האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

DOES DECISION QUALITY (ALWAYS) INCREASE WITH THE SIZE OF INFORMATION SAMPLES? SOME VICISSITUDES IN APPLYING THE LAW OF LARGE NUMBERS

by

KLAUS FIEDLER and YAAKOV KAREEV

Discussion Paper # 347 January 2004

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

Does Decision Quality (Always) Increase with the Size of Information Samples?

Some Vicissitudes in Applying the Law of Large Numbers

Klaus Fiedler and Yaakov Kareev

(University of Heidelberg) (The Hebrew University, Jerusalem)

Running Head: Sample size and decision quality

Author Note: The research reported in this article was supported by grants provided from the Deutsche Forschungsgemeinschaft to Klaus Fiedler and by the Israel Science Foundation, grant 712-1998 to Yaakov Kareev. Helpful comments by Matthias Bluemke on a draft of this paper are gratefully acknowledged. Correspondence should be addressed to Klaus Fiedler, Psychology Department, University of Heidelberg, Hauptstrasse 47-51, 69117 Heidelberg, Email: kf@psychologie.uni-heidelberg.de, or kareev@vms.huji.ac.il.

Abstract

Adaptive decision-making requires that environmental contingencies between decision options and their relative advantages and disadvantages be assessed accurately and quickly. The research presented in this article addresses the challenging notion that contingencies may be more visible from small than large samples of observations. An algorithmic account for such a "less-is-more" effect is offered within a threshold-based decision framework. Accordingly, a choice between a pair of options is only made when the contingency in the sample that describes the relative utility of the two options exceeds a critical threshold. Small samples – due to their instability and the high dispersion of their sampling distribution – facilitate the generation of above-threshold contingencies. Across a broad range of parameter values, the resulting small-sample advantage in terms of hits is stronger than their disadvantage in terms of false alarms. Computer simulations and experimental findings support the predictions derived from the threshold model. In general, the relative advantage of small samples is most apparent when information loss is low, when decision thresholds are high, and when ecological contingencies are weak to moderate.

Does Decision Quality (Always) Increase with the Size of Information Samples? Some Vicissitudes in Applying the Law of Large Numbers

In cognitive psychology and decision-making research, amount of information is considered to be a function that cannot take negative values. Apparently, adding information can only increase its overall value, implying a monotonically increasing function. Psychologists implicitly agree with economists that the utility of information cannot decrease, just as the value of other resources (e.g., money, property, power, legal rights) can never decrease when its amount increases. To be sure, the function that relates usefulness or total value of information to the amount of evidence may be negatively accelerated, and level off asymptotically; however, the slope of the function is assumed to remain non-negative.

For example, consider the typical form of a learning curve (see Figure 1). Performance increases monotonically with the number of learning trials, though the increment gets smaller as learning approaches an asymptote. The reason for concavity is easily understood. The increment that each trial adds is maximal at the beginning when there is little prior learning, when each trial is likely to contribute something new; however, as learning proceeds, the likelihood increases that further trials only reiterate what is already known. The concavity and the constantly positive slope would thus appear to be quite natural. Note that such a function is characteristic of information acquisition in general, and is not peculiar to human or animal learning. A similar function results when the overall reliability of a test is depicted as a function of test length, or number of items included (Kuder & Richardson, 1937). The same rationale applies to the accuracy of judgments or decisions as a function of the number of independent judges (Rosenthal, 1987). Increasing the number of test items or judges leads to the canceling out of errors, and a reduction in error with a corresponding increase in the proportion of systematic variance in the aggregate test score or judgment. In general, the function relating accuracy of predictions to sample size often resembles that of Figure 1.

Indeed, the logic underlying Bernoulli's law of large numbers appears so natural that we can hardly imagine that the monotonicity of information value might be questioned. Thus, if some researcher came up with data showing better learning following k trials in one experimental condition than following 2k trials in another, one would hardly postulate a new learning curve but try instead to find out what was wrong with these data: The two groups may differ in talent or motivation, or the researcher may have fallen prey to some artifact or uncontrolled factor such as fatigue arising from extended learning. Similar suspicion would arise if a psychometrician were to claim that a score based on a random subset of test items was more valid than a score based on all items, or if teachers were reported to discriminate more accurately between smart and poor students when they based their appraisal on a small rather than a large number of observations.

Can Less be More?

However, is the possibility that less information can be worth more or that decisions based on small samples can be superior to those based on larger ones really that far-fetched, or incompatible with logical and axiomatic underpinnings? Indeed, there is compelling anecdotal and experimental evidence that knowing or thinking too much can be of disadvantage. Just as the penalist in soccer should not dwell too long on where to kick the ball, consumers are typically most satisfied with their product choices when they make quick gut decisions based on minimal information. Additional product information can turn decisions into a torture and increase the likelihood of post-decision dissonance (Festinger, 1957). Or, to give another example, for an electronic search, an output of, say, 50 references is more informative and useful than one of 5000. Relevant research to back-up such anecdotal experience includes Wilson and Schooler's (1991) findings of impaired decisions with extended thinking, Forest and Feldman (2000) demonstration that lie detection performance decreases with amount of reflection, or increasing conflicts experienced when judgments are based on a large rather than small amount of evidence (Fiedler, Semin, Finkenauer & Berkel,

1995; Sande, Goethals & Radloff, 1988). Ambady and Rosenthal's (1992) work on the amazing diagnosticity of "thin slices" of behavioral evidence is also relevant in this respect.

Being confronted with such examples, one may still be inclined to dismiss the evidence regarding the benefits of scarce information as hardly bearing on the axiomatic view depicted at the outset. One may dismiss the performance impairment sometimes observed with high amounts of information as "just" due to cognitive overload, the inadequate administration and organization of resource-limited cognitive procedures applied to large arrays of information. Or one might reason that the examples reflect cases in which implicit, unconscious decisions are more appropriate than explicit, reflected decisions (Wilson, Dunn, Kraft, & Lisle, 1989).

However, again, the spontaneous tendency to discount such counter-intuitive findings or reconcile them with the axiomatic view may not be warranted. Recent research on adaptive cognition has revealed a number of less-is-more effects (Borges, Goldstein, Ortman, & Gigerenzer, 1999; Gigerenzer & Goldstein, 1996; Hertwig & Todd, 2003; Krauss & Wang, 2003; Martignon & Hoffrage, 1999) that cannot be reduced to inefficient cognitive procedures, as they also occur in computer simulations based on a fully rational, unbiased program. Another noteworthy phenomenon is evident in Elman's (1993) demonstration of the "importance of starting small". Using computer simulations of artificial language learning, Elman observed superior learning of a complex language system when the window size of the input utterances, or the short-term memory window of the learning individual, was limited rather than expanded in size in the early stage of learning. Allowing for more information in early learning impaired subsequent learning. Similarly intriguing evidence for less-is-more effects in language acquisition has been reported by Newport (1988, 1990).

Sample Size and Contingency Assessment

Our prime example, which is in the focus of the present research, refers to contingency assessment. Detecting environmental contingencies between signals and significant events, between causes and effects, or between situations and behaviors is a central module of

adaptive intelligence. As Kareev (1995, 2000; Kareev, Lieberman, & Lev, 1997) has shown, because of the skewness of the sampling distribution of correlations, and the fact that the skew is more pronounced the smaller the sample, the likelihood of detecting a contingency may, under certain conditions, be higher when the sample of observations is small rather than large – a phenomenon that we will elaborate on shortly. Again, one would assign almost axiomatic status to the assumption that accurate assessment of correlations increases with sample size. After all, sample size corresponds to number of learning trials, and additional trials can hardly reduce learning (cf. Figure 1). Imagine an animal exposed to the contingency between territories and danger. Choosing the less dangerous territory is crucial for survival in the long run. Is it not obvious that the probability of correct choices increases with the number of opportunities to learn about the environment? How could the animal ever be expected to reach a better decision with a smaller, rather than a larger sample of experience?

An intriguing answer to this paradox can be derived from a statistical principle that governs all probabilistic ecologies. When repeated samples of a certain size are drawn from a universe in which two variables (e.g., territories and danger) are actually correlated, the resulting distribution of sample correlations r around the "true" value ρ is not symmetrical but skewed. Most sample correlations are higher than the correlation in the universe. Thus, one of nature's nice features is that empirical snapshots often amplify actual effects. Even more amazing, this tendency is more pronounced for small samples. Small samples are thus particularly likely to accentuate and to expose the contingencies that actually exist in the world (Kareev, 1995). Furthermore, Kareev (2000) has shown that for binary variables this property of the empirical world has its maximum at a sample size of 7 ± 2. The interpretation which suggests itself is that evolution may have tailored human working memory span to cover exactly this maximally sensitive "window size" (cf. Kareev, 2000).

Regardless of the viability of such an evolutionary account (cf. Lewontin, 1979), the phenomenon suggests a serious violation of information monotonicity. Thus, an animal

exposed to a smaller number of learning trials may be more likely to choose the safer of two territories than one exposed to a larger sample of trials. This paradox should generalize across many biologically important decisions. Let us quickly add that the precise conditions under which this striking implication holds still have to be specified. But the possibility of a less-is-more effect in contingency assessment should be apparent, as supported by better contingency assessment performance in individuals with low rather than high memory capacity (Kareev et al., 1997). Note also that this less-is-more effect does not confound amount of information with complexity, memory organization, or the incomparability of cognitive procedures.

Critical Appraisal Within a Cognitive-Ecological Framework

Indeed, it should be realized that the small-sample advantage is logically independent of memory constraints because it originates in a statistical sampling process, which takes place in the environment, outside the individual's brain, prior to any cognitive processes. Kareev's main point is thus an ecological rather than a cognitive-psychologcal one. It suggests that any probabilistic ecology will generate samples that tend to overestimate existing contingencies. In contrast, whereas Kareev's framework is basically *ecological*, the monotonic learning curve in Figure 1 refers to *cognitive* laws of learning and memory. Critique and misunderstandings of less-is-more phenomena may arise from the failure to notice this distinction. In the remainder of this article, we develop an integrative, *cognitive-ecological* framework, taking into account both cognitive and ecological processes, within which both advantages and disadvantages of small samples can be reconsidered.

As outlined in Figure 2, the entire process of assessing an actually existing correlation or contingency can be decomposed into two stages, the ecological sampling process and the cognitive assessment and decision process. For convenience, consider the simplest case of a 2 x 2 contingency resulting from two dichotomous variables. For a sensible problem context, let us take the perspective of a teacher whose task is to figure out the performance differences that exist between students. The contingencies depicted in Figure 2 refer to different rates of

correct (+) and incorrect (–) responses given by two students, A and B. Let us further assume that the marginal proportion of correct responses in the population – i.e., the actual ability level – is rather high (e.g., a/(a+b) = .75) for Student A but rather low for Student B (e.g., c/(c+d) = .25). The actually existing contingency can thus be calculated as Δ = (a/(a+b) – c/(c+d)) = .75 – .25 = .50. The teacher's task calls for, first, gathering a sample of observations about the two students' performance; for simplicity, we assume random sampling. This sampling process (see upper part of Figure 2) can be defined as the transition from the latent population contingency Δ to the sample contingency Δ_{sample}. Note that, according to the above rationale, the superiority of A over B is more likely to be amplified in Δ_{sample} when the teachers' sample is small rather than large – an intriguing implication in its own right.

However, up to here, human memory has not come into play; the transition from Δ to Δ_{sample} reflects a statistical law applied to the probabilistic world. Replacing the human teacher by computer-based grading would not affect the logically antecendent transition from Δ to Δ_{sample}. Figure 2 (bottom) then represents the cognitive decision-making stage, which involves the transition of the sample contingency Δ_{sample} into the teacher's cognitive estimate Δ_{est}, as evident in differential judgments of Student A and B. There are various reasons why Δ_{est} can deviate from Δ_{sample}: Impaired perception, inhibited learning, memory overload, or competing cognitive processes and distracters. An informed analysis of the relation between sample size and contingency assessment requires that both stages, ecological sampling and cognitive decision making, be taken into account within a cognitive-ecological framework.

With respect to this analytical distinction between ecological sampling and cognitive processing, we can now locate two major critiques of Kareev's argument. The first objection comes from Juslin and Olsson (2004); it also pertains to the pre-cognitive, ecological transition stage, from the latent environment (i.e., population) to the sample. Juslin and Olsson claim that the seeming advantage of small over large samples is more apparent than

real. The "normal" advantage of large samples is evident as soon as the analysis is not confined to *hits* (i.e., detecting actually existing contingencies) but also takes into account *false alarms* (i.e., erroneously detecting non-existing contingencies). The second source of critique originates in empirical research on the cognitive stage of contingency learning (Fiedler, 1991, 1996; Fiedler, Russer & Gramm, 1993; Fiedler & Walther, 2003; Fiedler, Walther, Freytag & Plessner, 2002; Fiedler, Walther & Nickel, 1999), which provides evidence for improved contingency learning with a growing number of learning trials.

Taking Both Hits and False Alarms into Account.

Can it really be the case that small samples are more useful in detecting environmental laws than large samples? Or could Kareev's argument be flawed, reflecting only a one-sided, incomplete problem analysis? In response to this question – which is again reminiscent of the axiomatic status of a monotonic information function – Juslin and Olsson (2004) recently came up with the following argument. The small-sample advantage is only evident in the hit rate of correct sample-based decisions, but disappears when false alarms (i.e., incorrect inferences about environmental contingencies) are also taken into account. Although small samples may really exaggerate actually existing correlations, they also tend to produce illusory correlations that are not really there. After all, the skewed sampling distribution of r, from which the advantage of small samples emerges, presupposes that ρ is non-zero, that is, the analysis is confined to hits (i.e., the sample-based conclusion that a correlation exists when a non-zero ρ exists). However, in reality, when a sample contingency r is observed, the underlying ρ could be anything. Relative to the true ρ, a decision based on r that a correlation exists could represent a hit or a false alarm. Juslin and Olsson (2004) conducted computer simulations and concluded that if the underlying ρ is unknown so that decisions based on an observed r can represent hits as well as false alarms, the "normal" superiority of large samples reappears. Decisions were more accurate when based on large than on small random samples, at any level of ρ or any criterial value of r assumed to be required to make a decision.

However, although this critique seems compelling and logically sound, one should be cautious in discarding the less-is-more effect as due to an artifact or a one-sided emphasis on hits. One way to defend Kareev's standpoint would be to point out that specific problem contexts may render the benefits of hits more important than the costs of false alarms. In biological evolution, for instance, detecting dangerous predators or poisonous food is essential for survival, whereas the unnecessary effort involved in erroneously evading a harmless non-predator or avoiding an eatable piece of food would appear to be minimal. In general, to the extent that the costs or benefits of an environmental stimulus are higher than the costs or benefits of taking the right or wrong actions (cf. Swets, Dawes & Monohan, 2000), maximizing hits is more adaptive than minimizing false alarms, thus reinstating the assets of small samples. At least, the functional value of sample size could be expected to vary with the problem context and with domain-specific weighting of costs and benefits.

However, the viability of the small-sample advantage can be defended not only with reference to costs and benefits in certain ecological niches, but even at a more general level. When making choices – in the elementary case, binary choices between two options – we are not concerned with a precise estimation task (e.g., to estimate the true proportion of success with two options) but with the choice of that option that is better, or at least equivalent to other options, or simply "good enough". A "false alarm" or erroneous decision would not be one in which the r (between options and observed success) overestimates the true contingency ρ that has the same sign. Even when both options are equivalent (i.e., $\rho=0$), choosing one (i.e., assuming $\rho\neq0$) should not be considered wrong, but satisficing (Simon, 1956; Gigerenzer, 2001), given that one decision has to be made anyway. Whenever r deviates from ρ without reversing its sign, the resulting behavioral choice could still be considered a success within such a bounded-rationality approach. Only when the observed r reverses the sign of the true value could one talk of a genuine false-alarm decision. However, such reversals are quite unlikely even with small samples. Assuming $\rho = .50$, that is, assuming, for example, success

rates of .75 versus .25 for the two options, even a small sample of seven to ten items is rather unlikely to yield a reversal.

Within such a behavioral-choice framework, the buffering of false alarms becomes even more effective when we introduce the assumption that organisms do not make choices, or take action, at any moment in time, but only when the sampled evidence favors one option clearly enough. Thus, only when the absolute sample contingency r exceeds some critical threshold c will the organism choose Option A (if $r > +c$) or Option B (if $r < -c$). As long as the sampled evidence does not clearly suggest either action, the organism may continue sampling or draw a new sample. Given this safeguard of a decision threshold c, which has to be substantially different from zero to be psychologically effective, the definition of a false alarm becomes even more conservative, making it rather unlikely that even a small sample correlation exceeds $|c|$ in a direction opposite to that of ρ.

The diagram in Figure 3 demonstrates that such a framework predicts – under specifiable boundary conditions – that the advantage of small, unstable samples in terms of hit rates can be stronger than their vulnerability to false alarms. For small samples, the deviations of observed correlations r from ρ tend to be generally larger than for large samples. However, the difference is asymmetric; the hit rate advantage ($r > +c$; right side) for small samples is larger than the false alarm disadvantage ($\underline{r} < -c$; left side). Thus, granting $\rho > 0$ so that the distribution of r is not symmetrical around 0 but displaced toward the right, the very instability of small samples, results in more gains (hits) than losses (false alarms).

It is important to recognize one crucial difference between the present model and the similar model that led Juslin and Olsson (2004) to conclude that the posterior probability of correctly detecting environmental contingencies increases with sample size under all reasonable conditions. A major source of divergence originates in different definitions of hits and false alarms. Assuming a positive correlation, $\rho > 0$, Juslin and Olsson code an obtained r a hit if both $r > +c$ while $\rho > +c$, and a false alarm if $r > +c$ while $\rho < +c$. Thus, whenever the

population correlation ρ underlying a supra-threshold *r* is lower than the criterion c+, the outcome is considered a false alarm, even when both ρ and *r* are positive so that a correct choice is made. Using such a definition, which requires quantitative precision in addition to the correct sign of a correlation, small samples will of course produce many false alarms.

Within the present framework, in contrast, a false alarm is defined in terms of a wrong behavioral decision, that is, a false alarm is only coded if $r > +c$ although $\rho < 0$ (and if $r < -c$ although $\rho > 0$), whereas a hit means that $r > +c$ while $\rho > 0$ (or that $r < -c$ while $\rho < 0$). For any choice between two options, choosing the better option is a correct decision (a hit) and choosing the worse option is an incorrect decision (a false alarm). It is only the correct sign of the contingency that is relevant, regardless of the degree of over- or underestimation. In other words, the present approach pertains to correct *choices* rather than accurate *estimations*.

For example, if $c = +.5$ and $\rho = +.3$ or $+.4$, there is no reason to consider a choice based on an observed sample correlation of $r = +.6$ or $.+7$ a false alarm. After all, the sample has informed a correct choice. The assumption here is that consequences (i.e., benefits and costs) of decisions are only determined by whether the correct behavioral choice has been made.

Note that the above notion of a "correct choice" presupposes only two decision options, A and B, such that a positive contingency $\rho > 0$ implies that A is the correct choice. If there were more decision options, A, B, C, ... K, an optimal choice (of the best option) would require more than correctly assessing a 2 x 2 contingency. In that case, one would have to consider all pairwise contingencies, which is tantamount to determining an optimal ordering of all decision options. As it is unlikely that all pairwise comparisons exceed the threshold *c* at the same time, the present heuristic can be hardly used to find an optimizing solution for a large number of options. Rather, the heuristic is suitable for satisficing choices (Gigerenzer, 2001; Simon, 1956), that is, to decide whether a focused option A is at least as good as some comparison standard B, regardless of whether an even better option exists. Whenever decision

makers can resort to satisficing rather than optimizing strategies, particularly when choice problems are decomposed into sequences of binary choices, there is a reasonable chance of obtaining a small-sample advantage (see also Anderson & Doherty, 2004).

Taking Cognitive Processes of Learning and Memory Into Account

Having shown that small samples can, under specific assumptions, inform better decisions than large samples – in spite of "axiomatic feelings" and Juslin and Olsson's (2004) opposite conclusion – we now turn to the second source of criticism. Up to now, we have only been concerned with environmental sampling outside the human mind. The crucial question, however, is whether any advantage entailed in a small input sample will survive the cognitive learning and memory processes taking place inside the decision maker. To be sure, it should be clear that whenever decision makers use perfect actuarial strategies, registering data samples with perfect reliability and drawing a decision based on the calculated *r* in the sample, the model in Figure 3 applies, specifying conditions under which a small-sample advantage holds. However, in real life such an actuarial strategy may be unobtainable. Rather than registering all input observations objectively, individuals often have to learn and memorize information implicitly (Seger, 1994), whereby observations may be ill-defined, distributed over extended time, mixed-up with irrelevant information, and therefore hard to discern from background noise. In such a task context, when many stimuli compete with each other for attention and cognitive resources, small samples may be more vulnerable, and be particularly likely to be overlooked or overridden by other, more extended samples. It is hard to see why contingency learning should not obey normal learning functions, which increase monotonically with number of observations, or trials (Figure 1).

Supportive evidence for this contention comes from a series of experiments on contingency learning in various social cognition paradigms (for an overview, see Fiedler & Walther, 2003). Whenever the task is to figure out the proportion of positive outcomes (success, desirable behavior, etc.) associated with different target persons or groups,

performance increases monotonically with increasing numbers of observations. Pertinent evidence comes, in particular, from a series of experiments in a simulated school class setting (Fiedler, Walther, et al., 2002). A classroom with sixteen students was represented on the computer screen and participants played the role of a teacher who could observe the students' achievement (i.e., their correct and wrong answers to questions in different lessons) across an extended period of time. The computer program that guided the experiment determined actual parameters of correct responses for each student. Moreover, students differed in their motivation parameter, as reflected in the frequency of raising hands, thus producing variation in the size of samples available for each individual student. For each pair of students, then, a contingency could be computed between students (A vs. B) and success (+ vs. –), exactly as in Figure 2. For instance, when the ability parameters for A and B were .8 and .2, the resulting contingency was $\Delta = .80 - .20 = .60$. At the beginning, of course, differences between students, or pairwise contingencies, were completely unknown to teachers, who only gradually learned the achievement differences as the number of observations increased. At the end, teachers' average judgments reflected these differences quite accurately. But still, existing contingencies were manifested more clearly for those student pairs for which teachers had gathered large rather than small samples. Thus, when learning and memory are taken into account, contingency assessment seems to obey the same monotonic laws as all learning.

Note, however, that these findings from the simulated school-class differ in one crucial respect from the binary-choice situation of Figure 3. The teachers' judgment task is basically an *estimation* task rather than a *choice* task. Teachers estimate the proportion of correct answers they have received from different students, and the greater accuracy observed with larger samples refers to deviations of estimates averaged across teachers. There is little doubt that large samples lead to more accurate estimates than small samples, as evident in the lesser dispersion of the large-sample distribution in Figure 3. For a direct test of the possibility of a

small-sample advantage, one would have to analyze the teacher's pairwise choices between students in task contexts that call for a marked decision threshold.

Related to the distinction between estimation and choice is Hogarth and Einhorn's (1992) distinction between continuous updating and end-of-sequence judgments. The teachers' evaluation task involves continuous updating over time. Information sampling never comes to an end, rendering it quite unlikely that a finite sample is ever compared to a threshold. Continuous-updating is easily interpreted as an estimation task but hardly ever involves discrete choices. According to the model in Figure 3, any advantage of small samples should be more apparent in task settings that call for end-of-sequence judgments or choices based on the comparison of finite samples with a threshold or standard (e.g., deciding whether Student A is better than B, or exceeds some graduation threshold).

Evidence From Computer Simulations and Pertinent Experiments

In any case, it would appear to be an open empirical question, rather than a logical impossibility whether, under specific conditions, an advantage of small samples may survive the cognitive decision making process. The remainder of this paper is devoted to exploring this question. First, a straightforward computer simulation of the sampling stage is presented, based on the aforementioned definitions of hits and false alarms in a satisficing model. In a second results section, computer simulations of the cognitive decision-making stage are reported, using a simulation model (Fiedler, 1996) that orignally served to demonstrate the superiority of large samples in associative learning. The remainder of the article is then devoted to a comparison of the major simulation results with empirical evidence from several experiments involving binary choice, in which sample size was manipulated.

Simulation of Sampling Stage

Although the small-sample advantage depicted in Figure 3 exists on purely logical grounds – under auspicious conditions – the following computer simulation is quite informative about the quantitative degree of the less-is-more effect and its parametric

boundary conditions. The simulation approach was simple and straightforward. The model was confined to dichotomous variables. We used Δ, rather than ρ, because Δ is the normative measure of contingency between binary variables, as well as less vulnerable to small cell frequencies and Δ provides a natural model of a choice task, based on the comparison of the success rate of two options A and B (i.e., two conditional probabilities). The population correlation was represented as a 2 x 2 table containing 1000 observations. For instance, to represent Δ = .20, the population would include a = 300 positive outcomes for option A, b = 200 negative outcomes for A, c = 200 positive outcomes for B, and d = 300 negative outcomes for B, yielding $\Delta = a/(a+b) - c/(c+d) = 300/(300+200) - 200/(200+300) = .6 - .4 = .2$. To represent Δ = .10, the respective distribution of cases in the total population would include a = 275, b = 225, c = 225, and d = 275. For simplicity, and similarity with Juslin and Olsson's (2004), we used only equal marginal distributions, such that $(a+b) = (c+d)$, and symmetrical correlations, such that $(a-b) = (d-c)$. In this case, Δ is identical to ρ. For asymmetrical cases, ρ is the geometric mean of the two measures Δ and Δ' that result when either the two row proportions or the two column proportions are compared, respectively.

The sampling simulation involved drawing repeated random samples (with replacement) of a given size n from the N = 1000 cases comprising each population (i.e., each level of Δ) and assessing the sampling distribution of the resulting estimates Δ_{sample} across a total of 10000 trials. Five contingency levels were included (Δ = .1 vs. .2 vs. .3 vs. .4 vs. .5); samples size varied from n = 4 to 8 to 16 to 24 to 32. For convenience, the sign of the actual contingency was always positive so that hits were defined as observed contingencies that exceed the positive threshold ($\Delta_{sample} > +c$), whereas false alarms were defined as $\Delta_{sample} < -c$. The range of possible contingencies (–1.0 to +1.0) was subdivided into 21 categories (–1.0 to –.9; –.9 to –.8, ..., up to +.9 to +1). The boundaries of these categories can then be used to analyze the impact of different decision thresholds c, by considering the cumulative frequency

of cases falling in categories that are as extreme as or more extreme than *c*. The hit rate is the summed frequency of all categories > +*c*, whereas the false-alarm rate is the summed frequency of all categories < –*c*. Because the total number of simulated cases is 10000, the hit and false alarm rates in Table 1 can be interpreted on an "out of ten thousand" scale. The table covers a subset of data for three contingency levels, $\Delta = .1$, .2, and .4.

In general, across all levels of Δ, *c*, and *n*, hit rates are higher than false-alarm rates, and, at all *c* and Δ levels, both hits and false alarms are more frequent for small than for large samples, reflecting the higher dispersion of the sampling distribution. More importantly, Table 1 reveals that when the decision threshold *c* is higher than Δ, the advantage of small samples in terms of hits is typically larger than the disadvantage in terms of false alarms, as anticipated analytically in Figure 3. For instance, given $\Delta = .1$, at criterion levels higher than $|c| = .4$ (i.e., above the dotted line), the difference between hit and false alarm rates (i.e., between adjacent columns) *decreases* from left to right with increasing sample size.

To render this small-sample advantage more visible, the graphs in Figure 4 show the differential frequencies of hits minus false alarms of all sampled contingencies exceeding particular levels of *c*. Different charts are included for different Δ levels. In fact, the preponderance of decreasing curves indicates that over a wide range of the parameter space, the difference of hits minus false alarms decreases with increasing sample size – if the threshold *c* is strong enough. The small-sample advantage also depends on Δ, showing a maximum at intermediate Δ levels. At the upper end of contingency strength ($\Delta = .5$) the curves become flatter, but accuracy is rather high anyway, regardless of sample size. At the lower end ($\Delta = .1$), the paucity of systematic variance weakens small samples more than large samples; a small-sample advantage is only visible for extremely high values of *c*. Note that the lines hardly ever increase (i.e., hardly ever exhibit an advantage of the larger samples). We deliberately exclude the $\Delta = 0$ case, because if $\Delta = 0$ any decision would be equally

"correct" or satisficing. Thus, although given $\Delta = 0$ small samples lead to more $\Delta_{sample} > c$ cases than large samples, those "alarms" need not be considered wrong decisions. In any case, computer simulations demonstrate quite impressively that over a wide area of the parameter space small samples outperform large samples, and are hardly ever outperformed, when decision processes follow the assumptions of the threshold model depicted in Figure 3.

Simulation of Sample-Based Decisions

Let us now turn to the issue of whether any small-sample advantage can survive the cognitive process stage. Computer simulation can again answer the basic question of whether algorithms exist at all that produce better decisions from small than from large samples. However, while the simulation of drawing a sample from a universe was straightforward, simulating the cognitive process in the decision maker would appear to be more difficult. It is nevertheless quite easy to simulate a mental operation that corresponds to the Δ interpretation of a contingency: Scan the data base for all information relevant to a decision option A and sum over all associated evaluations. Then repeat the same operation for option B. The difference between the two summed evaluations affords a measure of Δ.

Exactly such an evaluative comparison underlies a simple connectionist feedforward model called BIAS (Fiedler, 1996), which has been shown to account quite well for the cognitive process of contingency assessment (Fiedler, 2000; Fiedler, Kemmelmeier & Freytag, 1999) and which was also used for the present study. The algorithm used for the BIAS simulation of contingency assessment is explained in Figure 5. All information is represented distributively, using 12-element vectors or patterns of binary features to denote variable levels. The ideal types on the left side of Figure 5 indicate the patterns denoting options A versus B, and positive (☺) versus negative (☹) evaluation. Each bivariate observation is a concatination of an option vector with an evaluation vector, however, degraded by a proportion of inverted vector elements reflecting noise. The first 12 columns of the matrix in Figure 4 represent observations that resemble the ideal combination of A (in the

upper segment) and ☺ (lower segment), though they are not identical to the ideal type. The respective numbers of items in each column block represent the distributions of $a = 12$, $b = 6$, $c = 6$, and $d = 12$ items in a 2 x 2 contingency table. The noise factor is introduced to acknowledge that some information is lost, either in the environment, or in memory.

Given such a representation of (degraded) stimulus information about an observed sample contingency, the simulation of a Δ-like algorithm proceeds as follows. For an overall evaluation of option A, the ideal type of A is used as a prompt and compared to all column vectors (in the upper segment). The degree of match, defined as the dot product between the prompt and all column vectors, is computed and each column is weighted (i.e., multiplied) with this dot product. This amounts to amplifying items which resemble A and reducing items unrelated or dissimilar to A; if a dot product is negative, the item weight is actually reversed. The weighted matrix is then summed across all columns and the evaluation segment of the resulting sum vector (i.e., the bottom segment) is correlated with the ideal type for positive evaluation. To the extent that the resulting correlation is positive, the evaluation of A can be assumed to be positive; a negative correlation with the ideal pattern of positivity indicates negative evaluation. The evaluation for option B is then computed in the same fashion, and the difference between the two resulting evaluative correlations for A and B provides the simulated measure of the contingency, quite analogous to Δ.

Using this algorithm, the simulation routine again started with a population defined by a 1000-cases 2 x 2 table, from which random samples of n cases were drawn. The resulting sample of $a + b + c + d = n$ observations for all combinations of A vs. B and ☺ vs. ☹ were translated into vectors composed of the corresponding ideals. A proportion of i vector elements were then reversed; the noise parameter was manipulated ($i = .1$ vs. .2 vs. .3) to simulate different degrees of noise in the environment. The algorithm described above was then applied to compute the subjectively experienced contingency Δ_{est} arising from the given sample drawn from an environment with a given Δ. This procedure was repeated 10000 times,

starting with new random vectors for the ideal types and different random productions of noise. Separate simulation series were run for different Δ values, samples sizes *n*, and noise parameters *i*. In each case, the 10000 simulated contingency judgments were then compared to varying decision thresholds *c*, to capture the rate of resulting hits and false alarms.

Figure 6 portrays the major results analogous to the sampling stage results in Figure 4. Again, it is evident that some of the curves exhibit a descending trend, indicating the possibility that under certain conditions the small sample advantage can even survive a cognitive process that underlies the constraints of monotonically increasing learning functions. To be sure, the residual small-sample advantage is clearly lower than for the environmental sampling stage. However, within reasonable confines, small samples still exert their assets. When the actual contingency is different from zero but not too high, when the noise parameter is low, and when the decision threshold *c* is substantial, small samples' tendency to increase the hit rate is stronger than their tendency to increase the rate of false alarms. Large samples, in contrast, demonstrate their superiority when *c* becomes lower and when the proportion *i* of noise in the system increases.

Discussion. This pattern, encountered in many related simulations, offers an account for the boundary conditions under which use of small samples has an advantage. Whenever behavioral choices are contingent on an evidence threshold that is high enough and when there is not too much information loss or unreliability in the data, then small samples are the equal of large samples, or even outperform them. This notable advantage of small samples is most visible for moderate contingency values between $\Delta = .1$ and $\Delta = .3$. However, it does not strongly depend on the value of Δ, as very high Δ values produce a ceiling effect (i.e., hardly any false alarms) and $\Delta = 0$ renders any decision satisficing. However, for lower decision thresholds (i.e., when a growing proportion of the entire distribution of outcomes leads to decisions) and for higher noise ratios (i.e., when the aggregation effect of large samples cancels out errors), the "normal" advantage of large samples is borne out inevitably.

It should be noted what these parameter boundaries mean, psychologically. High decision thresholds – the domain where small samples "excel" – can be expected to be applied to important decisions, that is, when significant consequences are at stake. Fortunately, the buffering function of a high threshold protects the organism under such problem conditions from too many false alarms and thereby warrants high success rates based on small samples. Another advantage may be that small samples increase the likelihood that action can be taken anyway, because the sampled evidence exceeds the threshold. Not being paralyzed in a passive, inactive state, may be of high functional value. However, the advantage of small samples for important decisions is lost when there is too much noise. When input data are invalid or unreliable, or likely to be lost in memory, this will be most detrimental for decisions based on small samples, whereas increasingly larger samples serve to filter out noise and to extract the systematic variance even under such conditions.

Thus, an organism that is about to profit from the small-sample advantage runs the danger of committing consequential mistakes in noisy environments. It is therefore of utmost importance for organisms to recognize the amount of noise in the input and to adjust sample size accordingly. Organisms that lack this sensitivity for the degree of noise will probably fail. Therefore, individual differences in decision-making ability may reflect to a larger degree performance with small samples than with large samples or, conversely, the superiority of good inductive decision makers should be manifested mainly on small-sample tasks.

Other inductive tasks

Before we turn to empirical results from decision-making experiments, it should be mentioned that the circumscribed small-sample advantage is not peculiar to the standard contingency paradigm alone, but is a general feature of inductive inferences in different paradigms, such as the detection of elementary proportions or conditional probabilities. For the purpose of the present article, let us briefly consider one other inductive task that might be called *competing-tendencies*. Competing-tendency tasks are structurally similar to the

assessment of contingencies, but distinct in lacking dimensional constraints. An implicit assumption in the 2 x 2 structure in Figure 2 is that the two levels of each dichotomous variable are mutually exclusive. An observation pertains either to Student A or to B; it is either correct or incorrect, such that the sum of all four frequencies in the 2 x 2 table amounts to the total number of observations. Observed outcomes (+ and –) refer exclusively to one student. Such exclusiveness constraints do not hold for many real-world tasks that call for a choice between competing but not mutually exclusive tendencies. For instance, with regard to professional choice, Student A may have two competing interests, biology and medicine. A counselor wants to find out which tendency is stronger. However, biology and medicine are not mutually exclusive; they overlap in contents. Any evidence that the student is interested in biology also provides some evidence for interest in medicine. Observations about the two tendencies are not additive, due to overlap. Many real-world decisions and choices fit the competing-tendency structure rather than the idealized case of strictly complementary dichotomies: What's the degree of interest in related professions? How much do I like different friends from the same group? How dangerous are various risks associated with the same environmental causes? What political goals are most important? Is the patient manic or depressive? Although competing tendencies in these decision problems do not exclude each other and mutually imply each other, they differ in strength and adaptive behavior calls for the assessment of the relative strength of competing tendencies.

Within this problem context – which may be much more common than it is prominent in statistics books – another advantage (and disadvantage) of small sample size becomes apparent. It can be shown that although large samples maximize the assessment of both (all) tendencies that are at work, small samples often differentiate better between strong and weak tendencies – which is crucial for making choices. This is because, over a wide part of the parameter space, insufficient learning due to paucity of observations hampers the assessment of weak tendencies more than the assessment of strong tendencies, which do not require large

samples. The ironic consequence is that small samples may allow for better discrimination than large samples, as illustrated in the difference between two learning curves – one easy to learn and one hard to learn – in Figure 7. Learning of the hard-to-discern tendency is retarded, that is, the lower curve starts to rise later, has a lower slope, and takes more trials to reach the asymptote. As one can easily see, the difference between the strong and the weak tendency is most pronounced for small and medium numbers of trials and diminishes gradually as the weak tendency is also understood after a large number of trials.

For a simple computer simulation, the BIAS model (Fiedler, 1996) was used once more. Each of two of competing tendencies, U and V, was represented by a 12-element binary vector, but the last six elements of the vector defining tendency U was identical to the first six elements defining tendency V. Consider the case when there are 5 observations for U and 3 observations for V. As already shown in Figure 5 above, BIAS adds a new column vector for each of the 8 observations to a stimulus matrix, in which the upper 12 rows represent the one (stronger) and the last 12 rows represent the other (weaker) tendency. However, due to the overlap, the total number of rows is not 24 (2 times 12) but only 18, because the middle six elements overlap. Thus, as a column vector for an U observation is appended in the upper segment, six of the 18 elements provide overlapping information about V in the lower segment, and vice versa. For each learning trial, again, we did not append a perfect copy of the U or V vector, but a noisy copy in which a proportion of i = .1, .2 or .3 elements had been inverted. To quantify learning strength, again, the vector defining a particular tendency was used as a prompt (i.e., each matrix column was weighted by its similarity (dot product) with the ideal tendency vector in the respective segment), and the resulting sum across all columns (in the segment of the tendency) was compared (i.e., correlated) with the ideal vector. To the extent that the resulting correlation approaches unity, the respective tendency has been learned perfectly. However, note that observations of one tendency will also influence the computation of the learning of the other tendency, due to overlap in the middle elements.

Figure 8 summarizes the major results, with reference to *differential learning* (i.e., the difference between learning the dominant and the inferior tendency). The dominance of one tendency over the other was varied from a ratio of 5:3 observations, respectively, to 6:2 to 7:1. The total number of observations (sample size) varied from n = 8 to 16 to 24 to 32 (multiples of 8 to conserve the above dominance ratios).

Figure 8 shows that *differential learning* decreases from small to large samples. Although learning of both tendencies increases slightly with sample size, the *dominance* of one tendency over the other decreases as learning proceeds, from 8 to 16 to 24 to 32 observations. This is evident in the number of simulation trials on which the learned difference between dominant tendency ("hits") and inferior tendency ("false alarms") exceeds various thresholds. As in the preceding simulations, the small sample advantage increases from lower to higher thresholds. It is also apparent that a rather low threshold of c = .2 is already sufficient to establish equality or even a small-sample advantage.

EXPERIMENTAL EVIDENCE FOR DECISIONS BASED ON VARYING NUMBERS OF OBSERVATIONS

Thus far, we have provided an algorithmic proof of existence for the seemingly paradoxical claim that, under specific conditions, small samples can lead to better decisions than larger ones. We have described a decision-threshold model from which the small-sample advantage can be deduced analytically, and have presented the results of simulation studies demonstrating that this curious advantage covers a considerable part of the relevant parameter space. Moreover, our simulations have shown that this phenomenon is not confined to the statistical sampling stage but can even survive the learning and memory processes in a human decision maker. Now the crucial question that suggests itself is whether this conclusion can be substantiated in real decision-making experiments. What empirical evidence is there for the

less-is-more effect for the impact of the parameters that were shown to moderate the advantages and disadvantages of small and large samples in computer simulations?

To be sure, an existence proof was already provided by Kareev (1995), showing that reduced information samples can in principle improve correlation assessment and subsequent decisions. In the present article, we report some preliminary evidence that expands the scarce evidence that exists on this intriguing topic (e.g., Kareev et al., 1997), and also explore some of the parameters or boundary conditions that simulations have shown to affect the optimal samples size. At the moment, we cannot provide truly comprehensive experimental evidence for the full parameter space used for the simulations – covering three levels of noise x six degrees of contingency x six samples sizes x multiple decision thresholds. Collecting such data would not only require a long-term research effort, but also presupposes ingenious means of manipulating the parameters directly or indirectly. However, in a few experiments we have already conducted we have included manipulations that can be clearly interpreted in terms of specific parameters. Altogether, these experiments corroborate the claim that small samples can be superior within the confines derived from the present model. Moreover, even under conditions in which performance with small samples were inferior, the differences were typically very modest. Thus any decision framework that also takes information costs – in terms of collecting it and deferred decisions – into account, in addition to correctness of decisions, will thus have a hard time to justify the investment in large samples.

Experiment 1

The task situation and stimulus materials used for this and the next two experiments were identical to the illustrative example used at the outset of this article. Experimental participants were presented with series of smilies and frownies, symbolizing positive and negative consumer reactions to a pair of products. Each trial was defined by four frequencies representing the respective numbers of smilies and frownies associated with two products to be compared, analogous to the 2 x 2 frame in Figure 2. Experiment 1 involved all six pairwise

comparisons that can be formed from a set of four products for which small samples were available (i.e., smily-frowny distributions of 6:2, 5:3, 4:4, 3:5). In addition, six pairs were formed from four other products characterized by the same ratios of smilies and frownies but twice as large samples (i.e., 12:4, 10:6, 8:8, 6:10). The contingency judgment task simply consisted of choosing the more preferred product from each pair. We were interested in whether preference decisions based on paired comparisons between small-sample products were more or less likely to be accurate than preference decisions based on larger samples representing exactly the same underlying contingencies.

Two different modes of stimulus presentation were used in two experimental groups, supposed to affect difficulty of encoding and two relevant model parameters. In the *successive-presentation* condition, all smilies and frownies pertaining to one target product of a pair were first presented, one at a time, on the left half screen (shaded in turquoise), before the smilies and frownies for the other product appeared on the right half screen (shaded in pink). In the *simultaneous-presentation* condition, the two sub-samples were intermingled, with smilies and frownies for both products (also projected on the left and right half screen) occurring in random alternation. This manipulation should affect the relative efficiency of small versus large samples for two reasons. First, simultaneously presented stimuli should be more difficult to encode and thus increase the degree of noise, as evident in reduced overall performance. However, apart from this main effect (or manipulation check), the higher noise level should create a relative advantage for large over small samples. Second, simultaneous presentation might force participants to draw decisions under higher uncertainty, that is, to admit a more lenient decision criterion, or threshold. Confidence and latency data should provide a check on the viability of this assumption. Both assumptions together, higher noise and lower confidence after simultaneous than after successive presentation, predict a relative advantage of large samples after simultaneous presentation and a relative advantage of small

samples after successive presentation of the two subsamples. The major theoretical prediction, therefore, was that of a presentation mode x sample size interaction.

Method

Participants and Design. Thirty male and female students of psychology and other subject matters participated at the University of Heidelberg either for payment or to fulfill a study requirement. They were randomly assigned to one of two experimental groups representing the mode of presentation factor (simultaneous vs. successive). The second design factor, sample size (large vs. small), was manipulated within participants.

Materials and Procedure. The entire experiment consisted of 22 contingency tasks, each describing several judges' positive or negative evaluations of two fictitious movies. For convenience, let the two movies be denoted A and B and positive (smilies) and negative reactions (frownies) be denoted + and –, respectively. Contingency assessment thus amounts to comparing the proportion of positive reactions to A and B; $\Delta = p(+/A) - p(+/B)$. As already mentioned, the computer screen was partitioned vertically into two halvess, shaded constantly in turquoise and pink. The names of the two target movies (i.e., meaningless alphanumerical labels) were presented at the top of the left and right half screens, and smilies and frownies appeared in randomly varying locations within the left (A) and right (B) half screen. The stimulus frequencies as well as the mode of presentation were manipulated. Immediately after all smilies and frownies for a given pair of movies had been presented, participants had to make a choice, using the left or right arrow key, as to which movie received the more favorable evaluation. Decisions latency and subjective confidence (indicated by moving the computer cursor on a 42-digit horizontal graphical rating scale) were also measured. No feedback was provided after each trial, but participants received a tabular feedback of their trial-by-trial performance at the end of the experiment, as part of the general debriefing. Minimal instructions used a cover story saying that consumer evaluations are often quickly conveyed by symbols and that the present study is concerned with the ergonomic value of two

quite natural symbols, smilies and frownies, thought to convey valence quite effectively. They were asked to work as carefully as possible, for the sake of the scientific investigation.

The contingency tasks were constructed by pairing two sets of four movies, each characterized by a specific distribution of smilies and frownies. For one set, small samples were available: 6+/2–, 5+/3–, 4+/4–, 3+/5–. For the other set, the same +/– proportions held, but samples were twice as large: 12+/4–, 10+/6–, 8+/8–, 6+/10–. The entire set of 22 trials included all six contingencies that could be formed from small samples (presented in positions 1, 6, 11, 15, 17, 19), six paired contingencies from large samples (in positions 3, 7, 14, 16, 18, 22), and the four contingencies relating small and large samples sharing an equal ratio (e.g., 6+/2– vs. 12+/4–). The remaining eight contingencies involved comparisons among small and large samples reflecting unequal but similar ratios (e.g., 6+/2– vs. 10+/6–). Only the first two sets of paired-comparisons are of interest for the present purpose.

Each trial started with the two half screens assuming their color and the names of the two movies being inserted at the top of the two half screens. All smilies and frownies pertaining to the same movie appeared at once. In the successive-presentation condition, the subsample for the left film was presented before the subsample for the other movie on the right. In the simultaneous-presentation condition, symbols pertaining to both movies (presented on different sides of the screen) appeared at the same time.

Results and Discussion

For a sensible and statistically quite powerful comparison of binary decisions based on small and large sample contingencies, we aggregated the performance across all six paired comparisons between small-sample targets and across all six paired comparisons between large-sample targets. The proportions of correct decisions (i.e., favoring the movie with the higher proportion of smilies) for comparisons of small and large sample contingencies, respectively, yielded two repeated measures to be analyzed in a two-factorial ANOVA, together with the between-participants factor, mode of presentation. Table 2 gives the relevant

means as a function of experimental conditions. Two significant effects emerged. A presentation mode main effect, $F(1,28) = 8.58$, $p < .01$, confirmed that simultaneous presentation causes more cognitive load and reduces overall performance. Pooling over sample sizes, the respective proportions of correct choices are .76 for successive and .53 for simultaneous presentation. This preliminary finding is consistent with the premise that simultaneous presentation serves to induce a higher degree of noise that should give a relative advantage to large samples. In contrast, successive presentation should serve to lower noise and should support a decision process that requires a binary choice between two coherent subsamples, presented one at a time, rather than a continuous updating of two parallel estimates. Consistent with this assumption, indeed, a significant disordinal interaction, $F(1,28) = 9.27$, $p < .01$, reflects better performance on small samples (.82) than large samples (.71) after successive presentation, but an advantage of large samples (.64) over small samples (.42) after simultaneous presentation. There was no sample-size main effect, $F(1,28) = 1.03$.

An analogous ANOVA on decision latencies corroborates the assumption of more uncertainty induced by simultaneous (M = 2.80) than successive presentation (M = 1.55), yielding a presentation mode main effect, $F(1,28) = 9.80$, $p < .01$. Latencies were also generally longer for small (2.37) than for large samples (1.98), $F(1,28) = 9.89$, $p < .01$. A sample size x presentation mode interaction, $F(1,28) = 4.64$, $p < .05$, reflects a stronger influence of presentation mode for small than for large samples (see Table 2).

The ANOVA of confidence ratings produced only one significant main effect – that for sample size, $F(1,28) = 7.88$, $p < .01$, indicating generally higher confidence when decisions were based on large (M = 25.29) than on small samples (M = 24.20), regardless of presentation mode. Although simultaneous presentation induced somewhat lower confidence (M = 24.08) than successive presentation (M = 25.40), the corresponding main effect fell short of significance, $F(1,28) = 2.79$, $p = .102$, suggesting that subjective confidence was not very sensitive to the between-participants manipulation of presentation mode.

Altogether these findings are consistent with the contention that small samples are not generally inferior to large samples. Small samples can even be superior under specific conditions, and the (relative) advantage of small samples can survive the cognitive laws of learning and memory. As predicted by the simulation results, small samples unfold their relative assets under conditions that keep information loss, or the noise factor in learning and memory, at a relatively low or moderate level. As noise increases and cognitive performance declines, the advantage of small samples is overridden by the aggregation advantage of large samples, which are needed for a difficult learning process.

The present preliminary study does not speak to several other parametric implications of our threshold-based decision framework. As to the Δ parameter, a within-participants comparison of different contingency levels would have lacked statistical power, given that each participant was only exposed to a single contingency at each level of Δ and sample size. Moreover, no evidence was obtained for the decision criterion *c*, which can be hardly manipulated in a straightforward fashion. To analyze the impact of *c* statistically, one would have to compare judgments based on different confidence levels, which again calls for a larger number of judgments within participants. In two other experiments, we therefore increased the number of data points by exposing each participant to 120 contingency judgment trials (40 at each level of Δ) in an attempt to base the empirical results on more solid ground and to learn more about the moderating role of the model parameters.

Experiment 2

The same basic task situation was used as in Experiment 1, except for the following notable changes. First, the stimulus series included 120 trials, 20 for each combination of sample size and Δ. Second, a generally higher sample size range appeared appropriate given that performance in (the simultaneous condition of) Experiment 1 was hardly above chance level and self-determined sample length under the present task conditions turned out to be larger than 20 observations (see Experiment 3). Sample size was therefore manipulated to

vary between two fixed levels, 16 versus 32 observations. Third, the 120 contingency tasks were not based on fixed 2 x 2 distributions, but each trial involved a new random sample drawn from three 1000-items populations representing true contingencies of $\Delta = .1$ or .2 or .4, respectively. Thus, unlike Experiment 1, the environmental sampling stage was not cut short, but the effective stimulus input was allowed to vary like true random samples of a given size.

Finally, four different presentation modes were used in different experimental groups, resulting from the combination of two binary factors, whether each presented smily or frowny was erased before the next symbol versus all smilies and frownies of a sample remained on screen, and whether the two subsamples (for the two products) appeared successively on the two sides of the screen versus simultaneously, with smilies and frownies from both subsamples appearing in randomly alternating fashion. Recall that in Experiment 1, the relative advantage of small samples had been confined to the successive presentation mode, which was interpreted in terms of reduced noise. If this interpretation is correct and small sample do not profit from a specific presentation mode *per se*, generally higher performance in Experiment 2 (due to larger samples eliminating noise) would speak against the contention that the small-sample advantage is peculiar to the specific case of successive presentation.

Method

Participants and Design. Ninety-two male and female students at the University of Heidelberg participated, either for payment or to meet a study requirement. They were randomly assigned to four experimental groups resulting from the manipulation of two between-participants factors, presentation time (symbols remain on screen vs. disappear) x order (successive vs. simultaneous). Two further factors were varied within participants, sample size (small = 16 vs. large = 32) and contingency level ($\Delta = .1$ vs. .2 vs. .4).

Materials and Procedure. The task situation and instructions were similar to those used in Experiment 1. Participants were told that each trial consists of smilies and frownies (reflecting positive vs. negative evaluations, respectively) associated with two products,

represented by two 5-digit strings on the two sides of the screen. The task was to make a decision as to which product is superior, in terms of the relative proportion of smilies in the population. Participants were explicitly told that stimulus samples were drawn from a larger population and that making a correct decision was defined in terms of a correct inference from the sample to the population. After each sample contingency was presented, participants could choose between three response options: Pressing the left arrow key on the keyboard to decide for the left product, pressing the right arrow key to decide for the right product, or pressing the downward key to indicate that no difference was discerned.

During a 2000 ms inter-trial interval the screen remained black. After the two sides of the screen assumed their color shading and the product labels were inserted at the top, similies and frownies appeared at a rate of 400 ms per symbol in random locations of the appropriate half screen. In the successive presentation-order condition, the left subsample was completed before the right subsample was started. In the simultaneous presentation-order condition, all items from both samples appeared in random order. In the symbol-erase condition, each smiley or frowny was erased immediately after its presentation period. In the symbol-remain condition, all previously presented symbols (on either side of the screen) remained on screen until the sample was complete. Then, however, all smilies and frownies were erased as the prompts for the three response options were inserted, in order to prevent participants from counting the symbols on the screen. If one of the two products was chosen, participants were also asked to indicate their confidence on a four-point scale ranging labeled 1 = very uncertain, 2 = hardly certain, 3 = quite certain, 4 = very certain. Each participant was paid 2.50 Euro plus five cent for each correct decision minus five cent for each wrong decision. The amount won or lost was doubled on trials with a confidence level of 3 or 4.

The presentation order of the 120 trials was randomized within participants. The 20 samples for each combination of sample size n (16 or 32) and Δ (.1 vs. .2 vs. .4) were drawn randomly, with replacement, from a 1000-item population (e.g., including, for Δ=.2, 300

smilies vs. 200 frownies for one product and 200 smilies vs. 300 frownies for the other product). Out of these 20 contingencies per condition, the actually superior product appeared on the left side on 10 trials, whereas on the other 10 trials the superior product appeared on the right side. The population distributions were always symmetrical (i.e., 275, 225, 225, 275 vs. 300, 200, 200, 300 vs. 350, 150, 150, 350 for Δ = .1, .2, and .4, respectively). The starting value for the computer's random generator was reset for each participant.

Results and Discussion

Hits and False Alarms in Large and Small Samples. For an initial check on the premise of the decision-threshold approach, we first inspected the effective stimulus contingencies that provided the input to the cognitive decision task, that is, the distribution of hits and false alarms produced by small and large samples at different Δ levels, as a function of variable decision thresholds *c*. Table 3 shows that, consistent with earlier simulations, small samples produced more hits than large samples, especially when the decision threshold *c* is moderate or high. Small samples also produced more false alarms than large samples, to be sure, but the false alarm costs were often smaller than the hit benefits, provided *c* is at least .4 or .5. This small-sample advantage *at the sampling stage* occurs if, and only if, the threshold *c* is higher than the actual contingency Δ in the population.

Correctness of Decisions. Let us now look at whether the original diagnostic value of small samples may, at least under specific conditions, survive the *cognitive decision_stage* in human decision makers. We first consider the overall correctness rates across all experimental conditions, regardless of subjective confidence and the presumed decision threshold. Even at this crude level of analysis, small samples already fare quite well, both absolutely and relative to large samples. Across all 92 participants of all four presentation mode conditions, the average within-participant correlation between the contingency-based decision (counting –1 and +1 for right and left product choices, respectively) and the *actual population contingency* (Δ varying from –.4 to –.2 to –.1, when the right product was superior, to +.1 to +.2 to +.4,

when the left product was superior) amounts to $r = +.72$ for large samples and $r = +.62$ for small samples, reflecting remarkable sensitivity to the choice implicated by environmental contingencies in a modest range. The respective average (within-participants) correlations between decisions and the *contingencies manifested in the sample* was .80 for large samples and .77 for small samples. These preliminary data also testify to the generally high motivation of the participants and the general reliability of the whole data array.

For a more systematic analysis of decision accuracy, still across all trials, regardless of decision threshold, we computed two indices within each participant and separately for each sample size x contingency level condition. The first index, *correctness*, is simply based on coding correct and incorrect decisions +1 and –1, respectively, and computing the average value on this scale, across all trials where a decision has been made. The second score, *confident correctness*, uses the same positive and negative coding for correct and incorrect decisions, respectively, but weighted by the corresponding confidence level. Accordingly, this score can vary between –4 and +4. Table 4 gives the means of both indices as a function of experimental conditions, together with other relevant statistics.

The overall level of accuracy is clearly evident in the positive range, suggesting that our attempt to raise performance above chance was successful. One can also see from Table 4 that correctness increased markedly with contingency levels. Correctness was consistently slightly higher for large than for small samples, although the difference was only in the range of .1 on the –1 to +1 correctness scale. A similar pattern was obtained across all presentation mode conditions (see Table 4). For the Successive/Symbol-Remain, Simultaneous/Symbol Remain, Successive/Symbol Erase, and the Simultaneous/Symbol Erase conditions, respectively, both a very strong contingency level main effect, $F(2,44) = 106.57$, 96.16, 71.61, and 88.80, all $p < .001$, (in the same order) and a significant sample-size main effect $F(1,22) = 7.02$, 21.72, 15.54, and 27.37 were obtained. A modest interaction term was found in two of the four analyses (Remain / First Left), $F(2,44) = 4.34$, $p < .05$, and Successive/Symbol Erase, $F(2,44)$

= 3.58, both *p*s < .05), mainly reflecting a ceiling effect for almost perfect performance at Δ = .4. When the two presentation mode factors were included in the ANOVA, no interaction emerged between the two main effects for sample size and contingency level and any presentation mode factor, suggesting a fairly general pattern.

Even at this overall level of analysis, independently of decision threshold, the accuracy obtained by increasing samples from 16 to 32 items was only modest. Given as many as 20 contingencies of the same type within participants, it is interesting to look at inter-individual differences. The number of participants out of 92 whose average performance was not higher for large than small samples was 36 at Δ = .1, 22 at Δ = .2, and 36 at Δ = .4. Note that it is uncontestable that accuracy increases with sample size when all trials are included (except for a few non-decisions), regardless of decision threshold. Nevertheless, the present results demonstrate that even this advantage is only modest and absent altogether in about one third of individual judges, in spite of a rather large number of observations within judges.

Subjective Confidence. Before we consider the analysis for confident correctness (i.e., correctness scores weighted by confidence), let us look at the pure confidence measures. In general, mean decision confidence increased from Δ = .1 (M = 2.80) to .2 (M = 2.93) to .4 (M = 3.33), as reflected in a contingency level main effect, $F(1, 176) = 219.567$, $p < .001$, in a four-factorial overall ANOVA. The sample size main effect was also significant, $F(1,88) = 4.91$, $p < .05$, reflecting somewhat higher confidence with small (M = 3.05) than with large samples (M = 2.98). The contingency level x sample size interaction, $F(1,176) = 12.47$, $p < .001$, indicates that the enhanced confidence for small samples is confined to Δ = .1 to .2 and disappears for Δ = .4. The relatively high confidence of small-sample decisions is consistent with the notion that small samples often provide a clearcut picture of evidence that can be assumed to often exceed some decision threshold – whatever the value of that threshold is. The only other significant result was due to higher confidence for remaining rather than removed samples, $F(1,88) = 8.18$, $p < .01$. We refrain from discussing this plausible finding.

Confident Correctness. Accordingly, weighting the correctness scores by subjective confidence should strengthen performance for small samples, relative to large samples. Indeed, the sample-size main effect was weaker and no longer significant in the ANOVA for the Symbol Remain / First Left condition, $F(1,22) = 3.20$. For the other three presentation mode conditions, the sample size main effect was significant, $F(1,22) = 13.10$, 9.68, and 29.23, for Symbols Remain / Alternating, Symbols Erase / First Left, and Symbols Erase / Alternating conditions, respectively. Two comments on these findings are in order. First, the degree of large-sample superiority is reduced when correctness is weighted by confidence, consistent with the notion that high confidence reflects having passed a tight decision threshold. Second, the residual superiority of large samples is highest in the most difficult Symbols Erase / Alternating condition, in which the noise ratio appears to be highest (see Table 4) but no longer significant in the Symbols Remain / First Left condition that is most likely to facilitate the encoding of small-sample packages without too much information loss.

Conditionalizing Correctness on Decision Threshold. In accordance with the decision-threshold model, high decision thresholds should render the assets of small samples most apparent. We thus computed conditional correctness scores, conditionalized on the threshold assumption, that is, counting only those trials on which the evidence given in the stimulus sample can be expected to have exceeded a certain decision threshold. The present data offer two ways of running such a conditionalized analysis. Either we can count only those trials in which the sample Δ exceeded a certain level, assuming that decision makers are sensitive to the sample contingency. Or, we count only those trials in which the judge's own subjective confidence rating exceeds some critical value, assuming that these post-hoc ratings provide a useful measure of the implicitly used decision threshold. Both operationalizations are imperfect, but both analyses shed some interesting light on the influence of the decision threshold on the relative performance triggered by large and small samples.

We first consider an analysis that includes all decisions made at the highest confidence level of 4. Analogous to the simulation analyses, we computed for all combinations of sample size and Δ within each participant the number of hits minus the number of false alarms (out of 20 trials) obtained with this confidence threshold. The resulting difference score turned out to have rather equal variance so that the analysis could be conducted immediately on the f(hits) – f(false alarm) scores. As evident from Table 4, only the huge contingency size main effects remained significant, whereas the sample size main effect was eliminated in three of the four ANOVAs conducted for each presentation mode condition. The only sample-size main effect, $F(1,22) = 5.48$, $p < .05$, for the Symbols Erase / Successive condition, was rather weak.

When similar f(hits) – f(false alarms) scores were computed within each participant (for all combinations of sample size and Δ) but conditionalized on a sample contingency of $\Delta_{Sample} \geq .4$, the sample size main effect vanished in all four presentation mode conditions (bottom of Table 4). Thus, both ways of operationalizing a high decision threshold serve to eliminate the modest large-samples advantage. Although at the end of the ecological sampling and cognitive decision-making process there is not a small-sample advantage, but merely a tie, this finding can be interpreted as a small-sample advantage in efficiency when information costs are taken into account. Moreover, the relative advantage of small samples became apparent and the normal superiority of large samples was most likely to be eliminated under conditions that could be derived from our threshold-based decision model.

Experiment 3

For still another way of operationalizing the interplay of sample size and decision thresholds, we ran another experiment. Rather than controlling sample size experimentally and letting confidence and evidence strength vary, we made an attempt to build the threshold into the decision task and to let sample size vary as a dependent measure. Thus, using a self-determined sampling paradigm, we instructed participants to sample data as long as necessary to make an informed (i.e., above-threshold) decision. Specifically, participants observed

growing samples of smilies and frownies, again referring to two target products that were drawn from a universe defining the objectively true contingency. At any point in time when the participant felt that the evidence was sufficient to make a choice, he or she could truncate the information search process. Note that this self-determined sampling paradigm is quite representative of the ecological task setting that motivates the decision-threshold model, assuming that organisms defer decisions until they have gathered satisficing information, and do not make a decision if the strength of the evidence in the sample is insufficient.

As sample size becomes a dependent variable in this paradigm – with large samples reflecting a longer period of uncertainty – we discover another intriguing aspect of the cognitive-environmental interaction that gives an advantage to small samples. If on easy trials the first few observations in a sample already provide clearcut evidence, then a confident and often correct decision can be based on few data. If, however, on difficult trials the initial evidence happens to be unclear, the required sample size will be larger while confidence and accuracy may often remain low. Thus, whenever information search is self-determined – so that decision makers design their own samples (Kareev & Fiedler, 2004) – it is likely that easy samples tend to be small whereas difficult samples tend to be larger. Within the framework of threshold-based decision making, this must not be discarded as an artifact. The correlation between sample size and difficulty is a natural consequence of a cognitive process that defers a decision until sufficient evidence is available.

Method

Participants and Design. Seventeen male and female students of the University of Heidelberg participated for payment (about 2 €). The only experimental variable, degree of contingency, was manipulated within participants. Each participants completed 120 trials, 30 at each level of Δ (0 vs. .1 vs. .2 vs. .4). The dependent variables of interest were correctness and sample size required to make a decision, as well as the relationship between both.

Materials and Procedure. The basic experimental task was similar to that of Experiment 1. Instructions explained that smilies and frownies represent positive and negative evaluations of two products, using the same basic cover story. Instructions emphasized the goal to make as many correct decisions as possible and to avoid wrong decisions. However, the total time would be restricted so that they should not lose more time on individual trials than necessary to make a correct decision. Thus the experimental procedure encouraged them both to try to be correct, and to use a sample as small as possible to achieve that goal. After a clear-screen inter-trial interval of 2000 ms, each trial started with the presentation of two products labels (random strings of five letters and numbers) on top of the left and right side of the screen, respectively. Then a smily or frowny was added every 500 ms, in a random location of the appropriate side of the screen, under the constraint that no other symbol already occupied that location. Symbols were not erased so that the growing record of smilies and frownies associated with the two products could be pursued on the screen.

Information search could be stopped at any time, by releasing the space bar, when the participant felt the evidence would be sufficient to make a decision. At that moment, all smilies and frownies were erased, but the product labels and screen colors remained as participants were offered three response options: choosing the left product (using the left arrow key), choosing the right product (right arrow key), or not making a decision (downward arrow key). Immediately afterwards, they were asked to indicate their subjective confidence by adjusting a light pointer on a 42-segment graphical rating scale.

Symbols were randomly sampled from a population defined by a certain Δ. Each participant was administered 120 trials, 30 at each of the four levels of Δ (0 vs. .1 vs. .2 vs. .4). Trials of the $\Delta = 0$ condition were treated as fillers and excluded from the analysis. The superior product was presented on the left and the right side on one half of the trials of each Δ condition. Order of presentation of the 120 trials was randomized.

Results and Discussion

Self-Determined Sample Size. Under the task conditions of this experiment (contingency levels, presentation duration and speed, motivation, memory overload etc.), the mean self-determined sample size required to make a decision, averaged across participants and trials within participants, was 28.1 for $\Delta = .1$, 26.9 for $\Delta = .2$, and 25.7 for $\Delta = .4$. The standard deviation of individual self-determined mean sample sizes was 5.35 for $\Delta = .1$, 5.27 for $\Delta = .2$, and 5.59 for $\Delta = .4$. The relative constancy of sample sizes across contingency levels suggests that judges were not particularly sensitive to the different amounts of information required to detect a contingency at different levels of Δ. The relative independence of sample size from from effect size (Δ) may reflect the aforementioned stopping rule. Information search may have been truncated whenever the first few observations of a sample revealed a clear-cut picture, and this primacy effect might have occurred at all three contingency levels. For an empirical check, one has to look at the sample contingencies associated with small and large numbers of trials.

Sample Size and Sample Contingency. Accordingly, we calculated, within each participant, the correlation between the size of the various samples gathered on different trials and the observed Δ in the sample (coded positive if the sign was correct). If this correlation was negative, small samples were indeed reflective of stronger sample contingencies – the kind of primacy effect depicted above. Furthermore, this would yield another reason for a small-sample advantage *at the sampling stage* – namely, when active information search renders sample size a dependent variable. Indeed, the average correlation between sample size per trial and (correct) contingency size in the sample (within each individual participant) turned out to be clearly negative, $\underline{r} = -.37$. Testing the vector of all 17 individual correlations against zero resulted in a highly significant $t(16) = -7.46$, $p < .001$.

Thus, the present experiment corroborates that in the context of self-terminated information search, when information is sampled until a satisficing amount of evidence is found *in the sample*, the resulting relation between samples size and contingency strength in the sample can be negative. That is, not only do small samples reflect actually existing contingencies more pronouncedly than large samples (that much has already been revealed by our statistical analyses and simulations), but it is also the case that people are satisfied with the evidence, terminate the search, and reach a decision. Obviously, if human decision makers assess the samples they have observed objectively (using notebooks or computers or other devices to avoid error or inaccuracy), their decisions will exhibit a small-sample advantage. However, let us now also consider the fate of small and large samples in a cognitive decision process that cannot rely on such external-memory devices. Note that under the brief and demanding exposure and learning conditions of the present experiment, participants had to capture contingencies from a single presentation of a series of stimuli (one per .5 sec), forcing them to rely on a highly fallible learning and memory process. Would the performance on small-sample trials, which allow for only a little learning under ill-suited conditions, still be higher or comparable to the performance on large-sample trials?

Sample Size and Correctness. For an empirical answer to this crucial question, we calculated, within each participant, the correlation between the correctness of decisions and the size of the sample drawn on corresponding trials. The pertinent results appear in Table 5. Across all contingency levels and without any restriction of reasonable sample size, the average overall correlation between sample size and correctness is essentially zero ($r = -.014$), reflecting that over a wide range of sample sizes (the middle 80% of the sample size distribution fell between 14.91 and 44.97 trials) performance was not inferior (to say the least) with smaller sample sizes.

Because some judges were obviously ill-calibrated, basing many decisions on too small sample sizes, we re-computed the size-correctness correlations within judges, counting only

those trials on which the sample sizes were in a reasonable range around the average of spontaneously sought sample sizes (i.e., between 20 and 32) and only for trials with a correlation of at least $\Delta = .2$. Note that this sample-size range is centered around the average self-selected sample sizes, which can be assumed to represent the actually chosen decision criterion. Under these more auspicious conditions for small samples, the average size-correctness correlation became $r = -.09$, and the vector of all 17 individual correlation actually had a central tendency that was significantly negative, $t(16) = -2.36$, $p < .05$. All other combination of sample size range and contingency level included in Table 5 revealed no significant differences from zero, suggesting a generally constant performance across considerable variation in sample size.

Interindividual Differences. Out of all 17 participants, the correlation between sample size and correctness was negative for 13 individuals across all trials with $\Delta \geq .2$ and for 11 individuals across all trials with $\Delta \geq .1$. We finally calculated the correlation across participants between individual size-correctness correlation and confident correctness scores (i.e., correctness weighted by average an individuals confidence). Those who were generally more correct tended to be those who were better on small than on large samples, as reflected in a negative correlation of $r = -.23$ for $\Delta \geq .1$, $r = -.20$ for $\Delta \geq .2$, and $r = -.39$ for $\Delta = .4$.

General Discussion

The starting point of the present article was the intriguing possibility – recently pointed out by Kareev (1995, 2000) – that small samples of observations may reveal existing environmental contingencies more clearly and more regularly than larger samples of observations. Due to the skew of the sampling distribution, a majority of sample correlations exceeds the population correlation, and this characteristic of the empirical world reaches, for binary variables, a maximum at about 7±2 observations, fitting the window size of human short-term memory (cf. Kareev, 2000). Moreover, because the sampling distribution is skewed only for existing, but not for zero correlations, the accentuating effect of small

samples is likely to be diagnostic of actual environmental contingencies. This demonstration adds a new and potentially important item to a growing list of less-is-more phenomena (Elman, 1993; Gigerenzer, 2001; Hertwig & Todd, 2003, Newport, 1988, 1990) that have become prominent in recent research on adaptive cognition.

Fascinating as the small-sample advantage might be, however, two sources of criticism seemed to challenge Kareev's argument and to reduce its psychological significance. On one hand, the seeming advantage of small samples over large ones may disappear as soon as both hits and false alarms are taken into account (Juslin and Olsson, 2004). On the other hand, any advantage of small samples, may be overridden by the effect of an opposing principle that governs the cognitive process fed by environmental samples. Whenever information processing is not error-free – due to noise or imperfect learning and memory – the beneficial effect of aggregation comes into play (Fiedler, 2000; Fiedler & Walther, 2003): Error is gradually reduced and existing statistical relationships become increasingly visible as the number of observations increases. Thus, whereas small samples tend to amplify existing contingencies in the environment, there is benefit in large samples, as evident in monotonically increasing learning curves (cf. Figure 1).

The present approach speaks to both sources of criticism and eventually arrives at the conclusion that the effect of the feature of the empirical world that was highlighted by Kareev (2000) is not at all illusory. Small samples may under certain conditions inform better decisions than large samples, and this advantage need not be confined to hits but persists even when false alarms are considered and when memory loss is taken into account.

We have proposed a threshold-based decision framework within which the confines of the small-sample advantage can be deduced, tested, and understood. The basic assumption underlying this framework is that adaptive intelligence relies heavily on binary choice – choosing the better of two options – based on a restricted sample of observations. The elementary model of this binary-choice situation is a 2 x 2 contingency between dichotomous

variables (2 options x 2 outcomes). Within a satisficing rather than an optimizing approach, it is often not necessary to estimate the strength of such contingencies precisely; rather, it suffices to figure out the sign of the contingency, in order to identify the better option. According to the threshold-based decision model, decisions are contingent on the sampled evidence favoring one option over the other to a sufficient degree. Our analyses demonstrated that sufficiently large differences are more likely to occur when the sample size is small rather than large, reflecting not only the greater lability and higher dispersion, but also the greater skew in the distributions of small samples. To be sure, the same lability of small samples may not only produce many hits (i.e., evidence exceeding the threshold for the correct option) but also false alarms (i.e., evidence exceeding the threshold for the incorrect option). However, our analyses indicate that the small samples' gains in terms of hits are larger than their losses in terms of false alarms (see Figure 3), provided the decision threshold is high enough. A similar conclusion was recently reached independently Anderson & Doherty (2004).

Note that the rationale derived within this decision-threshold framework for sometimes expecting better performance with small than with large samples is no longer the same as in Kareev's (2000) original phenomenon. Whereas Kareev (2000) was solely concerned with the enhanced *skew* of small-sampling distributions, in the framework presented here the *curtosis* of sampling distributions plays a major role: It is the high dispersion of sampling distributions obtained from small samples that facilitates observing many above-threshold contingencies. This characteristic is not confined to strong or very strong contingencies, which are required to produce a marked skew in sampling distributions. Rather, the present argument applies to all non-zero contingencies – weak, medium and moderately strong – that can be expected to hold in reality. The only restriction to the size of contingencies is that they must allow for an even higher decision threshold.

Several computer-simulation studies were reported to test this contention and to explore the generality of the small-sample advantage. The first set of simulations pertained to the pre-

cognitive sampling stage, that is, the transition of a latent population contingency into environmental samples. Three parameters were varied in these initial simulations: Sample size n, contingency strength Δ, and the decision threshold c. As predicted on analytical grounds, the difference between the numbers of above-threshold hits and false alarms actually decreased with increasing sample size (from 4 to 32) when the decision threshold c was substantial and when the environmental contingency Δ was different from zero but lower than c. For strong and very strong contingencies, the simulation results approached a ceiling effect anyway, that is, the difference between hits and false alarms was very large, and similar with both large and small samples. Overall, the small-sample advantage at this environmental sampling stage (i.e., the downward inclination of curves in Figure 4) covers a remarkable part of the parameter space.

The next set of simulations applied a connectionist algorithm that has been shown to describe quite well cognitive processes of judgment and decision making, in particular, the aggregation effect – a monotonic increase of probabilistic learning with number of trials. As the algorithm depends on the amount of noise in the input data, the noise ratio i was included as a fourth parameter. Even though learning always increases with the number of trials, and especially so for high-noise environments (thus counteracting the small-sample advantage in the environment), decision accuracy still did not generally increase with sample size. Rather, under conditions derivable from the decision-threshold model, the small-sample advantage outweighed or even overrode the learning effect, producing fairly horizontal or even slightly decreasing performance curves (Figure 6). In particular, this was evident in cases of low noise and decision thresholds higher than the environmental contingency.

We then went on to simulate a variant of cognitive contingency assessment, referred to as competing tendencies. Though much less popular than standard contingencies and hardly covered in statistics books, competing-tendency tasks are quite common in everyday decision-making. The crucial feature of this class of tasks is that information about the two options is

not independent but, due to meaning overlap, all information bears on the other option as well. To the extent that many comparisons and dichotomies are not strictly exclusive (e.g., a student's performance in math and physics; two computer brands sharing the same technical modules; love and hate), small samples may more effectively discriminate between the superior and the inferior option than do large samples (which result in about equal learning for the weaker as for the stronger option; cf. Figure 7). In fact, an additional set of simulations for the competing tendencies variant resulted in some marked decrement in the discriminating ability between the strong and the weak tendencies as sample size increased.

Altogether, these findings demonstrate that an advantage of small over large samples is neither impossible on purely logical grounds nor a miracle from a psychological point of view. Whenever the assumptions of the threshold-based decision model are met – that is, when satisficing, above-threshold choices are called for rather than accurate quantitative estimation – then a smaller sample may well lead to a better decision than a large sample, just because the small-sample contingency is more likely to exceed the decision threshold (in the correct direction) than the large-sample contingency. As long as the sampled data are assessed reliably, with no information loss, using external memory devices (notebooks, recorders etc.), the advantage of small samples is immediately manifested in better decisions. If, however, contingency assessment relies on fallible human learning and memory processes, then the advantage of small samples is reduced due to the competing advantage that large samples have in learning in noisy environments (i.e., the aggregation effect). Even then, however, it was shown that the less-is-more phenomenon may survive the cognitive stage of decision making under notable auspicious conditions.

Systematic experimental work on decision performance as a function of sample size lags behind the wide set of conditions explored in the simulation studies. Up to now, we have run only a few experiments designed to measure contingency-based decision making in task situations that speak to the decision-threshold model. Although the stimulus presentation

conditions were quite demanding in these experiments (e.g., a single brief exposure of sampled items), resulting in a good deal of noise and suboptimal contingency learning, no substantial advantage of large over small samples was observed. Moreover, when specific conditions derived from the threshold model could be assumed to be met – namely, low noise, sufficiently high contingencies and rather high thresholds – the small samples led to equally good and, at least sometimes, even better decisions than large samples.

It should be mentioned that these findings appear to be quite stable as they were also obtained in several other data sets not reported here. This conclusion even holds for the "home domain" of the typical aggregation effect, the simulated classroom experiments mentioned at the outset (Fiedler et al., 2002). Participants ('teachers') who had to discern the ability and motivation parameters of 16 students in a simulated school class from observed samples of correct and incorrect responses, and from the frequency of raising hands, clearly performed better for those students for which larger samples of performance were available. However, when decision-threshold assumptions were imposed on post-hoc analyses of the same data, by calling for above-threshold sample contingencies between pairs of students and the proportion of correct responses in the observed data, small samples more often exceeded a threshold than did large samples. And teachers' above-threshold judgments of the difference between pairs of students were about equally accurate for large and small samples, provided a sufficiently strong threshold.

Nevertheless, experimental evidence is still rather scarce and clearly lags behind the insights gained by computer simulations. Although the few experiments reported here are encouraging, further experimental evidence is clearly needed. One obvious goal would be to think of ways of manipulating the threshold and noise parameters in natural ways. Another important extension would be to include information cost in an experiment designed to control the precise payoffs (benefits and costs) of hits and false alarms of contingency-based binary choices. Finally, it would be particularly interesting to see whether a clearcut small-

sample advantage can be demonstrated under less demanding exposure and encoding conditions than in the experiments reported above.

Several psychological and pragmatical implications of these findings suggest themselves. From an economical point of view, the challenging idea that the added value of further information can be negative – an implication hard to reconcile with common notions of rational choice – has to be tackled. The necessary costs for personnel selection decisions or investment choice may be reduced considerably by setting an optimal threshold and relying on small samples rather than large but overly conservative samples. A similar point can be made for scientific research. To the extent that research is not governed by rules of precise estimation but by rules of detecting strong effects, the present approach suggests that increasing sample sizes may not always be of advantage in scientific research.

However, apart from such practical and methodological considerations, the main motive underlying the present investigation is theoretical. Our research on the advantages and disadvantages of small and large samples highlights the value and need for a genuinely cognitive-ecological theory approach. The interaction of cognitive strategies and constraints on one hand and properties of the information ecology on the other hand produces emergent models of adaptive cognition that can go way beyond the insights of a traditional, purely cognitive approach. What appears like a miracle or impossibility from a rationalist point of view – namely, that less can be more – turns out to be a natural product of cognitive-ecological interaction.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin, 111*, 256-274.

Anderson, R.B., & Doherty, M.E. (2004). *A criterion-specific advantage for small samples in the detection of correlation*. Unpublished research.

Borges, B., Goldstein, D. G., Ortmann, A., & Gigerenzer, G. (1999). Can ignorance beat the stock market? In G. Gigerenzer, P.M. Todd, and the ABC Group (Eds.) *Simple heuristics that make us smart* (pp. 59-72). Oxford: Oxford University Press.

Elman, J.L. (1993). Learning and development in neural networks: The importance of staring small. *Cognition, 48*, 71-99.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Fiedler, K. (1991). The tricky nature of skewed frequency tables: An information loss account of distinctiveness-based illusory correlations. *Journal of Personality and Social Psychology, 60*, 24-36.

Fiedler, K. (1996). Explaining and simulating judgment biases as an aggregation phenomenon in probabilistic, multiple-cue environments. *Psychological Review, 103*, 193-214.

Fiedler, K. (2000). Illusory correlations: A simple associative algorithm provides a convergent account of seemingly divergent paradigms. *Review of General Psychology, 4*, 25-58.

Fiedler, K., Kemmelmeier, M., & Freytag (1999). Explaining asymmetric intergroup judgments through differential aggregation: Computer simulations and some new evidence. *European Review of Social Psychology, 10*, 1-40.

Fiedler, K., Russer, S., & Gramm, (1993). Illusory correlations and memory performance. *Journal of Experimental Social Psychology, 29*, 111-136.

Fiedler, K., Semin, G.R., Finkenauer, C., & Berkel, I. (1995). Actor-observer bias in

close relationships: The role of self-knowledge and self-related language. *Personality and Social Psychology Bulletin, 21*, 525-538.

Fiedler, K., & Walther, E. (2003). *Stereotyping as inductive hypothesis testing*. New York: Psychology Press.

Fiedler, K., Walther, E., Freytag, P., & Plessner, H. (2002). Judgment biases in a simulated classroom – a cognitive-environmental approach. *Organizational Behavior and Human Decision Processes, 88*, 527-561.

Fiedler, K., Walther, E., & Nickel, S. (1999). The autoverification of social hypotheses: Stereotyping and the power of sample size. *Journal of Personality and Social Psychology, 77*, 5-18.

Forest, J.A., & Feldman, R.S. (2000). Detecting deception and judge's involvement: Lower task involvement leads to better lie detection. *Personality and Social Psychology Bulletin, 26,* 118-125.

Gigerenzer, G. (2001). The adaptive toolbox. In G. Gigerenzer & R. Selten, (Eds.), *Bounded rationality and the adaptive toolbox* (pp. 37-50). Cambridge, MA: The MIT Press.

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

Hertwig, R., & Todd, P.M. (2003). *The benefits of cognitive limits: Why bigger isn't always better*. Unpublished manuscript.

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

Juslin, P., & Olsson, H. (2004). *Capacity limitations and the detection of correlations: Comment on Kareev (2000)*. Manuscript submitted for publication.

Kareev, Y. (1995). Through a nerrow window: Working memory capacity and the detection of covariation. *Cognition, 56*, 263-269.

Kareev, Y. (2000). Seven (indeed, plus minus two) and the detection of correlations. *Psychological Review, 107,* 397–402.

Kareev, Y., & Fiedler, K. (2004). *On the accentuation of contingencies: The sensitive research designer versus the intuitive statistician*. Manuscript submitted for publication.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126*, 278-287.

Krauss, S., & Wang, X.T. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General, 132*, 3-22.

Kuder, G.F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151-160.

Lewontin, R.C. (1979). Sociobiology as an adaptionist program. *Behavioral Science, 24*, 5-14.

Martignon, L., & Hoffrage, U. (1999). Why does one-reason decision making work? A case study in ecological rationality. In G. Gigerenzer, P.M. Todd, and the ABC Group (Eds.) *Simple heuristics that make us smart* (pp. 119-140). Oxford: Oxford University Press.

Newport, E.L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of American Sign Language. *Language Science, 10*, 147-172.

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science, 14*, 11-28.

Rosenthal, R. (1987). *Judgment studies: Design, analysis, and meta-analysis*. Cambridge: Cambridge University Press.

Sande, G.N., Goethals, G.R., & Radloff, C.E. (1988). Perceiving one's own traits and others': The multifaceted self. *Journal of Personality and Social Psychology, 54*, 13-20.

Seger, C. (1994). Implicit learning. *Psychological Bulletin, 115*, 163-196.

Simon, H.A. (1956). Rational choice and the structure of environments. *Psychological Review, 63*, 129-138.

Swets, J., Dawes, R.M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1*, Whole No. 1.

Wilson, T. D., Dunn, D. S., Kraft, D., & Lisle, D. (1989). Introspection, attitude change, and attitude-behavior consistency: The disruptive effects of explaining why we feel the way we do. In L. Berkowitz, (Ed.) *Advances in experimental social psychology*, *22*, 287-343. San Diego, CA, US: Academic Press.

Wilson, T.D., & Schooler, J.W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*, 181-192.

Table 1

Number of Hits and False Alarms (FA) out of 10000 Simulated Sampling Trials as a Function of Sample Size *n*, Three Levels of Contingency Δ, and Absolute Decision Criterion Values |*c*|

	n = 4		*n* = 8		*n* = 12		*n* = 16		*n* = 24		*n* = 32	
|*c*|	Hit	FA	Hit	FA	Hit	FA	Hit	FA	Hit	FA	Hit	FA
						Δ = 0.10						
1.0	906	395	83	19	8	0	0	0	0	0	0	0
0.9	906	395	83	19	46	1	3	1	0	0	0	0
0.8	906	395	489	91	153	9	30	2	6	1	1	0
0.7	2199	1261	783	256	382	67	113	7	49	2	4	0
0.6	2199	1261	1038	391	637	139	322	45	147	9	46	1
0.5	3469	2132	2160	894	1391	364	903	155	395	32	227	11
0.4	3469	2132	2367	1049	2037	681	1728	395	1062	134	771	59
0.3	4559	3159	4119	1717	3362	1070	3097	760	2553	406	2084	261
0.2	4559	3159	4575	2592	4639	2118	4176	1523	4235	1214	3769	781
0.1	4559	3159	5488	3322	5532	2843	5788	2666	6067	2341	6169	2149
						Δ = 0.20						
1.0	1221	291	158	1	23	0	6	0	0	0	0	0
0.9	1221	291	158	1	99	0	38	0	2	0	0	0
0.8	1221	291	866	37	288	1	120	1	27	0	7	0
0.7	2719	940	1280	135	751	16	307	4	124	0	43	0
0.6	2719	940	1629	201	1120	59	806	17	390	0	227	0
0.5	4186	1552	3008	558	2163	158	1691	60	974	4	715	0
0.4	4186	1552	3262	666	3083	351	2894	158	2240	36	1962	8
0.3	5228	2608	5230	1172	4630	587	4549	365	4266	130	4004	54
0.2	5228	2608	5707	1819	5881	1303	5673	812	6185	470	6089	224
0.1	5228	2608	6564	2425	6743	1838	7212	1562	7785	1102	8128	831
						Δ = 0.40						
1.0	2414	84	600	0	123	0	41	0	0	0	0	0
0.9	2414	84	600	0	436	0	217	0	48	0	10	0
0.8	2414	84	2135	3	1053	1	627	0	343	0	141	0
0.7	4157	400	2845	18	2151	3	1255	0	1042	0	602	0
0.6	4157	400	3335	34	2921	8	2598	0	2322	0	1905	0
0.5	5960	723	5271	113	4653	21	4448	4	3888	1	3723	0
0.4	5960	723	5468	147	5695	48	6011	12	5937	3	6346	0
0.3	6785	1473	7400	347	7303	110	7570	44	7942	4	8229	1
0.2	6785	1473	7780	658	8269	289	8461	127	9053	27	9274	6
0.1	6785	1473	8377	998	8844	474	9247	340	9634	110	9822	56

Note: Below dotted lines, the difference of hits – false alarms decreases with sample size

Table 2

Summary of Results Obtained in Experiment 1 as a Function of Experimental Conditions

	Simultaneous		Successive	
Dependent Measure:	Small	Large	Small	Large
Correct Choice	.42	.64	.82	.71
Latencies	3.12	2.47	1.62	1.49
Confidence	23.66	24.51	24.74	26.06

Table 3

Number of Hits and False Alarms, Relative to Different Criterion Values *c*, as a Function of Contingency Levels Δ and Small (16) Versus Large (32) Sample Size

			c=.1	*c*=.2	*c*=.3	*c*=.4	*c*=.5	*c*=.6	c=.7	*c*=.8	*c*=.9	*c*=1
Δ = .1	Small	Hit	1028	840	623	423	225	120	55	19	6	0
		FA	476	343	202	101	40	18	6	4	1	0
	Large	Hit	1086	819	480	236	63	12	1	0	0	0
		FA	390	227	72	27	3	0	0	0	0	0
Δ = .2	Small	Hit	1260	1084	860	602	358	235	131	44	18	3
		FA	304	205	110	56	23	5	3	1	0	0
	Large	Hit	1419	1229	870	536	210	78	18	0	0	0
		FA	155	70	16	2	0	0	0	0	0	0
Δ = .4	Small	Hit	1633	1566	1442	1220	916	691	436	215	78	10
		FA	47	19	12	6	4	1	0	0	0	0
	Large	Hit	1726	1690	1551	1326	942	539	208	43	10	0
		FA	7	2	0	0	0	0	0	0	0	0

Note: At *c* levels right of the bold vertical lines, the difference of hits – false alarms is higher for small than for large samples

Table 4

Mean Correctness, Confident Correctness, and Conditional Correctness Scores Obtained in Experiment 2, as a Function of Experimental Conditions

	Small Samples			Large Samples			Sample size effect
Presentation mode	Δ = .1	Δ = .2	Δ = .4	Δ = .1	Δ = .2	Δ = .4	*F*(1,22)
Correctness							
Remain / First Left	.25	.57	.91	.38	.66	.93	7.02*
Remain / Alternating	.40	.53	.89	.46	.76	.97	21.72*
Erase / First Left	.25	.57	.86	.46	.73	.91	15.54*
Erase/ Alternating	.20	.48	.80	.28	.70	.93	27.37*
Confident Correctness							
Remain / First Left	0.73	1.80	3.18	1.06	1.99	3.24	3.20
Remain / Alternating	1.22	1.63	3.09	1.27	2.30	3.36	13.10*
Erase/ First Left	0.74	1.70	2.88	1.31	2.07	3.04	9.68*
Erase/ Alternating	0.55	1.21	2.36	0.71	1.86	2.90	29.23*
f(hits) – f(False Alarms) Conditional on Confidence = 4							
Remain / First Left	2.17	5.92	12.63	4.38	6.58	12.42	1.56
Remain / Alternating	3.38	5.96	10.83	3.29	6.33	12.21	1.52
Erase / First Left	2.35	4.96	10.04	4.74	5.48	11.13	5.48*
Erase / Alternating	1.83	3.29	7.00	1.17	3.96	8.13	0.85
f(hits) – f(False Alarms) Conditional on Sample Δ ≥ .4							
Remain / First Left	2.13	3.63	10.58	2.13	2.21	9.79	2.66
Remain / Alternating	2.00	3.71	9.92	1.75	2.17	10.21	1.78
Erase / First Left	1.48	3.39	8.96	4.78	2.04	9.52	2.67
Erase / Alternating	1.75	3.63	9.17	2.46	2.54	10.33	0.25

Table 5

Mean Within-Judge Correlations Between Sample Size and Correctness of Decisions as Function of Contingency Levels Δ and Limitations of Sample-Size Range

	Range of Population Contingencies Included		
Sample Size Range	$\Delta \geq .1$	$\Delta \geq .2$	$\Delta = .4$
Unlimited	–.014	+.024	+.011
> 20	–.049	–.055	–.012
20 – 32	+.008	–.088*	–.021

Figure Captions

Figure 1. The typical form of a monotonically increasing, negatively accelerated learning curve.

Figure 2. Two transitions involved in environmental contingency assessment: Ecological sampling and cognitive processing.

Figure 3. Sampling distributions within a threshold-based decision framework.

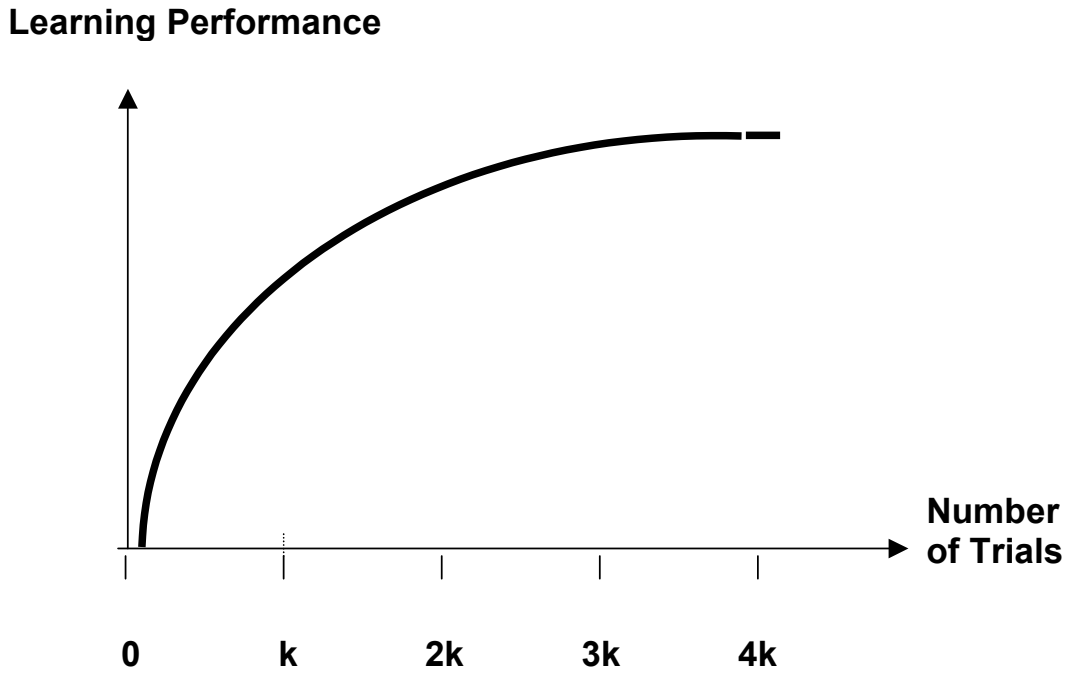
Figure 4. Simulation of sample contingencies as a function of population contingency, sample size, and decision criterion.

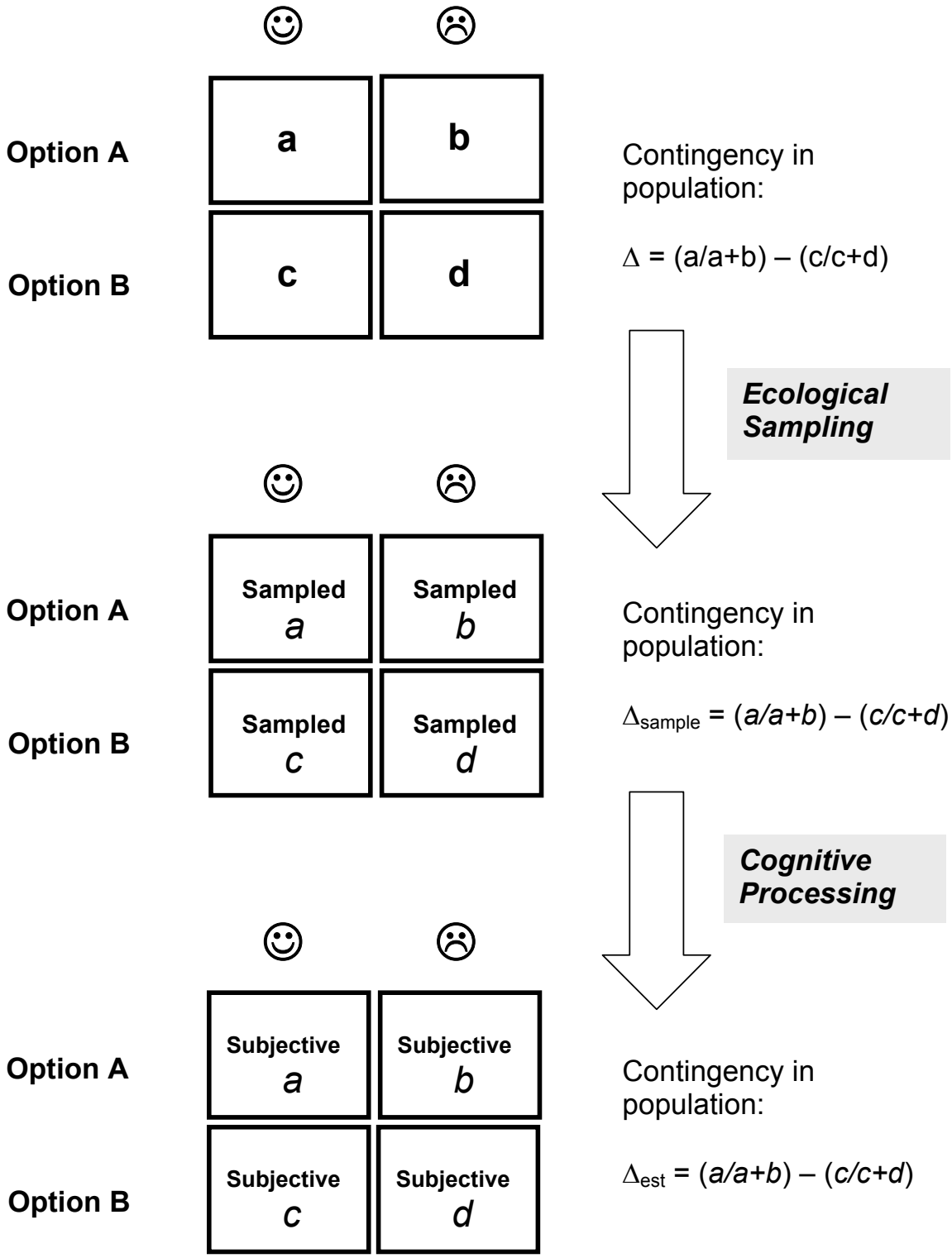
Figure 5. BIAS model simulation of the cognitive process of contingency assessment.

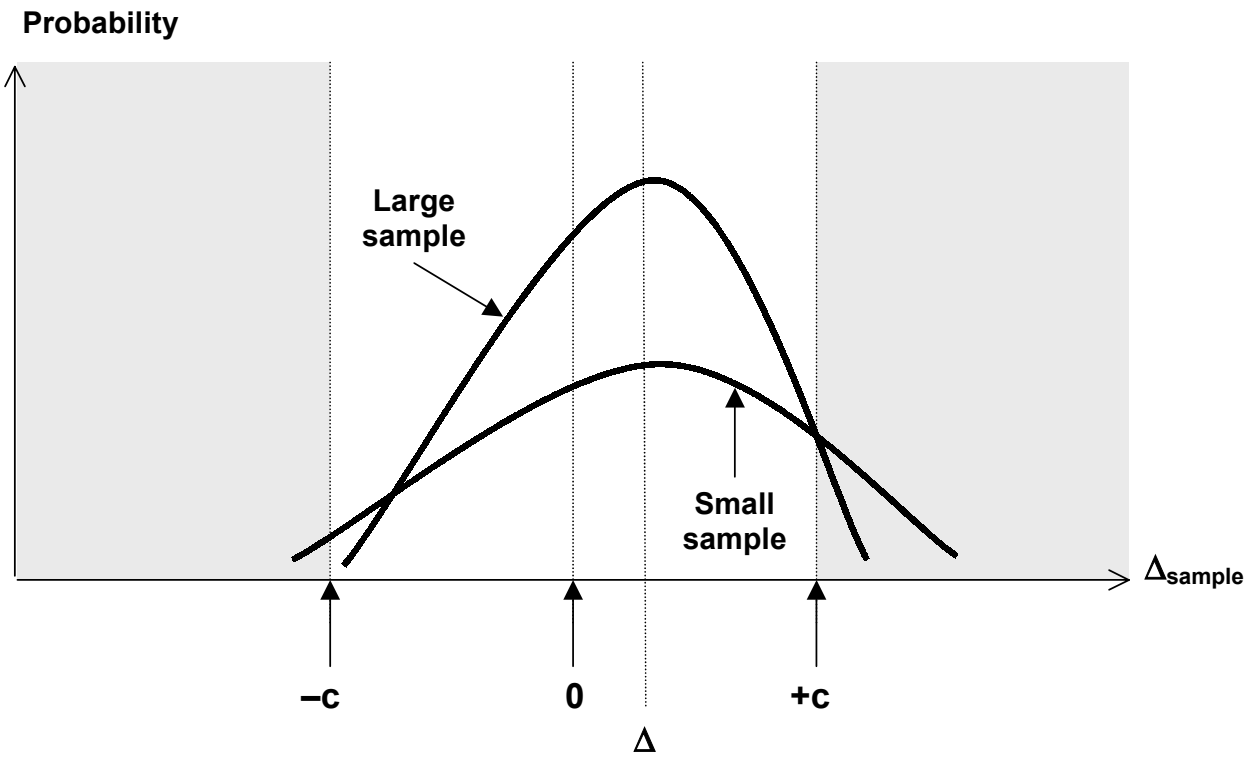
Figure 6. Simulation of cognitive decision accuracy as a function of population contingency, sample size, noise ratio, and decision criterion.

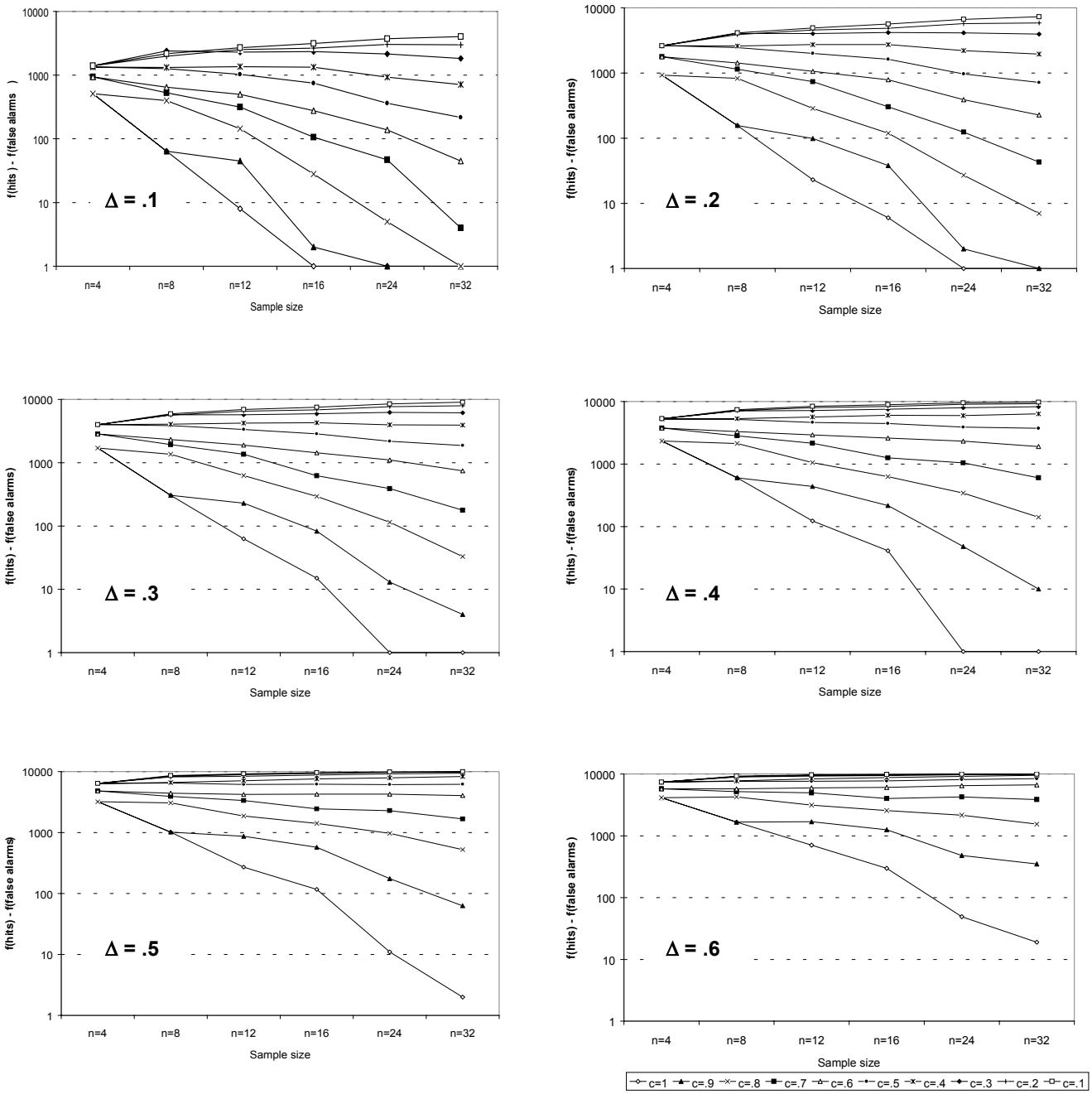
Figure 7. Illustration of learning curves for two competing tendencies.

Figure 8. BIAS model simulation of for two competing tendencies in a sample.

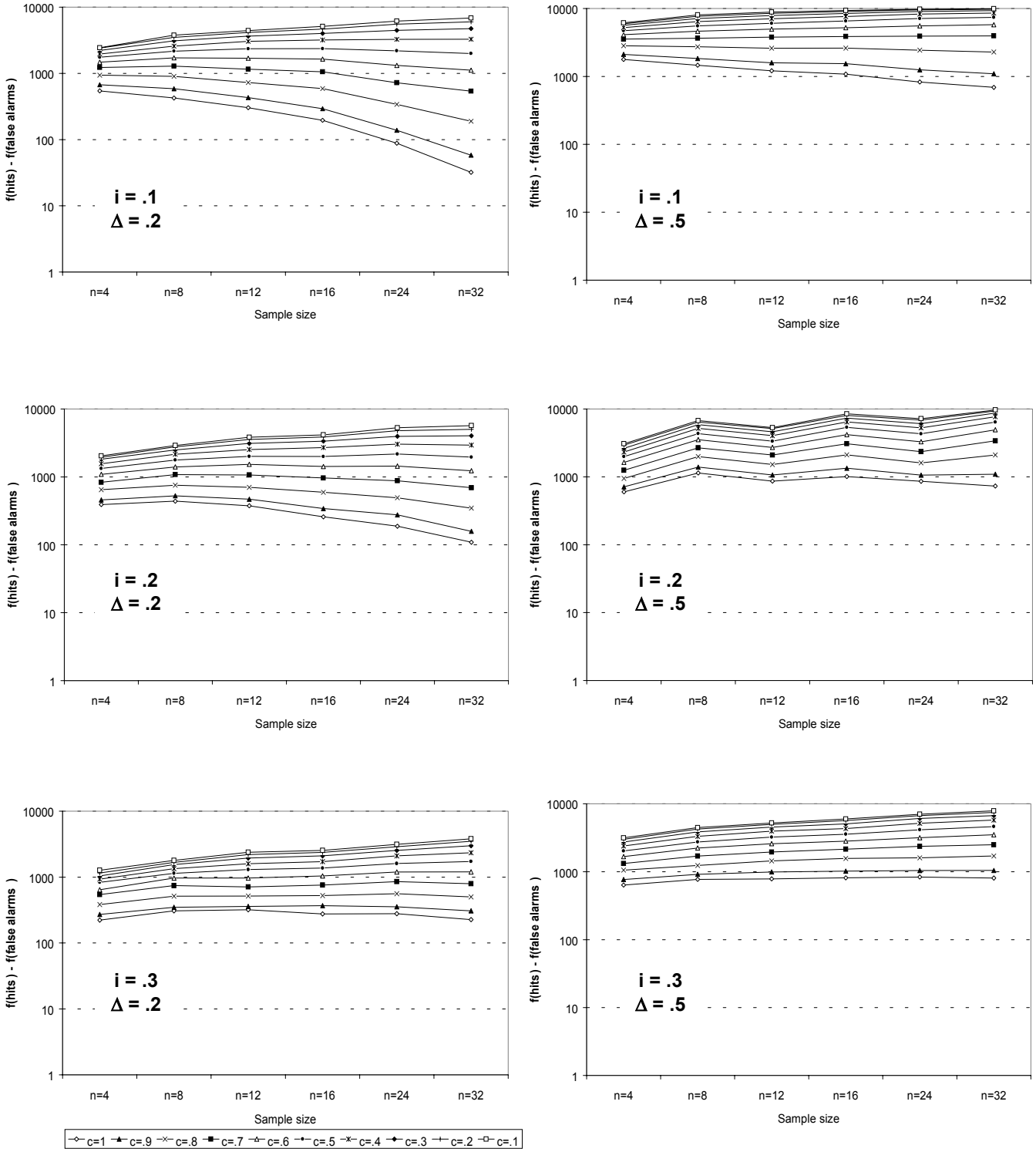

Learning Performance
Number of Trials
0
k
2k
3k
4k

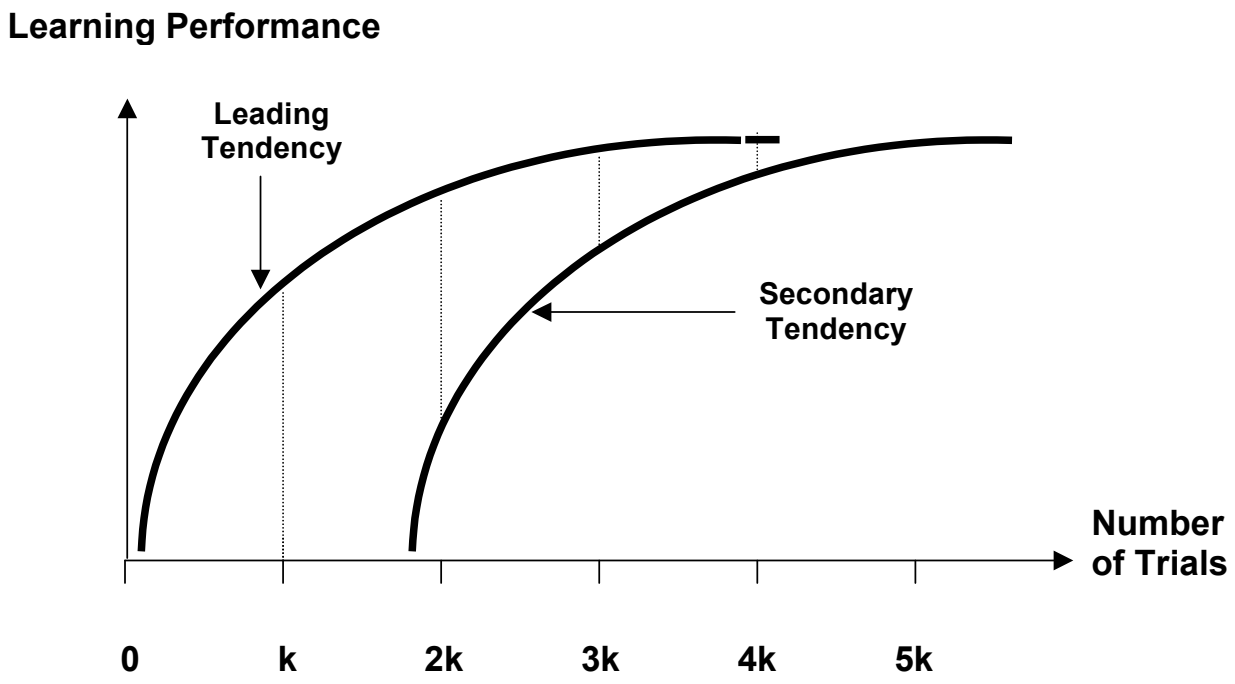

Option A
Option B
a
b
c
d
Contingency in population:
Δ = (a/a+b) – (c/c+d)
Ecological Sampling
Sampled a
Sampled b
Sampled c
Sampled d
Contingency in population:
Δsample = (a/a+b) – (c/c+d)
Cognitive Processing
Subjective a
Subjective b
Subjective c
Subjective d
Contingency in population:
Δest = (a/a+b) – (c/c+d)

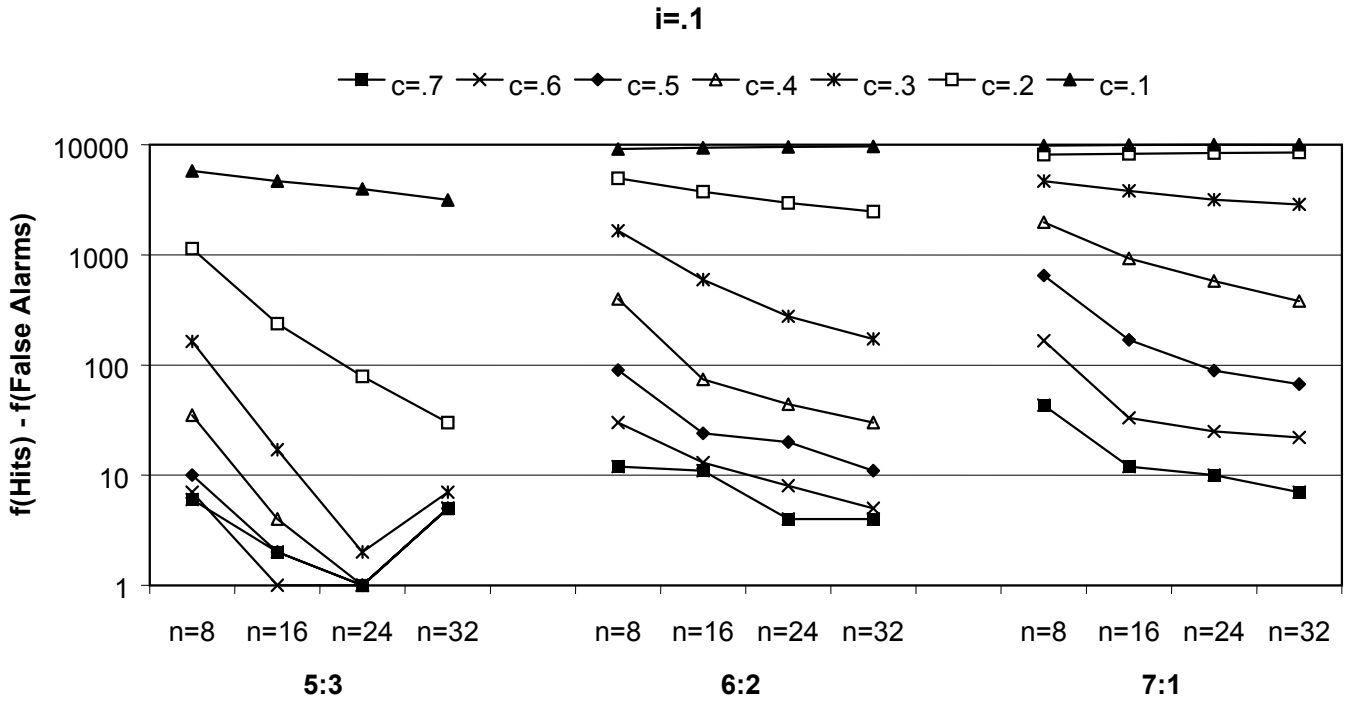

Probability
Large sample
Small sample
Δsample
–c
0
+c
Δ


10000
1000
100
10
1
f(hits) - f(false alarms)
Δ = .1
n=4
n=8
n=12
n=16
n=24
n=32
Sample size
f(hits) - f(false alarms)
Δ = .2
Δ = .3
Δ = .4
Δ = .5
Δ = .6
c=1
c=.9
c=.8
c=.7
c=.6
c=.5
c=.4
c=.3
c=.2
c=.1

Ideal		Types					
A	B	☺	☹	12 A☺	6 A☹	6 B☺	12 B☹
-	+			-+--+---+++-	+-----	++-++-	++-++++++-+
+	-			++++++-++++	+--+++	-+--+-	----------++
+	-			-+++-+-+-++-	+++++-	++--+-	----+----+--
-	+			----------+-	------	++++++	+-+++-+----+
+	+			++++++--++++	+-++++	+++++-	-----++++++-
-	-			+++-+-+-+---	--+-++	+-+++-	---+-+-++-+-
+	+			-----++++++--	+-++++	++++++	++++++-+++-+
		-	+	+---+++----+	++++-+	----+-	-+++++++-+-
		+	-	+++-++++++-+	------	--+++-	---+++-++---
		-	-	-++--+-+-+--	+-+-+-	-----+	++--+-------
		-	+	----+-+++---	+++-+-	+-----	-+-+---+++-+
		+	-	---+++++++++	--+---	+++++-	+-+---+---+-
		+	+	-+++-++++-++	++++-+	-+++++	-++++--+++-+
		+	-	++-+++-+-++-	------	-+-++-	+------+---+


10000
1000
100
10
1
f(hits) - f(false alarms)
i = .1
Δ = .2
i = .1
Δ = .5
i = .2
Δ = .2
i = .2
Δ = .5
i = .3
Δ = .2
i = .3
Δ = .5
n=4
n=8
n=12
n=16
n=24
n=32
Sample size
c=1
c=.9
c=.8
c=.7
c=.6
c=.5
c=.4
c=.3
c=.2
c=.1


Learning Performance
Leading Tendency
Secondary Tendency
Number of Trials
0
k
2k
3k
4k
5k


i=.1
c=.7
c=.6
c=.5
c=.4
c=.3
c=.2
c=.1
f(Hits) - f(False Alarms)
10000
1000
100
10
1
n=8
n=16
n=24
n=32
5:3
6:2
7:1