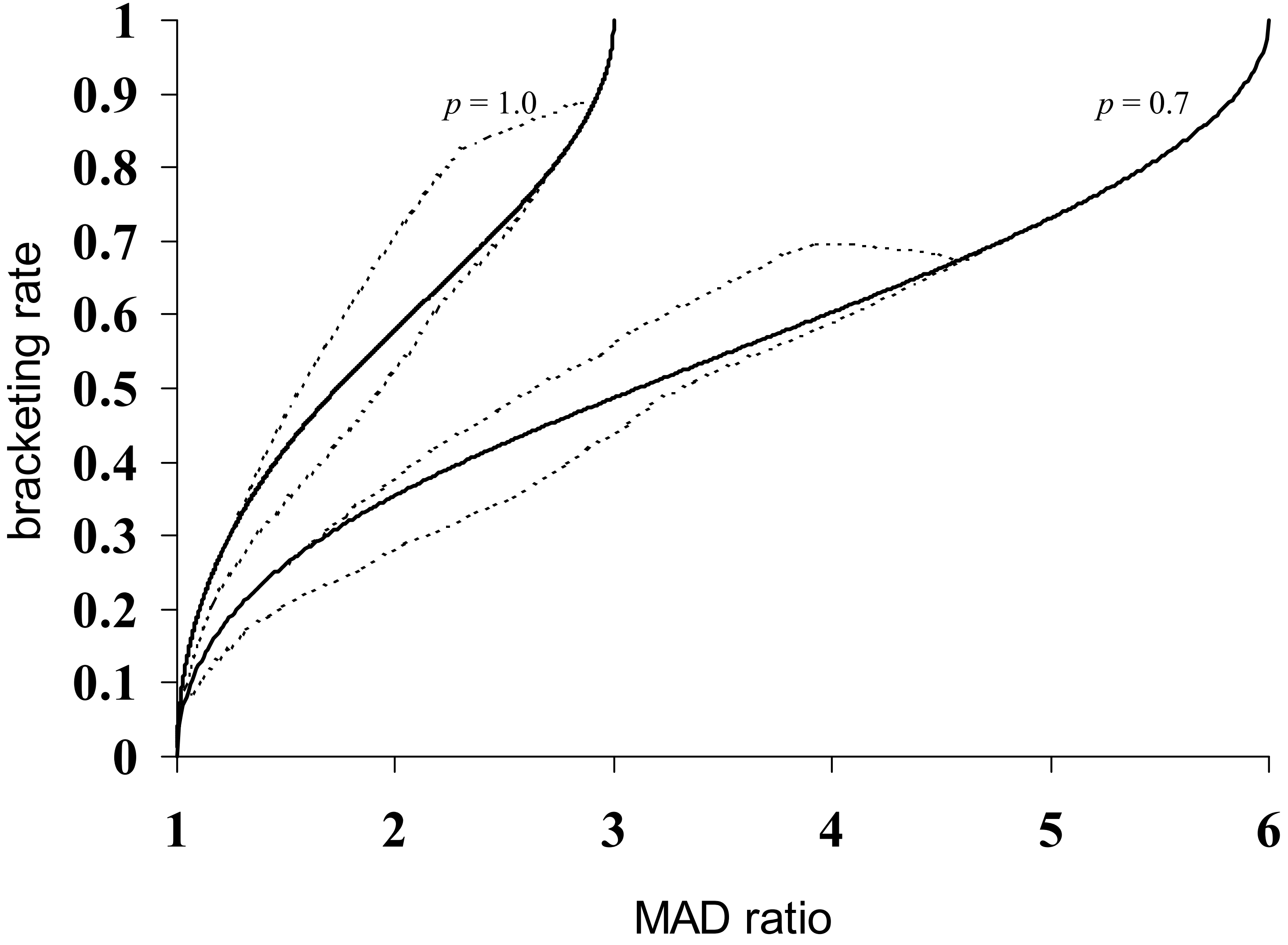
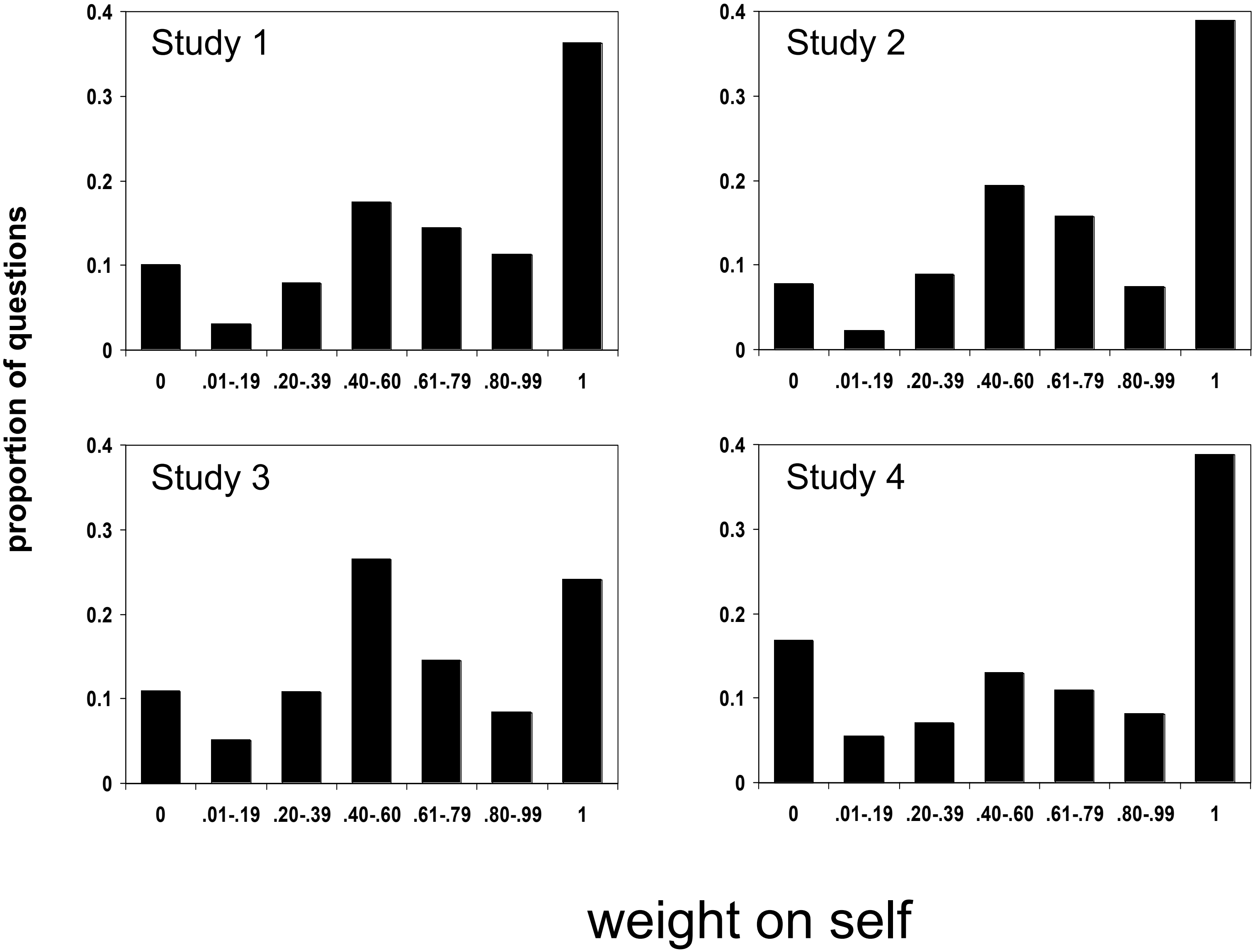

bracketing rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
1.5
2
2.5
3
3.5
4
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1
1.5
2
2.5
3
3.5
4
MAD ratio

Figure 5



bracketing rate
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
p = 1.0
p = 0.7
1
2
3
4
5
6
MAD ratio

Figure 6



proportion of questions
Study 1
Study 2
Study 3
Study 4
0.4
0.3
0.2
0.1
0
0
.01-.19
.20-.39
.40-.60
.61-.79
.80-.99
1
weight on self