Small Samples Do Not Cause Greater Accuracy—But Clear Data May Cause Small Samples: Comment on Fiedler and Kareev (2006)

Laurel Evans and Marc J. Buehner
Cardiff University

Fiedler and Kareev (2006) have claimed that taking a small sample of information (as opposed to a large one) can, in certain specific situations, lead to greater accuracy—beyond that gained by avoiding fatigue or overload. Specifically, they have argued that the propensity of small samples to provide more extreme evidence is sufficient to create an accuracy advantage in situations of high caution and uncertainty. However, a close examination of Fiedler and Kareev’s experimental results does not reveal any strong reason to conclude that small samples can cause greater accuracy. We argue that the negative correlation between sample size and accuracy that they reported (found only for the second half of Experiment 1) is also consistent with mental fatigue and that their data in general are consistent with the causal structure opposite to the one they suggest: Rather than small samples causing clear data, early clear data may cause participants to stop sampling. More importantly, Experiment 2 provides unequivocal evidence that large samples result in greater accuracy; Fiedler and Kareev only found a small sample advantage here when they artificially reduced the data set. Finally, we examine the model that Fiedler and Kareev used; they surmised that decision makers operate with a fixed threshold independent of sample size. We discuss evidence for an alternative (better performing) model that incorporates a dynamic threshold that lowers with sample size. We conclude that there is no evidence currently to suggest that humans benefit from taking a small sample, other than as a tactic for avoiding fatigue, overload, and/or opportunity cost—that is, there is no accuracy advantage inherent to small samples.

**Keywords:** sampling, information gathering, small sample advantage, law of large numbers

There has been little argument about the effect of sample size on choice since Bernoulli (1713) published the proof of his law of large numbers, which states the intuitive: that the sample average,

$$\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n),$$

converges toward the population average as sample size increases:

$$\lim_{n \to \infty} \bar{X}_n = \mu.$$

We can distinguish between the strength of the evidence (the sample average) and its weight (the sample size). If, for example, we were looking at the average rating of a restaurant by its patrons, a high rating (high strength) would be pleasing, especially if we knew it was the result of a great number of votes (high weight).

Of course, when a person is gathering information, there is a natural opportunity cost for taking a large sample—one could be doing other valuable things—and there are also the issues of memory restriction and mental fatigue. After all, there is little point in taking a sample that is so large that one begins to forget earlier items, as this is precisely what tends to happen (Erev & Barron, 2005; Hertwig, Barron, Weber, & Erev, 2004, 2006). The law of large numbers, however, only makes a statement about accuracy; if one’s sole goal is to make a correct choice, and fatigue, memory, and opportunity costs are not a problem, then one cannot do better than to sample as much as possible.

Yet, in the past four decades, and more so in recent years, there has been an upsurge of evidence suggesting that not only do people tend to take small samples (Hertwig et al., 2004; Tversky & Kahneman, 1971) but also that small samples are not as inaccurate as they might seem. Kareev and his colleagues, in particular, pioneered efforts to understand how the skew of sampling distributions of correlations affects the need to take a large sample (Kareev, 1995, 2000; Kareev, Lieberman, & Lev, 1997), although there has been argument about the varying roles of hits, misses, false alarms, and correct rejections (see Juslin, Fiedler, & Chater, 2006; Juslin & Olsson, 2005; Kareev, 2005). Hertwig and Pleskac (2008) also pointed out the diminishing returns of a large sample. Take the typical two deck gamble, in which the participant draws...
potential payout cards from each of two decks and tries to determine which will give the better payout. Hertwig and Pleskac identified several natural strategies people might take. If the natural mean heuristic strategy is used, taking a sample of 10 cards from each deck results in an 84% chance of choosing the deck with the higher expected utility; taking double that amount, meanwhile, results in 88%—only a 4-point increase. The results are similar for Bayesian strategies and for a strategy based on prospect theory: There is a curve of diminishing returns. In all cases, a larger sample always results in increased accuracy, thanks to the law of large numbers, but a surprisingly small sample will work well for much less effort. Thus, taking a small sample is not nearly as poor a strategy as one might initially think.

However, Fiedler and Kareev (2006) have gone one step further, arguing that taking a small sample is sometimes actually a better strategy from an accuracy-alone standpoint. Fiedler and Kareev have claimed that their research “is concerned with a distinct reversal of Bernoulli’s (1713) law of large numbers” (p. 884) and that “small samples can inform more correct [emphasis added] choices than large samples” (p. 884). Fiedler and Kareev have made these claims in the context of discussing the “recent research [that] has revealed several ways in which performance may decrease with increasing amount of information, quite independent of fatigue or overload [emphasis added]” (p. 883).

Summary of Fiedler and Kareev (2006)

Fiedler and Kareev (2006) first presented a model of choice to show that, theoretically, a small sample accuracy advantage\(^2\) can occur—and can occur for humans assuming this model is the one people use. Their research revolves around a problem much like the two deck gamble, in which bits of information (either “positive” or “negative”) are gathered about two potential job candidates. Fiedler and Kareev noted that this information is more extreme when samples are small, and they suggested that this property can be exploited in a model of choice. Imagine, for example, that one sees 90% positive information for Candidate A and 50% positive information for Candidate B—the difference between the two candidates is then given as \(0.9 - 0.5 = 0.4\); this is known as the contingency of the sample, or \(\Delta_{sample}\). We can see that \(\Delta_{sample}\) tends to be more extreme than for large samples (see Figure 1); this is because small samples tend to underrepresent rare events (see Hertwig & Pleskac, 2008, for a discussion and mathematical proof of this “amplification effect”). For instance, if Candidate A has, in the population, 90% positive information, but a person only draws a small sample (say four items), it is likely that he or she will see 100% positive information, by chance failing to draw any of the relatively fewer pieces of negative information. This amplified proportion would make the contingency in our example increase from \(0.4\) to \(0.5\), and in general many contingencies are likely to become more extreme. Fiedler and Kareev used this information to argue that, given a fixed threshold point, \(t\), “small samples, compared with large samples, gain more from correct choices (on the right) than they lose because of incorrect choices (on the left)” (p. 887; see Figure 1). Crucial to their conclusion is that a decision maker’s threshold for choice is always determined by a sole, fixed contingency—which does not change with sample size. In essence, Fiedler and Kareev assumed in this decision model that people’s choice thresholds (and, presumably, their confidence as well) rely solely on the contingency seen (the evidence’s strength), ignoring any effect sample size (weight) might have.

Bolstered by the theoretical possibility of a small sample accuracy advantage, Fiedler and Kareev (2006) conducted two empirical studies to search for the advantage in human behavior. In the first study, participants were asked to choose between two potential job candidates. They sampled pieces of evidence (which could either be positive or negative) until satisfied enough to stop and make a choice. Fiedler and Kareev found a positive or zero correlation between sample size (\(n\)) and a measure of correctness in Block 1, and they found some negative correlations between \(n\) and correctness in Block 2. Fiedler and Kareev also found negative correlations between the sample contingency, \(\Delta_{sample}\), and \(n\), concluding that “the smaller the sample resulting from self-determined information search, the more clear-cut the resulting evidence pointing in the correct direction” (p. 895). The second experiment was similar to the first, except that participants viewed samples of fixed size. In this case, participants always had better correctness scores for the large compared with the small samples. Fiedler and Kareev then “[controlled the threshold assumption statistically] by including only those trials on which \(\Delta_{sample}\) exceeded specific decision thresholds” (p. 899), finding that then the small sample accuracy advantage emerged.

---

1 It should be noted that this task deals with preference and not with inference, as the participants were choosing which payout deck they would prefer—they might, for example, prefer a risky, high-payoff deck even if it has a lower expected utility. However, assuming one values the highest expected utility, as most do (Hertwig et al., 2004), a small sample can yield surprisingly good results.

2 Fiedler and Kareev (2006) referred to this advantage as the small sample advantage throughout, but we prefer to refer to it as the small sample accuracy advantage to avoid confusion with advantages associated with mental fatigue or opportunity costs.
Scope of Our Critique

We argue that, first of all, Fiedler and Kareev’s (2006) evidence does not support the hypothesis that humans make use of a small sample accuracy advantage. Most importantly, we draw attention to the fact that, when presented with a small versus a large sample (in Experiment 2), participants were always more accurate with the large sample. We can think of no better test of their hypothesis, and the results clearly point to a large sample advantage. However, even assuming some error in this experiment, we do not find evidence for the small sample accuracy advantage in Experiment 1 either. Here, participants were allowed to truncate their own samples; they were more accurate or equally accurate with large than small samples in Block 1, and they were less accurate with large samples in Block 2. Naturally, this evidence is correlational and could be the result of any number of causes. We argue that a likely alternative explanation for the negative correlation in Block 2 is that mental fatigue set in after a long and tedious experiment, and overall, we contest Fiedler and Kareev’s claim that small samples cause greater accuracy, instead arguing that clear data cause smaller samples. We stress that, even if this causal structure is incorrect, given the results of Experiment 2, it is unlikely that there is a small sample accuracy advantage.

We also briefly examine the decision model laid out by Fiedler and Kareev (2006), which suggests that people employ a fixed, contingency-based decision threshold when making sampling-based choices. Cahan (2010) already provided an excellent discussion of how this decision model is statistically invalid, explaining how “better information (provided by large samples) can be translated into worse decisions if the decision rule is faulty” (p. 840). We point out that Fiedler and Kareev’s model is also not a good model of human behavior. We discuss reasons why the model does not produce optimal decision making, in addition to evidence that human choice making is better described by a different model.

Discussion of Fiedler and Kareev’s (2006) Evidence

We begin by discussing Fiedler and Kareev’s (2006) Experiment 2, which we think is the strongest test of the small sample accuracy advantage hypothesis. For each trial in this experiment, participants were asked to choose the superior of two products on the basis of a fixed amount of evidence presented to them on a computer screen; they were also allowed to make no choice if no evidence for the small sample (Griffin & Tversky, 1992). Therefore, decision makers hardly ever refrain from making a choice” (p. 898). When discussing their analysis, Fiedler and Kareev pointed out that very few “no choice” decisions were made (on average 4.54 out of 60 trials), saying, “As in a sunk-cost situation, having expended cognitive effort in a sample that may be longer or shorter than a self-determined sample, decision makers hardly ever refrain from making a choice” (p. 898). In other words, they justified their application of the “threshold assumption” by suggesting that participants were not adhering to their natural (high) thresholds, as it would in a self-determined sample. Fiedler and Kareev justified their selection of this subset in terms of a “threshold assumption,” as their hypothesis specifically involves situations in which thresholds are high: “A small-sample advantage can only be expected for relatively high decision thresholds, that is, situations in which organisms wait for an unusually clear-cut picture of evidence before they make a choice” (p. 887). When discussing their analysis, Fiedler and Kareev pointed out that very few “no choice” decisions were made (on average 4.54 out of 60 trials), saying, “As in a sunk-cost situation, having expended cognitive effort in a sample that may be longer or shorter than a self-determined sample, decision makers hardly ever refrain from making a choice” (p. 898). In other words, they justified their application of the “threshold assumption” by suggesting that participants were not adhering to their natural (high) thresholds, as it would in a self-determined sample. Fiedler and Kareev’s post hoc selection criteria therefore discarded trials in which Δsmall did not exceed the thresholds they picked, as if participants had not made choices on those trials.

First, although we agree that the sunk-cost explanation is plausible, we must note that it is purely speculation and therefore does not warrant excluding data. Even if it were true, one would need to have evidence that humans determine their thresholds by Δsmall alone to justify acting in their stead by looking at a subset of trials determined by Δsmall values. However, as we argue later, not only did Fiedler and Kareev (2006) fail to provide such evidence but there already exists evidence that people in fact rely on other factors in addition to Δsmall (Griffin & Tversky, 1992). Therefore, it is erroneous to apply such a cutoff rule post hoc.

In sum, it appears that Fiedler and Kareev (2006) have indeed identified an interesting analytic rule such that, if one applies it to a subset of data, small samples in this subset tend to be related to higher accuracy. However, as their evidence shows, participants did not constrain themselves to such a subset, and Fiedler and Kareev did not provide sufficient reason to apply their rule post hoc and to draw conclusions about human behavior.

Fiedler and Kareev (2006) also analyzed the Δsmall measure, claiming that it too provides some insight into whether there is a small sample accuracy advantage. They looked at r(Δsample, n),

shown small versus large samples, participants would be more accurate with small samples—and the evidence clearly shows that they are more accurate with large samples. Fiedler and Kareev conceded that this “indicates] the normal superiority of large samples” (p. 899). They pointed out that small samples do fare reasonably well, noting that, in certain conditions, “the performance difference between small and large samples was rather modest” (p. 899), a finding in line with the evidence that small samples can perform surprisingly well (Hertwig & Pleskac, 2008), although not as well as large samples.

Fiedler and Kareev (2006) then examined only those trials on which Δsample exceeded a given contingency. When doing so, they found that small samples were associated with higher accuracy—higher Δchoice, and also higher f(correct) − f(incorrect) and higher Δconchoice, two other measures that they developed for measuring accuracy3 (see Fiedler & Kareev, 2006, Tables 6 and 7). Fiedler and Kareev justified their selection of this subset in terms of a “threshold assumption,” as their hypothesis specifically involves situations in which thresholds are high: “A small-sample advantage can only be expected for relatively high decision thresholds, that is, situations in which organisms wait for an unusually clear-cut picture of evidence before they make a choice” (p. 887). When discussing their analysis, Fiedler and Kareev pointed out that very few “no choice” decisions were made (on average 4.54 out of 60 trials), saying, “As in a sunk-cost situation, having expended cognitive effort in a sample that may be longer or shorter than a self-determined sample, decision makers hardly ever refrain from making a choice” (p. 898). In other words, they justified their application of the “threshold assumption” by suggesting that participants were not adhering to their natural (high) thresholds, as it would in a self-determined sample. Fiedler and Kareev’s post hoc selection criteria therefore discarded trials in which Δsmall did not exceed the thresholds they picked, as if participants had not made choices on those trials.

First, although we agree that the sunk-cost explanation is plausible, we must note that it is purely speculation and therefore does not warrant excluding data. Even if it were true, one would need to have evidence that humans determine their thresholds by Δsmall alone to justify acting in their stead by looking at a subset of trials determined by Δsmall values. However, as we argue later, not only did Fiedler and Kareev (2006) fail to provide such evidence but there already exists evidence that people in fact rely on other factors in addition to Δsmall (Griffin & Tversky, 1992). Therefore, it is erroneous to apply such a cutoff rule post hoc.

In sum, it appears that Fiedler and Kareev (2006) have indeed identified an interesting analytic rule such that, if one applies it to a subset of data, small samples in this subset tend to be related to higher accuracy. However, as their evidence shows, participants did not constrain themselves to such a subset, and Fiedler and Kareev did not provide sufficient reason to apply their rule post hoc and to draw conclusions about human behavior.

Fiedler and Kareev (2006) also analyzed the Δsmall measure, claiming that it too provides some insight into whether there is a small sample accuracy advantage. They looked at r(Δsample, n),

3 The equation f(correct) − f(incorrect) is the difference in frequencies of correct and incorrect choices, which is very similar to Δchoice. We discuss Δconchoice later.
finding virtually zero correlation for all population contingencies, \( r = -0.02, 0.23, \) and \(-0.035\) for \( \Delta_{\text{population}} = -0.1, 0.2, \) and \(0.4\), respectively. Fiedler and Kareev took this as evidence that “across all choices, when the threshold assumption is not met, there is no small-sample advantage” (p. 899), having explained earlier that “negative correlations between \( \Delta_{\text{sample}} \) and \( n \) indicate a small-sample advantage” (p. 895).

We note that a zero correlation between \( \Delta_{\text{sample}} \) and \( n \) is exactly what one would expect to find in this experiment. By the law of large numbers, in a fixed sample of randomly drawn evidence, \( \Delta_{\text{sample}} \) will approach \( \Delta_{\text{population}} \) as \( n \) grows (see Figure 1). What does this mean for the correlation between \( \Delta_{\text{sample}} \) and \( n \)? Consider a plot where \( \Delta_{\text{sample}} \) is represented on the abscissa, and \( n \) on the ordinate. The law of large number dictates such a plot to look like a scatterplot of a normal distribution with mean equal to \( \Delta_{\text{population}} \). In other words, when sample size is small, \( \Delta_{\text{sample}} \) can range widely, from low negative to high positive, and when sample size is large, we would expect a \( \Delta_{\text{sample}} \) value near to \( \Delta_{\text{population}} \). Because this distribution always centers around \( \Delta_{\text{population}} \) for both low and high \( n \), the correlation between \( \Delta_{\text{sample}} \) and \( n \) should always be around zero. We can think of no reason this correlation would be informative about the accuracy of small samples.

Finally, we discuss the confidence reported by participants. Fiedler and Kareev (2006) found a significant main effect of sample size on confidence, \( F(1, 88) = 4.91, p < .05 \), reflecting higher confidence with small than with large samples, and a Contingency Level \( \times \) Sample Size interaction, \( F(1, 176) = 12.47, p < .001 \), indicating that higher confidence for small samples occurred for the weaker population contingencies \((1.1 \) and \(2.2)\). Fiedler and Kareev said that “the enhanced confidence of small-sample choices is consistent with the fact that small samples often provide a clear-cut picture of evidence” (p. 899). However, given that the evidence from this experiment does not suggest a small sample accuracy advantage, we cannot conclude that the small samples were more “clear-cut” to the participants, in terms of providing better information. Perhaps these results were instead caused by a kind of mental overloading, as there is more to process in a large sample, and this may make it more difficult to be confident in one’s appraisal. Such difficulty would likely be enhanced in the conditions that involved the faces disappearing after presentation but could still exist even in the other conditions. Thus, a small sample is more clear-cut, in this case, in that it is easier for the participant to process.

We now move on to discuss Fiedler and Kareev’s (2006) Experiment 1. This experiment was similar to Experiment 2, except participants had control over the size of the sample. They held down a button to see generated evidence (also smiley and frownie faces), letting go when they were ready to make a choice. Each face appeared on screen for 500 ms before disappearing again, and again, evidence was drawn randomly (with replacement) from populations with contingencies \((1.1, 2.2, \) or \(4.4)\). We do not consider the data from \( \Delta_{\text{population}} = 4.4 \), for which there was a ceiling effect for accuracy.

For the first trial block, Fiedler and Kareev (2006) found a significantly positive correlation between \( \Delta_{\text{choice}} \) and \( n \) for \( \Delta_{\text{population}} = -0.1, \) \( r = 0.13, p < .05 \), and no correlation for \( \Delta_{\text{population}} = -0.2 \). This indicates the usual accuracy advantage of large samples. For the second trial block, they found significantly negative correlations between \( \Delta_{\text{choice}} \) and \( n \) for \( \Delta_{\text{population}} = -0.1 \) and \( \Delta_{\text{population}} = -0.2 \), at \( r = -0.19 \) and \( r = -0.10 \), respectively (both \( ps < .05 \)). Participants were less accurate with larger samples. The small sample accuracy advantage suggested by Fiedler and Kareev (2006) thus can at best explain 50% of the evidence derived from correlating sample size and correctness in Experiment 1; the remaining 50% of the evidence is directly contradicting their argument. More importantly, the evidence they presented is correlational in nature, and given our assessment of Experiment 2, it is very unlikely that the causal relationship between sample size and accuracy is driven by a small sample accuracy advantage. We return to this issue later when we consider an alternative explanation.

Fiedler and Kareev (2006) also looked at the correlation between \( n \) and a measure they called \( \Delta_{\text{con choice}} \), which is \( \Delta_{\text{choice}} \) multiplied by the participant’s reported confidence for that trial. They reported significant negative correlations in Block 1 for \( \Delta_{\text{population}} = -0.2, \) \( r = -0.21 \), and in Block 2 for \( \Delta_{\text{population}} = -0.1, \) \( r = -0.21 \), and for \( \Delta_{\text{population}} = -0.2, \) \( r = -0.44 \) (all \( ps < .05 \)). However, \( \Delta_{\text{con choice}} \) is two separate measures (confidence and accuracy) combined. Although theoretically this kind of measure “weights” accuracy in terms of confidence, it is difficult to determine the meaning of a correlation between such a double measure and \( n \). With careful examination, one can see that there is little this measure can tell us except about confidence. Note that, in all cases, mean \( \Delta_{\text{choice}} \) scores were equal to or greater than \( -0.1 \) (see Fiedler & Kareev, 2006, Table 2), indicating that people were far more often accurate than inaccurate. Because most data for the correlation between \( \Delta_{\text{choose}} \) and \( n \) would then come from accurate trials, the relationship between \( \Delta_{\text{con choice}} \) and \( n \) would likely be driven by the relationship between confidence and \( n \) on those trials. In other words, the accuracy/inaccuracy relationship quite possibly plays little role because there are too few inaccurate trials to affect the correlation much. Therefore, all this measure appears to tell us is that people are more confident with a smaller sample, which has already been documented in the literature (Vickers, Smith, Burt, & Brown, 1985). Again, we later discuss different possible causes for this relationship.

Finally, we discuss Fiedler and Kareev’s (2006) \( \Delta_{\text{sample}} \) results. Earlier, we argued that, in a fixed sample of randomly drawn evidence (i.e., Experiment 2), the correlation between \( \Delta_{\text{sample}} \) and \( n \) is uninformative and should always be close to zero. In Experiment 1, however, participants self-truncated the sampling process, and we therefore consider whether and how a correlation between \( \Delta_{\text{sample}} \) and \( n \) might be meaningful under these circumstances.

In Experiment 1, Fiedler and Kareev (2006) found significantly negative correlations between \( \Delta_{\text{sample}} \) and \( n \) in all conditions; in Block 1, \( r = -0.23, -0.45, \) and \(-0.51, \) for \( \Delta_{\text{population}} = -0.1, 0.2, \) and \(0.4, \) respectively, and in Block 2, \( r = -0.42, -0.54, \) and \(-0.65, \) for \( \Delta_{\text{population}} = -0.1, 0.2, \) and \(0.4, \) respectively (all \( ps < .001 \)). Fiedler and Kareev claimed that this “reflects a pervasive small-sample advantage” (p. 895), saying, “the smaller the sample resulting from self-determined information search, the more clear-cut the resulting evidence” (p. 895). They also noted that there was a significant main effect of trial block, \( F(1, 34) = 15.68, p < .001 \), indicating stronger negative correlations in Block 2, and they suggested that this reflected “adaptive learning.” This was also reflected in an analysis of variance of sample sizes at the different \( \Delta_{\text{population}} \) levels in both blocks; there was a significant Contingency Level \( \times \) Trial Block interaction, \( F(2, 68) = 21.39, p < .001 \), indicating that a
wider range of sample sizes was used in the second trial block—reduced \( n \) for strong contingencies and increased \( n \) for weak contingencies.

Fiedler and Kareev’s (2006) interpretation of the negative correlations between \( \Delta_{\text{sample}} \) and \( n \) unequivocally argues that (a) small samples cause higher \( \Delta_{\text{sample}} \), which in turn (b) results in small samples being more clear-cut. In other words, Fiedler and Kareev’s rationale behind the second argument (b)—small samples are more clear-cut—clearly assumes that higher \( \Delta_{\text{sample}} \) equates with clear-cut evidence. Fiedler and Kareev in fact provided no evidence for this latter assumption; furthermore, it is contradictory to long-held notions of sample size and statistical power. We would like to put forward a different interpretation for the negative correlations between \( \Delta_{\text{sample}} \) and \( n \) in Experiment 1, an interpretation that is based on rational principles of sampling and statistical power that can explain the evidence from Experiment 1 equally well.

Our alternative explanation involves a causal structure that is directly opposite to what Fiedler and Kareev (2006) have suggested: Instead of small samples causing clear-cut evidence, clear-cut evidence causes small samples. Considering what may be driving the correlation between \( \Delta_{\text{sample}} \) and \( n \) in a self-truncated sampling task, we suggest that when people see a high enough \( \Delta_{\text{sample}} \) early (i.e., at small \( n \)), they are likely to consider the evidence clear and stop early to make their choice; when \( \Delta_{\text{sample}} \) is too low, they continue sampling. Because \( \Delta_{\text{sample}} \) tends to approach \( \Delta_{\text{population}} \) as \( n \) grows (see Figure 1), \( \Delta_{\text{sample}} \) will have a tendency to shrink in magnitude with sampling, and therefore large \( n \) will likely be paired with lower \( \Delta_{\text{sample}} \). As we discuss in depth later, with larger \( n \) people are more likely to accept a comparatively lower \( \Delta_{\text{sample}} \), as clear evidence (Griffin & Tversky, 1992) and stop to make a choice. Taken together, this behavior means a high \( \Delta_{\text{sample}} \) (either positive or negative) is likely to be paired with a small \( n \), and any \( \Delta_{\text{sample}} \) of lower magnitude is likely to be paired with a larger \( n \). However, because there is always a difference between the candidates, any high \( \Delta_{\text{sample}} \) the participant sees is more likely to be in favor of the better candidate than the worse candidate (see Figure 1); therefore, the high \( \Delta_{\text{sample}} \) seen is likely largely positive (rather than largely negative). Taken together, this means there will be many high, positive \( \Delta_{\text{sample}} \) values paired with small \( n \); only a few high, negative \( \Delta_{\text{sample}} \) values paired with small \( n \); and the lower magnitude \( \Delta_{\text{sample}} \) values will be paired with large \( n \). This should result in a negative correlation, as was found—and it is entirely consistent with our claim.

Our alternative causal explanation can also account for the results for adaptive learning. In this case, perhaps people are getting better at encoding the flashing faces, so they are able to spot the very clear-cut evidence earlier and, thus, stop viewing evidence earlier. By encoding we mean an attentional process: The participants learn to direct their attention more efficiently around the screen. In other words, as they become familiar with the environment, participants learn to ignore irrelevant background information and focus on the faces. Once they learn to focus in this way, the result may be a better memory of the faces and, therefore, a better ability to stop as soon as clear evidence is detected. However, as we discuss later, it is still possible that a different part of their memory ability—their ability to retain information—will erode as the memory system becomes fatigued.

Also, clear-cut evidence is, of course, more likely to lead to a confident choice, so our explanation can also account for the negative relationship between \( \Delta_{\text{choice}} \) and \( n \). In other words, it is possible that when people see clear evidence early, they are likely to become confident and therefore stop early (with a small sample). They would then tend to have lower confidence with larger samples precisely because they are not confident enough to stop early and must often continue sampling until just barely sure (or perhaps they may even stop eventually to guess). Even if they like to continue until they are very certain, there are bound to be situations in which they tire and accept a sample with less certainty.

Finally, we return to consider the correlations between \( \Delta_{\text{choice}} \) and \( n \). As we mentioned, the relationship between \( n \) and accuracy goes from positive/nonsignificant in Block 1 to negative in Block 2, and this cannot be explained solely by our suggestion of causality. One possible interpretation of this change is that the large sample advantage holds initially but is corroded over time by mental fatigue. Consider this: We have just suggested that decision makers may stop early and make a choice when the early sample is clear-cut. Conversely, when a small initial sample is not clear, they will likely continue to sample to achieve more clarity. This behavior may produce optimal performance at first, but after many trials, the ability to remember items may erode; extended sampling thus may not produce the desired clarity when the decision maker is unable to make use of all the obtained evidence. More specifically, we hypothesize that participants, worn-out by having solved many choice problems already, will not gain as much from extended sampling as they would ordinarily, and they will revert to guessing simply to terminate an apparently difficult choice problem. We do not claim that they will be fatigued enough to take smaller samples overall (they do not), but rather we suggest that their performance related to large samples declines slightly, because of difficulty remembering all the information they have seen. Essentially, over the course of many decisions, choices based on large samples may move from being based on initially mixed data (that likely become clearer with time) to being based on data that may well remain mixed because it is difficult to remember all of it. Meanwhile, small samples, by their very nature, are easier to remember, and so performance related to small samples remains as it was when participants were fresh—even improves, according to our earlier argument that encoding ability may improve. Thus, small sample performance improves slightly, whereas large sample performance declines slightly, and in this way large sample performance could become poorer than small sample performance, even while overall accuracy remains the same (as it does in this experiment). This explanation would account for the initial positive correlation between sample size and correctness as well as the subsequent negative correlation. We leave it to future research to determine the extent to which this hypothesis is supported.

In sum, Fiedler and Kareev (2006) found that, when they presented participants with either small or large samples (Experiment 2), participants were always more accurate with large ones. When they allowed participants to truncate their own samples (Experiment 1), participants were generally more accurate with large samples for Block 1 but were more accurate with small samples for Block 2. These data together are not consistent with the predictions of the small sample advantage hypothesis, which would suggest small sample superiority in all cases. Experiment 2, being
a true test of causality, indicates that large, rather than small, samples cause greater accuracy. Experiment 1 provides only correlational data, which may be explained in any number of ways; we point out that, rather than small samples causing clear-cut evidence, the opposite causal structure is equally plausible—clear-cut evidence may cause small samples. Taking the two experiments together, we suggest that large samples are superior except when mental fatigue erodes one’s memory of earlier evidence—although we hasten to add that, even if fatigue is not a factor, the results of Experiment 2 strongly deny that a small sample accuracy advantage is.

Discussion of Fiedler and Kareev’s (2006) Model

We turn now to the decision model that inspired Fiedler and Kareev (2006) to search for the small sample accuracy advantage in human behavior. We argue first that their model does not produce optimal decision-making, and second that it does not appear to describe human behavior.

On the surface, it is a very intuitive model: A person samples information until some measure of the data seen reaches an internal threshold, at which point he or she stops and makes a choice. The model is also in line with the current literature on models of choice (see, e.g., Busemeyer & Townsend, 1993; Pleskac & Busemeyer, 2007; Ratcliff & Starns, 2009), which features models with just such a fixed, internal threshold. Fiedler and Kareev (2006) suggested that the data measure that reaches threshold in their choice situation is the contingency the sampler sees. With this assumption in place, they showed on statistical grounds that a small sample would give an accuracy advantage to a decision maker using their model and their measure of accuracy—in this case $f(\text{correct}) - f(\text{incorrect})$, which is the difference between the number of correct and incorrect choices. They showed this small-sample superiority by analyzing a computer simulation of their model, in which they tallied the number of correct and incorrect choices out of 10,000 simulated trials for sample sizes $n = 4$, 8, 12, 16, 24, and 32, across $\Delta_{\text{population}} = .1$, .2, .3, .4, .5, and .6 (see Fiedler & Kareev, 2006, Table 1 and Figure 4). Their results clearly indicate that $f(\text{correct}) - f(\text{incorrect})$, when observed at high decision thresholds, was higher for small than for large samples.

However, Fiedler and Kareev (2006) never justify using the contingency as a basis for their threshold, either in terms of evidence about optimality or about human behavior; they simply applied the threshold and reported the results. We observe that, if one wants to make optimal decisions, the contingency is not the appropriate measure for the threshold; instead, we suggest that a contingency-based threshold should lower with sample size. When applied, this kind of threshold overall produces the best $f(\text{correct}) - f(\text{incorrect})$.

To explain our suggestion, we first note that an optimal decision maker should take into account the probability a given choice will be correct. In this way, she can determine the likelihood of correctness for a sample and act accordingly—stopping if in making a choice she is very likely to be correct, or sampling further if not. By the law of large numbers, the more one samples, the greater the likelihood of observing a value closer to the population contingency—thus, the better the sampler is informed about which choice is correct. We can clearly see this trend in Fiedler and Kareev’s (2006) simulation data (with probability of correctness obtained by dividing correct choices by total choices made). Their results show that the probability of a correct choice increases with sample size (see Fiedler & Kareev, 2006, Table 1), for example, increasing from $.64$ ($n = 4$) to $.98$ ($n = 32$) for $\Delta_{\text{population}} = .1$, $t = .6$. Furthermore, the more extreme $\Delta_{\text{sample}}$ one sees, the more likely one is to be correct; the probability of correctness also, then, should increase with threshold. This too is evident from Fiedler and Kareev’s data: For example, a 38% chance of correctness at $t = .1$ increases to 81% at $t = 1.0$ for $\Delta_{\text{population}} = .1$, $n = 8$. These two facts together—that likelihood of correctness increases with both sample size and threshold—are the reason that a lowering threshold is an optimal strategy for accuracy. For small samples, one can obtain a decent probability of accuracy if the threshold is set high, and, if sampling needs to continue, one can steadily lower the threshold according to what will produce the same likelihood of correctness. As an example, observing Fiedler and Kareev’s Table 1, if one would like a 70% accuracy rate when $\Delta_{\text{population}} = .1$, then one’s threshold should be .8 for $n = 4$ (which yields 906 correct answers out of a total of 1,301 answers, or 70% correct) and can lower to .1 for $n = 32$ (which yields 6,149 correct answers out of a total of 8,318 answers, or 74% correct). Although higher thresholds will still produce higher accuracy for large samples, low thresholds work just as well. Therefore nothing is lost, in terms of probability of correctness, by lowering one’s threshold.

However, probability of correctness is not the best measure for determining overall good performance on this type of choice task. We wholeheartedly agree with Fiedler and Kareev (2006) when they argued that the ratio of correct to incorrect answers—which contains information equivalent to probability of correctness—is lacking as an overall measure of accuracy. Fiedler and Kareev said,

In the context of this satisficing-choice model the difference, rather than the ratio, of hit and false alarm rates provides the relevant measure of [the expected value], the overall sum of all benefits and costs. Gaining $+8$ [value] from 10 correct and 2 incorrect choices is clearly superior to gaining $+4$ [value] from 5 correct and 1 incorrect decisions in the same time period. (p. 886)

Thus, we now show that not only does a lowering-threshold strategy not lose anything in terms of chance to obtain a correct answer but it also outperforms the fixed-threshold strategy when using the difference measure, $f(\text{correct}) - f(\text{incorrect})$.

Fiedler and Kareev (2006) argued that when applying the difference measure, one obtains a small sample accuracy advantage. We think that they applied this measure to their data unfairly, as they only considered comparisons between small and large samples when thresholds were high. Making a comparison only within this subset of the data results in small samples performing decently, whereas large samples tend to result in very few choices. For example, for $n = 32$, $\Delta_{\text{population}} = .1$, and a threshold of .4, the simulation made 771 correct choices, 59 incorrect choices, and 9,170 no-choices; $f(\text{correct}) - f(\text{incorrect}) =$ 712. Meanwhile, for $n = 4$, $\Delta_{\text{population}} = .1$, and a threshold of .4, the simulation made 3,469 correct choices, 2,132 incorrect choices, and 4,399 no-
choices, yielding \( f(\text{correct}) - f(\text{incorrect}) = 1,337 \). Thus, for high thresholds, smaller samples tended to result in higher \( f(\text{correct}) - f(\text{incorrect}) \). When viewing the entire table, however, it is clear that small samples perform well at high thresholds, and large samples perform well at low thresholds. For example, at \( n = 32, \Delta_{\text{population}} = .1 \), the simulation results for a threshold of .1 instead of .4 are 6,169 correct choices and 2,149 incorrect choices—these yield a higher \( f(\text{correct}) - f(\text{incorrect}) \) than is found anywhere in the data for that population contingency. At higher population contingencies, \( f(\text{correct}) - f(\text{incorrect}) \) is even higher; for \( n = 32, \Delta_{\text{population}} = .4 \), and a threshold of .1, the simulation makes 9,822 correct choices and only 56 incorrect choices, yielding \( f(\text{correct}) - f(\text{incorrect}) = 9,766 \)—this is compared with just 3,723 correct and 0 incorrect choices for a threshold of .4, the threshold Fiedler and Kareev suggested. Most importantly, if we consider an example decision maker who desires a 70% chance of correctness, \( f(\text{correct}) - f(\text{incorrect}) \) for her smaller samples is meaningless—because in the majority of times she will continue to sample. Overall she is likely to perform as if at \( n = 32, t = .1 \), because in the worst case, she can always sample up to 32 and lower her threshold down to .1—and, as we mentioned, \( f(\text{correct}) - f(\text{incorrect}) \) is higher for these values than anywhere else in the data, and this holds for all population contingencies reported.

In sum, it is evident that an optimal model of contingency based choice should employ a threshold that lowers as sample size increases; forcing the model to stick to a higher threshold for both small and large samples will result in it making much fewer choices for large samples, when those omitted choices could easily be made with a high probability of accuracy. Clearly, the optimal solution is to take advantage of both high and low thresholds, using the former for small and the latter for large samples.

Of course, although evolution often directs human strategies toward the most useful, there is no guarantee of this, so despite the dynamic threshold being superior to the fixed threshold, empirical evidence is needed to determine which humans use. There exists just this kind of evidence: Griffin and Tversky (1992) showed that although confidence relies more on evidence strength than sample size, sample size is still a factor. They asked participants to judge the likelihood that a spinning coin will fall heads when it is known the coin is biased at 3/5 for one side but unknown which side. Griffin and Tversky thus suggested that “people focus on the strength of the evidence—at least, as they perceive it—and then make some adjustment in response to its [sample size]” (p. 413). This dual reliance on both strength and sample size clearly points to a threshold that lowers as sample size increases. The regression results indicate that if sample size is low, then strength should be high to obtain high confidence (i.e., to make a choice); meanwhile, if the sample is large, then strength can be lower.

Note that the only difference between Griffin and Tversky’s (1992) scenario and Fiedler and Kareev’s (2006) scenario is that in the former, strength is determined by a simple proportion, whereas in the latter, it is determined by the difference between two proportions (the contingency). We assume, then, that there should be very little difference in behavior between the two scenarios (although future work could clarify this matter). Therefore, we propose that, even in Fiedler and Kareev’s case, one’s confidence should also always be based on both contingency (strength) and sample size.

Let us imagine the “adjustment for sample size” that occurs, but let us do so in a contingency scenario. One should be somewhat confident in a .3 contingency, for example, but this confidence then becomes adjusted on the basis of the size of the sample; the larger the sample, the more the confidence is boosted. Thus, one may be willing to make a choice when seeing a large sample with a contingency of .3 but not when seeing a small sample with the same contingency—one’s threshold is lower for large samples. To give a more detailed example, imagine seeing samples of size two for each candidate; say Candidate A is 50% positive (one smiley and one frownie face), and Candidate B is 100% positive (two smiley faces). In this case, the absolute value of the contingency is 1.5—1! = .5, but most people would probably require a higher contingency to reach threshold and make a choice (if willing to choose at all at sample size two)—that is, their choice threshold is higher than .5 for the sample size of two. Now imagine seeing samples of size 30 for each, with the same proportions—50% positive for A, and 100% positive for B. The contingency is exactly the same (.5), but this time one might be more likely to choose B—if so, his or her threshold for this sample size is at or below .5, making it lower in this large sample scenario than for the small sample scenario. Thus, the kind of “adjustment for sample size” that Griffin and Tversky (1992) found evidence for is precisely the lowering of one’s threshold as sample size grows, even in a contingency scenario.

We should point out, of course, that Griffin and Tversky (1992) clearly argued that evidence strength exerts a stronger influence on confidence than sample size. More specifically, one of Griffin and Tversky’s main points is that decisions, while adjusted for sample size, are not adjusted very much, resulting in overconfidence with small samples and underconfidence with large samples. It is then still theoretically possible that the threshold may be high enough for small samples to have a large \( f(\text{correct}) - f(\text{incorrect}) \), and then lower slightly but not enough for large samples to have a similar or greater \( f(\text{correct}) - f(\text{incorrect}) \). However, we hope our earlier discussion of Fiedler and Kareev’s (2006) evidence is convincing; certainly, people do not appear to gain any advantage from smaller samples. Future research should determine the extent to which thresholds are lowered in sampling situations; however, given the evidence, it is already clear that they will be lowered enough to negate any small sample accuracy advantage that might occur.

To summarize, Fiedler and Kareev (2006) discovered a very interesting situation: Because small samples tend to produce extreme contingencies, an algorithm designed to make choices based on contingencies alone (and ignoring sample size) will do best, when its threshold is high, taking small rather than large samples. However, we have shown that for organisms able to make use of sample size, determining choice threshold with the contingency

---

5 Griffin and Tversky (1992) did not explicitly claim significance for the factor of weight in their article, although their later conclusions imply it. Our efforts to clarify this matter with Griffin were unsuccessful.
alone is a suboptimal strategy. Furthermore, Griffin and Tversky (1992) have presented data that clearly imply that humans in fact make at least partial use of the stronger strategy. Although a dynamic-threshold model by itself would not completely eliminate the possibility of a small sample accuracy advantage, our earlier discussion of the evidence against a small sample accuracy advantage should be sufficient to convince the reader that it does not exist.

Conclusions

Fiedler and Kareev (2006) set out to “[examine] the rather provocative and counterintuitive hypothesis that the quality of decisions may, under specifiable conditions, decrease with increasing amount of information” (p. 899), independent of extraneous factors such as fatigue or opportunity costs. We have examined in detail both the empirical data they presented and the decision model that inspired their research. Overall, we found that their evidence does not support this hypothesis in human behavior and is instead more consistent with large sample superiority, perhaps tempered by fatigue in extreme conditions. We have additionally discussed why their model, though providing a fascinating thought experiment, is not the optimal model for organisms, like humans, that are able to keep track of sample size. Finally, we have cited evidence from Griffin and Tversky (1992) that suggests that humans appear to adopt a dynamic (lowering) threshold, and Fiedler and Kareev’s evidence from Experiment 2 reveals that because people are more accurate with large samples, their thresholds must lower enough to negate any small sample accuracy advantage.

Overall, we agree with Nick Chater when he said, “A rational agent with more information can always expect to do at least as well . . . as an agent with less information; because the agent with more information can, in the worst-case scenario, throw that information away” (Juslin et al., 2006, p. 116). We are not suggesting that small samples should be avoided at all costs—on the contrary, in situations in which efficiency is important, or fatigue or overload may set in, sampling quickly certainly has value. Even when ignoring those parameters, small samples fare surprisingly well thanks to the diminishing returns of larger samples (Hertwig & Pleskac, 2008). However, in the end, it is clear from the evidence discussed that unless fatigue is an issue, humans simply are more accurate with large samples.

References


Received October 9, 2008
Revision received June 8, 2010
Accepted June 16, 2010