האוניברסיטה העברית בירושלים

THE HEBREW UNIVERSITY OF JERUSALEM

CHOOSING BETWEEN ADAPTIVE AGENTS: SOME UNEXPECTED IMPLICATIONS OF LEVEL OF SCRUTINY

by

YAAKOV KAREEV and JUDITH AVRAHAMI

Discussion Paper # 436 October 2006

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/

Choosing Between Adaptive Agents:

Some Unexpected Implications of Level of Scrutiny

Yaakov Kareev and Judith Avrahami

Center for the Study of Rationality and

School of education

The Hebrew University of Jerusalem

ABSTRACT

Even with ample time and data at their disposal, people often make do with small samples, which increases their risk of making the wrong decision. A theoretical analysis indicates, however, that when the decision involves selecting among competing, adaptive agents who are eager to be selected, an error-prone evaluation may be beneficial to the decision maker. In this case, the chance of an error can motivate competitors to exert greater effort, improving their level of performance—which is the prime concern of the decision maker. This theoretical argument was tested empirically by comparing the effects of two levels of scrutiny of performance. Results show that minimal scrutiny can indeed lead to better performance than full scrutiny, and that the effect is conditional on a bridgeable difference between the competitors. We conclude by pointing out that error-prone decisions based on small samples may also maintain competition and diversity in the environment.

Choosing Between Adaptive Agents: Some Unexpected Implications of Level of Scrutiny

Although the sampling error of any statistic increases with decreasing sample size, people often make do with small-sample data even when more data could be collected easily (e.g., Burnett, Kareev, & Avrahami, 2005; Busemeyer, 1985; Fiedler & Kareev, in press; Fried & Peterson, 1969; Hertwig, Barron, Weber, & Erev, 2004; Kareev & Fiedler, 2006; Weber, Shafir, & Blais, 2004). One account of this behavior is that short-term memory capacity limits the number of items (i.e., the size of the sample) that can be considered simultaneously. According to this account, people surely would consider larger samples if they only could (see, e.g., Simon, 1990). Another account (Busemeyer, 1985; Payne, Bettman, & Johnson, 1988) focuses on the trade-off between cost and accuracy inherent in sampling: Although people realize that larger samples result in more accurate estimates (Bar-Hillel, 1979; Sedlmeier, 2005), the costs in time, effort, and lost opportunity do not justify the gain in accuracy. This latter account is consistent with recent analyses showing that for choices between alternatives, accuracy is quite high even with small sample sizes (Fiedler & Kareev, in press; Hertwig & Pleskac, 2006; Johnson, Budescu, & Wallsten, 2001). These analyses can also explain why people are often more confident in the accuracy of small-sample-based estimates than is objectively warranted (e.g., the Law of Small Numbers, Tversky & Kahneman, 1971). Finally, it has been shown that systematic biases that characterize small-sample-based estimates of correlation and variance may help in the early detection of correlations and foster an optimistic view of the world (Kareev, 2004).

Common to all these explanations is the view that, in and of itself, sampling error is undesirable, to be reduced whenever possible. Here we propose a new view of sampling error and draw attention to an implication of it that has hitherto hardly been considered: We argue that there exists a class of choice situations in which the uncertainty accompanying the use of small-sample data may affect the environment in ways that are to the sampler's benefit, rather than detriment. These desirable outcomes come about because of, rather than in spite of, the larger chance of committing an error. The class of situations we refer to are those in which sampling is conducted in order to make a performance-based choice between adaptive agents who are eager to be selected by the sampler. We propose that in such situations, the uncertainty produced by an error-prone evaluation can lead the competing agents to exert greater effort, and hence to perform better. Consequently, the sampler, although risking an occasional incorrect choice, reaps the benefits of operating in a more favorable environment—one in which the agents' overall performance is higher. Thus, the very same aspect that renders small samples a liability in the assessment of stable, indifferent alternatives may prove an asset in the assessment of adaptive agents whose performance can improve with increased motivation. Moreover, we argue that this very sampling error can maintain competition and diversity, to the further benefit of the sampler.

In the next section of this article, we lay out the rationale of the claim that the uncertainty induced by error-prone sampling may elicit greater effort on the part of competing agents who are eager to be selected by the sampler. We then present the results of three experiments designed to test implications of that analysis. We conclude by discussing further implications of the frequent use of small-sample data.

THEORETICAL ANALYSIS

The situation we consider is one in which a decision maker is to select one of a number of agents on the basis of their performance. The decision maker's utility is a monotonically increasing function of the agents' performance, and an agent's own utility is higher if selected. Examples of typical cases are choosing a service provider (e.g., a store, a broker) or choosing which employee deserves a bonus. What size sample should the decision maker employ to reach a decision? Surely, a large sample is more likely than a small sample to result in an accurate assessment, and hence to lead to the choice of the more deserving agent; small-sample data carry a higher risk of choosing the wrong agent. However, when the overall level of performance, rather than the comparison among the competitors, is of prime interest, the sampler should consider the effect of sample size not only on the accuracy of the choice, but also on the level of performance by the competing agents. To understand the effect on performance, one needs to consider that the sampling error inherent in the use of samples obviously introduces an element of uncertainty into the choice process, with the uncertainty being larger, the smaller the sample.

Decision making under uncertainty has been much studied in the past (e.g., Kahneman, Slovic, & Tversky, 1982), but the uncertainty in question has typically been that inherent in an environment indifferent to the implications of its uncertainty. Uncertainty in behavior, generated for strategic reasons, has mostly been studied in game theory in the context of interactions calling for the application of mixed strategies (e.g., Aumann & Hart, 1992–1994; for applications in psychology, see, e.g., Rapoport & Budescu, 1992). However, how uncertainty in decision making affects the behavior of agents who compete to be chosen has, to the best of our knowledge, never been considered in psychology. In contrast, such effects have been considered in economics. For example, studies of labor markets concluded that uncertainty in income increases labor supplies, or efforts (e.g., Block & Heineke, 1973). Now consider the effect that large- and small-sample scrutiny might have on competing agents' motivation to exert effort. When the evaluation of the agents' performance is based on a large sample, the true differences among them are likely to be revealed, and the truly superior agent is likely to be selected. Under such a sampling scheme, the weaker agents may see no chance of being selected, and hence no reason to try harder; realizing this, the stronger agents need not fear losing, hence they too will not try any harder. In other words, full, highly accurate scrutiny may fail to induce greater effort and an increase in performance. Now consider the possible effect of little scrutiny—the use of a small number of observations to determine the winner of the competition. Because the use of small-sample data can result in an error, weaker agents stand a chance of being selected; this being the case, they may decide to try harder, to increase the overlap in the distributions of their performance and the performance of the stronger agents. Realizing this, the stronger agents will fear losing and increase their efforts as well, so as to maintain their edge over the weaker agents. Thus, by introducing an element of uncertainty into the decision process, the decision maker may occasionally reward an agent who is not the best, but may often reap the benefits of overall better performance by all agents.[1]

A number of researchers have carried out game-theoretic analyses of this line of argument and have concluded that it is valid. The titles of the publications, such as "More Monitoring Can Induce Less Effort" (Cowen & Glazer, 1996) and "Competitive Prizes: When Less Scrutiny Induces More Effort" (Dubey & Wu, 2001), attest to the conclusions their authors reached. Similarly, Dubey and Haimanko (2003), analyzing the size of the sample a principal should take to assess the performance of competing agents, stated, "We show that the principal will do best to always choose a small sample size" (p. 1).

These analyses indicating the potential benefits of uncertainty in inducing better performance are all theoretical. They thus leave open the question whether people act as predicted by the theory. Do competing agents indeed exert more effort in the face of little scrutiny than in the face of much scrutiny? We conducted three experiments in an attempt to answer this question. In the first, we compared people's performance under conditions of minimal and full scrutiny. In the second, we tested a key assumption of the model—namely, that the effect of minimal scrutiny depends on the initial difference in performance between the competing agents. The third experiment tested whether or not the effect would also be observed when there is no competition.

EXPERIMENT 1

Experiment 1 was designed to test if level of scrutiny affects people's performance. Participants, competing in pairs, solved simple numerical addition problems. The members of each pair were paid the same amount for participation. In addition, they were promised that the person who solved more problems correctly would be paid an extra amount (a bonus).

Following a 1-min practice session, they performed the main task, which took place over six 1-min sessions. The experimental manipulation involved the level of scrutiny: In the full-scrutiny condition, the number of correctly solved problems in all sessions was assessed, whereas in the minimal-scrutiny condition, the number of correctly solved problems in only one session, selected at random, was assessed. In both conditions, the bonus was awarded to the participant who had correctly solved more problems in the session or sessions inspected.

The members of each pair were unaware of each others' ability. We expected them to assume that an overlap in their ability was likely. According to the reasoning outlined earlier, both competitors, irrespective of their ability, would exert greater effort in the face of little than in the face of much scrutiny. In other words, we expected that for the effect to occur, competitors did not have to know their standing relative to one another.

Method

Materials and Task. Participants were presented with a seven-page booklet with a blank cover page. Each of the seven pages had 42 problems involving the addition of two two-digit numbers (e.g., 27 + 56). The problems on each page were arranged in seven rows and six columns. To avoid ceiling effects, we used enough problems per page so that even skilled performers could not solve all of them within the 1 min allotted.

Procedure. Participants performed the task in pairs in a quiet room. They were recruited individually, and care was taken to ensure that the members of each pair were unfamiliar with one another. Upon entering the room, they were told that they would be required to solve correctly as many arithmetic problems as possible, that they would be paid 10 New Israeli Shekels (NIS, about $2.20) for their participation, and that the person who solved more problems correctly would be awarded a bonus of another 10 NIS. The participants were each handed a booklet and informed that it consisted of seven pages (the first of which a practice page) with simple addition problems, and that they would have 1 min to solve as many problems as possible on each page. They were also told that the number of problems on each page had been chosen so that no one could solve them all within the allotted time.

After working on the practice page for the allotted time, the participants were informed how the winner of the competition would be determined. In the full-scrutiny condition, participants were told that the experimenter would count the number of correct answers on all six test pages of the booklet. In the minimal-scrutiny condition, participants were told that the experimenter would role a die—separately for each participant—and check performance only on the test page whose number came up on the die. In both conditions, the bonus was to be awarded to the player with the larger number of correct answers. The experimenter timed performance, instructing the participants when to start and when to stop work on each page.

Participants. Participants were 80 students at the Mount Scopus campus of the Hebrew University of Jerusalem. They volunteered to participate in the experiment in return for the payment offered. Members of a pair were always of the same gender. There were 10 female and 10 male pairs in each scrutiny condition.

Results and Discussion

A comparison of the mean number of correct solutions (with score on the practice page serving as a covariate) revealed a significant difference between performance under minimal scrutiny (mean correct solutions per page = 18.82) and performance under full scrutiny (mean = 17.57), $F(1, 77) = 5.46$, $MSE = 5.64$, $p_{rep} = .93$.[2] This result is in line with the argument put forward in the introduction: Performance under minimal scrutiny was superior to that under full scrutiny.

Not surprisingly, performance was much worse on the practice page than on the test pages—mean performance was 3.02 items lower on the former than on the latter. This very large effect was undoubtedly partly due to the effect of competition and partly due to practice effects. This effect is orthogonal, of course, to our research question.

For little scrutiny to have a motivating effect, both competitors should realize that under small-sample-based, error-prone assessment, the true order of the competitors might not be revealed. For that to happen, the mean difference in ability between the competitors should be small enough for the distributions of their performance to overlap. Obviously, if the difference between competitors is too big, even the uncertainty introduced by small-sample-based assessment would not change their relative standing.[3] We therefore predicted that when competitors know each other's ability, the effect of minimal scrutiny will be sensitive to that difference, and inversely related to it. In contrast, we expected the difference in ability to have no systematic effect on performance under full scrutiny, because such scrutiny should expose the true ordering of abilities irrespective of the initial difference between the competitors. Experiment 2 was designed to test these predictions by having participants who were cognizant of each other's practice score compete under conditions of either minimal or full scrutiny.

EXPERIMENT 2

Method

The method of Experiment 2 was identical to that of Experiment 1 with one exception: In Experiment 2, after participants solved the practice page, their performance was checked and announced. Thus, each participant had some indication of his or her ability relative to that of the other.

Participants were 120 students at the Mount Scopus campus of the Hebrew University who volunteered to participate in the experiment in return for the payment offered. Participants were unfamiliar with each other, and none of them had participated in Experiment 1. The members of each pair were of the same gender; 15 pairs of each gender performed the task under minimal scrutiny, and 15 pairs of each gender performed the task under full scrutiny.

Results and Discussion

The main analysis in this experiment called for regressing test performance on the absolute difference between competitors' practice scores, to find out if participants' performance was indeed negatively correlated to the difference in their initial ability in the minimal-scrutiny condition and unrelated to this difference in the full-scrutiny condition. For the minimal-scrutiny condition, the regression had a significant negative slope: *r* = -.397, *F*(1, 58) = 10.83, *MSE* = 6.47, p_{rep} = .98. In contrast, for the full-scrutiny condition, the correlation was slightly positive, and not significant: *r* = .135, *F*(1, 58) = 1.08, *MSE* = 7.12, p_{rep} = .64.

These results, which are fully in line with our predictions, qualify and refine the minimal-scrutiny effect: For minimal scrutiny to have a motivating effect, the ability of the competing agents should be either assumed or known to be sufficiently close for the error inherent in small samples to possibly eliminate the true difference between the competitors. When the difference in ability is large, the effect is no longer evident. The results of Experiment 1 indicate that when competitors are given no information about one another's ability, they assume that their performance distributions might overlap, and do try harder under minimal scrutiny.

EXPERIMENT 3

The game-theoretic literature we discussed earlier focused on the role of uncertainty on the behavior of competing agents. It could be argued that the source of the effect was not the close competition, but the mere uncertainty about the critical session, with each session potentially being the critical one. We therefore wondered whether the effect of uncertainty would be evident in the performance of a single, noncompeting agent as well. Would an individual who is rewarded in line with the level of his or her performance also exert greater effort when assessment is based on little, rather than full, scrutiny?

To answer this question, we had participants perform the same addition problems employed in the previous two experiments, but this time participants were tested alone. Amount of scrutiny was manipulated as before, with participants in the minimal-scrutiny condition rewarded according to their performance on a single, randomly chosen page, and participants in the full-scrutiny condition rewarded according to their average performance per page.[4]

Method

The task and procedure were identical to those employed in Experiment 1, except that each participant was tested individually and that the payment was 1 NIS (about $0.22) per correct solution. Payment was determined either by performance on a single, randomly determined page (minimal-scrutiny condition) or by average performance (full-scrutiny condition).

Participants were 48 students at the Mount Scopus campus of the Hebrew University of Jerusalem. There was an equal number of males and females in each condition.

Results and Discussion

An analysis of the mean number of correct responses (with practice score as a covariate) revealed that performance in the two conditions did not differ significantly (21.53 vs. 22.32 correct responses for the minimal- and full-scrutiny conditions, respectively, $F(1, 45) = 1.04$, $MSE = 7.09$, $p_{rep} = .63$.[5]

This finding strongly suggests that it is not uncertainty in what part of the performance would be assessed that brings about better performance under minimal scrutiny. Rather, it is in the context of competition that the motivating effect of uncertainty is evident.

GENERAL DISCUSSION

The results of the first two experiments demonstrate that the use of small-sample data to determine which of two competing agents is to win a competition can bring about greater efforts and better performance by all agents than the use of large-sample-based scrutiny. Although the decision maker using small-sample data may err at times and award the reward to the less deserving competitor, that very same chance of committing an error may create an environment that is overall superior to that which will prevail if larger samples are employed.

The results of the third experiment indicate that the effect is not to be found without competition.

Let us now turn to further implications of using small-sample data as the basis for making choices. Note that the chance of preferring a weaker competitor over a stronger one—even if relatively small—increases the weaker competitor's chances of survival. Thus, unavoidable, occasional errors can sustain competition. Large-sample-based, errorless assessment leading to a consistent preference for the best competitor over the rest of the field, although desirable in the short run, would result in the elimination of the weaker competitors, leaving the sole survivor with little incentive to improve, or even to try to maintain the high level of performance that rendered him or her the winner of the competition. By using an error-prone selection process, the decision maker can avoid ending up, inadvertently, at the mercy of a monopoly.

Another, related benefit of a small-sample-based selection process is that it may help maintain diversity. The very same competitors who are relatively weak under the present conditions, and who would be eliminated and become extinct (or go out of business) under full scrutiny, may have some qualities that would render them superior under other conditions—when a different service is required or when some hidden quality (e.g., high yield in a dry year) is called for. Maintaining a diversity of resources is of paramount importance for survival. According to this line of reasoning, an error-prone selection process is an efficient mechanism for sustaining diversity. Thus viewed, limited short-term memory capacity may be a built-in, structural characteristic that forces such a process. This is not to say that organisms should choose to commit errors, because this might eliminate the motivational effect of the error-prone selection process.

What we propose, then, is that the use of small-sample data may constitute an efficient compromise. On the one hand, it ensures that, on average and in most cases, the truly superior option is selected. On the other hand, it may introduce a degree of fallibility that may help bring about a more desirable environment.

ACKNOWLEDGMENTS

Yaakov Kareev and Judith Avrahami, Center for the Study of Rationality and School of Education, The Hebrew University of Jerusalem. Work reported in this study was supported by Grant 843/03 from the Israel Science Foundation. We thank Adi Duchin and Shani Michaeli for their help in collecting the data.

REFERENCES

Aumann, R.J., & Hart, S. (Eds.). (1992–1994). *Handbook of game theory with economic applications*. Amsterdam: Elsevier Science Publications.

Bar-Hillel, M. (1979). The role of sample size in sample evaluation. *Organizational Behavior and Human Performance*, *24,* 245–257.

Block, M.K., & Heineke, J.M. (1973). The allocation of effort under uncertainty: The case of risk-averse behavior. *The Journal of Political Economy*, *81*, 376–385.

Burnett, R.C., Kareev, Y., & Avrahami, J. (2005, November). *Sampling and choice under competition for resources*. Poster presented at the annual meeting of the Psychonomic Society, Toronto, Ontario, Canada.

Busemeyer, J. (1985). Decision making under uncertainty: A comparison of simple scalability, fixed-sample, and sequential-sampling models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*, 538–564.

Cowen, T., & Glazer, A. (1996). More monitoring can induce less effort. *Journal of Economic Behavior and Organization*, *30,* 113–123.

Dubey, P., & Haimanko, O. (2003). Optimal scrutiny in multi-period promotion tournaments. *Games and Economic Behavior, 42,* 1–24.

Dubey, P., & Wu, C. (2001). Competitive prizes: When less scrutiny induces more effort. *Journal of Mathematical Economics*, *36*, 311–336.

Fiedler, K., & Kareev, Y. (in press). Does decision quality (always) increase with the size of information samples? Some vicissitudes in applying the law of large numbers. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Fried, L.S., & Peterson, C.R. (1969). Information seeking: Optional versus fixed stopping. *Journal of Experimental Psychology, 80*, 525–529.

Hertwig, R., Barron, G., Weber, E.U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.

Hertwig, R., & Pleskac, T.J. (2006). *The game of life: Frugal sampling makes it simpler*. Unpublished manuscript, University of Basel, Basel, Switzerland.

Johnson, T.R., Budescu, D.V., & Wallsten, T.S. (2001). Averaging probability judgments: Monte Carlo analyses of asymptotic diagnostic value. *Journal of Behavioral Decision Making*, *14*, 123–140.

Kahneman, D., Slovic, P., & Tversky, A. (1982). Judgment under uncertainty: Heuristics and biases. Cambridge: Cambridge University Press.

Kareev, Y. (2004). On the perception of consistency. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 259–283). San Diego, CA: Academic Press.

Kareev, Y., & Fiedler, K. (2006). Nonproportional sampling and the amplification of correlations. *Psychological Science*, *17*, 715-720.

Killeen, P.R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*, 345–353.

Knoeber, C.R., & Thurman, W.N. (1994). Testing the theory of tournaments: An empirical analysis of broiler production. *Journal of Labor Economics*, *12*, 155–179.

Lazear, E.P., & Rosen, S. (1981). Rank-order tournaments as optimum labor contracts. *The Journal of Political Economy*, *89*, 841–864.

Payne, J.W., Bettman, J.R., & Johnson, E.J. (1988). Adaptive strategy selection in decision-making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 534–552.

Rapoport, A., & Budescu, D.V. (1992). Generation of random series in 2-person strictly competitive games. *Journal of Experimental Psychology: General*, *121*, 352–363.

Sedlmeier, P. (2005). Intuitive judgments about sample size. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition* (pp. 53–71). New York: Cambridge University Press.

Simon, H.A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41*, 1–20.

Tversky, A., & Kahneman, D. (1971). Belief in law of small numbers. *Psychological Bulletin*, *76*, 105–110.

Weber, E.U., Shafir, S., & Blais, A.-R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, *111*, 430–445.

FOOTNOTES

[1] Several conditions have to be met for this analysis to hold: The competing agents must assume that there exist true differences between them in ability, that there is some within-agent variance in performance, and that the differences in ability are small enough for the resulting distributions of performance to overlap.

[2] Killeen (2005) introduced $\underline{p}_{rep}$ as a measure of replication, defined as the probability of obtaining an effect of the same sign as in the original experiment. The $\underline{p}_{rep}$ values of .88, .95, and .99 correspond to $\underline{p}$ values of .05, .01, and .001, respectively.

[3] Indeed, in tournament theory (e.g., Knoeber & Thurman, 1994; Lazear & Rosen, 1981) mechanisms—whether handicapping or matching)—are often designed to address this issue. Matching, or handicapping to achieve matching, is of course very common in sports tournaments.

[4] It should be noted that to test the effect of uncertainty on a single agent, reward must be commensurate with performance (per-piece payment). Conditioning a reward on surpassing some criterion level would turn the situation into one of competition (if the criterion is unknown) or one in which the exertion of effort levels off once a known criterion is achieved.

[5] Although overall performance was better in Experiment 3 than in Experiment 1 (21.92 vs. 18.20 correct solutions per page, respectively), this difference was entirely due to an initial difference between the two groups, as evidenced in their practice scores (18.62 vs. 15.18, respectively), which were not rewarded in either experiment. In any case, this difference in performance does not bear on our topic.