Moving from the Lab to the Field: Exploring Scrutiny and Duration Effects in Lab Experiments[*]

Kenneth L. Leonard[&]
Melkiory C. Masatu, MD[!]
February 15, 2007

Abstract

The most important issue facing contemporary experimental economists is the generalizability of results from the lab to the field. This study provides novel insights from more than 1200 doctor/patient consultations, in which scrutiny and duration of treatment were varied. In line with recent evidence in the economics literature, we find evidence that highlights the importance of scrutiny (Levitt and List, 2007), and that such effects are relatively short-lived (Gneezy and List, 2006). Such insights are fundamental to the proper interpretation of results from laboratory experiments.

JEL Classification: I1, C9, O2
Keywords: Framing Effect, Scrutiny, Experimental Economics, Social Preferences

[*] This work was funded by a seed grant from the Maryland Population Research Council (MPRC). The data were collected in collaboration with Dr. Masatu (CEDHA, Tanzania) and the assistance of R. Darabe, M. Kyande and S. Masanja. We thank A. Lange, J. List and K. McConnell for suggestions on earlier drafts.

[&] 2200 Symons Hall, University of Maryland College Park, MD 20742 kleonard@arec.umd.edu
[!] Centre for Educational Development in Health, Arusha

The validity of experimental economics as a field rests on the idea that results are generalizable; that they apply in the field as well as in the lab (Levitt and List, 2007). The extraordinary control over the decision-making environment that defines the lab is a useful tool only if this control does not, by itself, change behavior and, more importantly, the reactions of subjects to changes in incentives. If any feature of the laboratory changes the way subjects tradeoff risk versus reward, effort versus leisure, or altruism versus selfishness, then the validity of the results are suspect. In this letter we look at the regular activities of subjects in the field and examine the impact of introducing an observer to the otherwise unchanged setting. This small step from the field towards the lab significantly alters the behavior of subjects, causing them to act more socially responsible, but only for a limited time.

We observe over 1200 doctor-patient exchanges (consultations) for 108 doctors in Arusha region of Tanzania, measuring the diagnostic quality (effort) provided in the following "experimental design". The study visited every health facility in the study area. None of the doctors had advance notice of the research team's arrival. The research team explained that they were "studying health care quality", that all observations were for research only and that the data on individual doctors could not be identified by other health officials. The research team doctor then took a seat in the corner of the consultation room and observed the doctor for the remainder of the day, or until he had observed 30 consultations. The research team doctor did not interact with the patients or the doctor during this period. Quality was measured using objective checklists derived from national protocol paired to each patient's presenting symptoms.

Deleted: region

Deleted: , visiting facilities at random times. No

Effort (as measured by adherence to national protocol) increased by 20% immediately after the research team arrived, even though no other aspects of the decision-making environment changed. In particular, the doctor is in his own office, seeing his regular patients and faces no new explicit incentives to provide effort. Because doctors are responding to the presence of another doctor and doctors share a professional training, increased effort can be interpreted as an increase in professional, ethical or socially responsible behavior.[1] In a different setting, List (2006) shows a similar result; that card sellers in a laboratory setting are more pro-social than card sellers in their natural setting who do not know that they are being observed. We find that the presence of an observer alters behavior even when the setting is otherwise unchanged.[2]

Equally important, the doctor returns to pre-observation effort even while the observer remains; effort falls markedly with the passage of time and equals pre-observation effort within 1 to 2 hours (10 to 15 consultations). This short-lived increase in effort shows that (1) the presence of an observer does not change material incentives, (2) additional scrutiny does make subjects behave in a more professional or socially responsible manner, and (3) when observers are passive this display of social responsibility is short-lived.

Incentive and Inspiration driven effort

To frame the discussion of scrutiny and effort we introduce a simple model in which subjects are motivated by both pecuniary and moral considerations (where moral means that an activity has no pecuniary reward, but contributes directly to the utility function)[3]. Doctors choose effort *(a)* for each consultation (indexed *t*) and gain utility from wealth, *W(a)* and morality, *M(a)* and experience disutility from effort, *c(a)*. Both wealth and morality are increasing in effort

[1] The measure of quality is designed to be monotically increasing in professional behavior (Leonard and Masatu, 2006) and has been validated in other settings (Das, Hammer and Leonard, forthcoming).

[2] This is a classic example of the Hawthorne effect (Mayo, 1933; Benson, 2000) and has been observed frequently in health studies (see, for example, Campbell, Maxey and Watson, 1995) as well as experimental studies (Harrison and List, 2004, pg. 1034, Levitt and List (2007)).

[3] Moral behavior, in this definition, is contingent on the individual's identity or self-view (Akerlof and Kranton, 2000). Since all doctors in our study have been professionally trained, this identity is usefully summarized by the word, 'professional.'

because all doctors face some demand for quality from their employers (with implied pecuniary benefits) and because all doctors have professional education (with implied non-pecuniary benefits). Increases in the level of scrutiny (s) can increase the return to effort if doctors think the observer might affect their reputation (and therefore their wealth) or if the presence of another doctor increases the psychic benefit of professional behavior. Finally, both W and M demonstrate diminishing marginal returns to cumulative effort. Indexing consultations by t and defining the cumulative provision of effort up to consultation t, A_t as $\sum_{\tau=\underline{t}}^{t-1} a_\tau$ we have $W_t = W(a_t, s_t, A_t)$ and $M_t = M(a_t, s_t, A_t)$. The doctor maximizes utility subject to effort:

$$\begin{aligned} \underset{a}{\text{Max}}\, U_t(a_t, s_t, A_t) = & \quad M(a_t, s_t, A_t) + W(a_t, s_t, A_t) - c(a) \\ s_t = & \begin{cases} s_1 & \text{if } t \leq 0 \\ s_2 & \text{if } t > 0 \end{cases} \qquad s_1 \leq s_2 \end{aligned} \tag{1}$$

Scrutiny is never zero, but scrutiny with an observer present (*t>0*) must be at least as high as scrutiny without an observer *(t≤0*). This study focuses on the changes in optimal effort due to increases in scrutiny ($\frac{\partial a^*}{\partial s}$) and the changes in effort over time, due to increases in cumulative effort ($\frac{\partial a^*}{\partial A_t}$). With optimal effort at each point in time, we get:

$$\begin{aligned} \frac{\partial a^*}{\partial s} &= \frac{-\left(\frac{\partial^2 W}{\partial a \partial s} + \frac{\partial^2 M}{\partial a \partial s}\right)}{\frac{\partial^2 W}{\partial a^2} + \frac{\partial^2 M}{\partial a^2} - \frac{\partial^2 c}{\partial a^2}} \\ \frac{\partial a^*}{\partial A_t} &= \frac{-\left(\frac{\partial^2 W}{\partial a \partial A_t} + \frac{\partial^2 M}{\partial a \partial A_t}\right)}{\frac{\partial^2 W}{\partial a^2} + \frac{\partial^2 M}{\partial a^2} - \frac{\partial^2 c}{\partial a^2}} \end{aligned} \tag{2}$$

Assuming that the second order condition is negative, the impact of scrutiny depends on an incentive-scrutiny effect ($\frac{\partial^2 W}{\partial a \partial s}$), an inspiration-scrutiny effect ($\frac{\partial^2 M}{\partial a \partial s}$), an incentive-accumulation effect ($\frac{\partial^2 W}{\partial a \partial A}$) and an inspiration-accumulation effect ($\frac{\partial^2 M}{\partial a \partial A}$). Thus, scrutiny increases effort by subjecting the doctor to additional rewards or punishment ($\frac{\partial^2 W}{\partial a \partial s}$>0) and/or by changing the psychic benefits of effort, ($\frac{\partial^2 M}{\partial a \partial s}$ >0). In addition, effort will decrease over time if it has a diminishing marginal contribution to wealth ($\frac{\partial^2 W}{\partial a \partial A}$ <0) and/or morality ($\frac{\partial^2 M}{\partial a \partial A}$ <0). This latter affect can be described as tiring—the cost of effort increases over time—or wealth or moral satiation—the return to wealth or morality falls with increases in the stock. The presence and sign of these affects are questions for empirical investigation.

Data

In 2002 and 2003, doctors on our research team observed the consultations of 939 patients, evaluating 96 doctors on their adherence to prescribed medical protocol (Leonard and Masatu, 2005). To investigate the reaction to scrutiny directly, we also collected data from a small sample of 12 doctors in 11 facilities, interviewing 320 patients about protocol adherence immediately after their consultations[4]. Thus, we have data for three types of patients: patients who had consultations before the team arrived at the facility, patients consulted after the team arrived and whose consultations were also observed by the research team, and patients consulted after the team arrived whose consultations were not observed by the research team. For the larger sample, the observer is always present. Figure 1 shows the percentage protocol adherence for these three types of patients in the sub-sample and for all patients in the full sample, and Table 1 shows statistical tests of these patterns.

Formatted: Indent: First line: 0.5"

The dotted line in Figure 1 represents doctors who were never observed, and shows that effort is relatively flat. Column 1 of Table 1 examines the behavior of these same doctors and shows that the coefficients for changes in effort before the observer arrives at the facility (**not present t|t≤0**), after the observer arrives at the facility **(present t|t>0)** and the coefficient

[4] Leonard and Masatu (2006) show that these short-term patient responses are well correlated with adherence as measured by observers.

representing the discrete change in effort when an observer arrives (**present (yes/no))** are not significantly different from zero.

The dashed line in Figure 1 shows effort for doctors who are eventually observed. Before the researcher arrives, these doctors are similar to unobserved doctors: Column 2 shows that before the research team arrives, effort does not change with the order of observation (**not present t|t≤0**). However, when the observer arrives, effort increases significantly (**present (yes/no)).** In addition, effort falls as the observer remains; the coefficient for changes in effort after the team arrives (**present t|t>0)** is negative and significant. Importantly, adherence after a long period of observation is almost the same as adherence in the absence of an observer.

The solid line shows the behavior of doctors in the larger sample, as judged by the researchers. Because there is no data on effort before the observer arrives, we cannot document the jump in quality, but the graph shows the decline with continued scrutiny; validated in Column 3 of Table 1.

Discussion

The data show that there is a scrutiny effect ($\frac{\partial a^*}{\partial s} > 0$) and that effort declines with cumulative effort under additional scrutiny ($\frac{\partial a^*}{\partial A}\Big|_{s=s_2} < 0$) but that effort does not decline under normal circumstances ($\frac{\partial a^*}{\partial A}\Big|_{s=s_1} = 0$). The increase and subsequent decrease of effort is unlikely to be caused by extrinsic economic incentives for at least three reasons. First, doctors are told that their behavior has no financial consequences. More importantly, Leonard and Masatu (2007) show a strong scrutiny effect for doctors who work in the public sector and are therefore virtually insulated from any form of censure (they cannot be fired, and their salaries and postings are determined by seniority). Second, since both the observer and the subject are doctors with similar training and views of quality, the subject would not expect financial compensation for displaying both extraordinary and ordinary effort to the same observer. Quality is a fixed attribute of doctors, not something rewarded on a piece-rate basis. A doctor who demonstrates that he can provide high quality but chooses not to do so, is unlikely to be well regarded. Third, the fact that unobserved doctors show no decline in effort over time makes it unlikely that doctors who believed scrutiny subjected them to differential extrinsic incentives would choose to reduce their effort over time; there is no evidence that doctors get tired from treating their patients.[5]

Thus, the changes in the behavior of doctors in this sample are caused primarily by an inspiration-scrutiny effect and by a subsequent inspiration-tiring effect. Gneezy and List (2006) describe a similar pattern in fundraisers and data entry personnel in an experimental setting. Participants who receive an irrevocable gift respond by working harder, even though the extra effort is not financially rewarded. This extra effort is of short duration and gift-recipients quickly revert to ordinary effort. This behavior is not due to subjects who believe extra effort will elicit another gift. To parallel our setting, we suggest that the gift itself changes the nature of the relationship between subject and observer, creating an opportunity for the subject to increase his or her moral stock by impressing the observer.

We submit that it is not particularly surprising that participants change their behavior when observed. Given these findings the real question is, "Why do participants receive decreasing marginal returns from actions designed to impress others or to increase the stock of moral satisfaction?" If a different observer observed subjects in each interaction, we could surmise that, after having impressed a certain number of observers, the subject is no longer interested in impressing new observers. However, doctors in our sample and subjects in Gneezy and List (2006) always face the same observer. The subject appears to impress an observer for a

[5] We can examine behavior separately for physical examination, history-taking, health education and politeness. Even though doctors should get 'tired' of providing physical examination before they get 'tired' of being polite, we find no evidence of differential patterns of provision in these inputs.

short period, bank the gain in moral satisfaction (from acting the way they would like to be seen) and then return to previous behavior as if they no longer care about the observer's opinion. Specifically, doctors in our sample discard the hard won image of being professional by displaying poor quality at the end. This behavior depends on a utility function in which the subject can feel good about the fact that an observer used to have a good opinion of them and ignore the observer's current opinion.

A less complicated explanation of this behavior is that the decision-making frame changes when subjects do not receive any positive feedback for extraordinary effort or negative feedback for ordinary effort. Initially, subjects expect feedback and therefore each new consultation takes place in a slightly different experimental context; one in which the subject's belief that they will receive feedback is falling. This changing belief changes the return to effort; doctors discover that without the expected feedback, projecting a professional image has fewer direct rewards.

Do these observations call into question the nature of a laboratory experiment in which, by definition, there will always be an observer present? Specifically, does the presence of an observer fundamentally alter a subject's tradeoff between wealth and morality? We know that observers can alter subjects' behaviors by prodding them, instructing them, or otherwise suggesting what is good or bad.[6] However, we find that neutrality and passivity also alter the experimental frame. Among doctors in Tanzania, the arrival of another doctor brings to mind the ethical and professional training of medical school, no matter what the observer wants to imply. In our work, we can observe the normal behavior of doctors and it is clear that they are changing their behavior. Furthermore, the similar training of subject and researcher means that changes in behavior are easy to predict and observe.

[6] See Levitt and List (2007, pp 162) for a discussion and several examples.

In a typical experimental setting, however, experimentalists do not observe 'normal' behavior and it is not always easy to predict how the experimental frame will interact with the utility function. The fact that the experiment does not have an explicit frame does not mean that the subject does not infer one. However, our study does show that the lack of feedback can alter the frame over time, and in particular make subjects more likely to act as they would in the absence of the researcher. Thus, when changes in behavior of subjects over time can be isolated from other effects such as 'learning how to play the game', the experimentalist can infer that the impact of scrutiny is falling with time; that behavior is closer and closer to 'normal' behavior. In short duration experiments, however, subjects are likely to display far more socially responsible behavior than they would in the absence of an observer.

References

Akerlof, G.A., and R. E. Kranton, "Economics and Identity," *Quarterly Journal of Economics*, 2000, 115(3): 715-53.

Benson, P. G., "The Hawthorne Effect," in W. E. Craighead and C. B. Nemeroff, eds., *The Corsini Encyclopedia of Psychology and Behavioral Science*, 3 ed., Vol. 2, NY: Wiley, 2000.

Das, Jishnu, Jeffrey Hammer, and K.L. Leonard "The quality of medical advice in low income countries," forthcoming, *Journal of Economic Perspectives*.

Campbell, JP, VA Maxey, and WA Watson, "Hawthorne Effect: Implications for Prehospital Research," *Annals of Emergency Medicine*, 1995, 26 (5), 590-594.

DiNardo, John and Justin L. Tobias, "Nonparametric Density and Regression Estimation," *Journal of Economic Perspectives*, 2001, 15 (4), 11-28.

Gneezy, Uri and List, John A, "Putting behavioral economics to work: Testing Gift Exchange in Labor Markets using Field Experiments," *Econometrica*, September 2006, 74 (5), (1365-1384).

Harrison, Glenn. W. and John. A. List, "Field Experiments," *Journal of Economic Literature*, December 2004, 17.

Leonard, Kenneth L. and Melkiory C. Masatu, "Outpatient process quality evaluation and the Hawthorne Effect," *Social Science and Medicine*, 2006, 63 (9), 2330-2340.

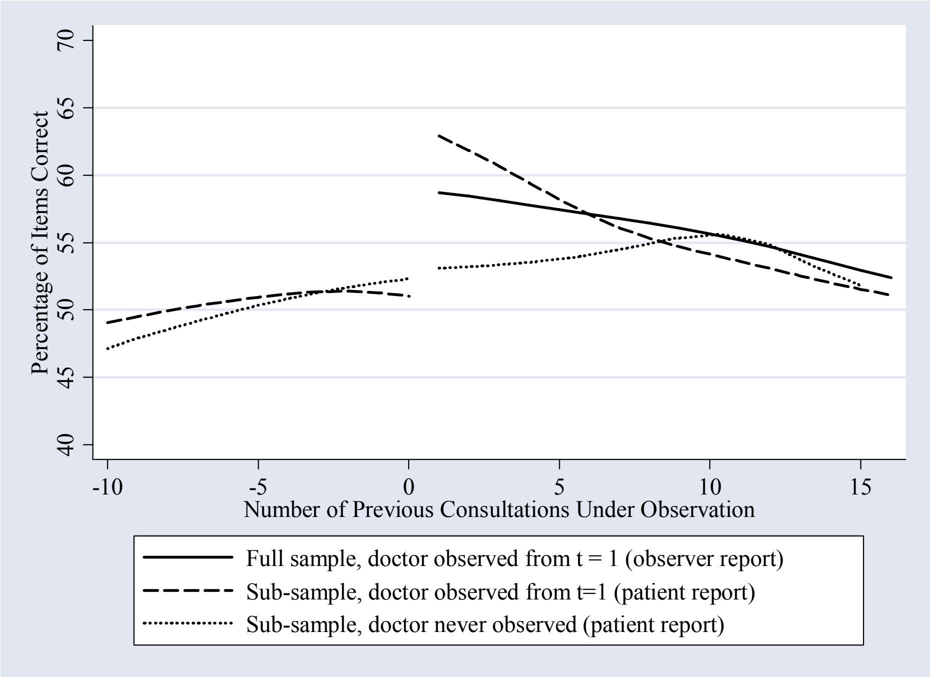
Leonard, Kenneth L and Melkiory C. Masatu, "Reexamining the gap between medical ability and practice using the Hawthorne effect," 2007, Mimeo, University of Maryland.

Levitt, Steve and John List, "What do Laboratory Experiments Measuring Social Preferences Tell us about the Real World?," *Journal of Economic Perspectives*, 21 (2) 2007.

List, John A. "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," Journal of Political Economy, 114(1) 2006.

Mayo, Elton, *The Human Problems of an Industrial Civilization*, New York: MacMillan, 1933.

Figure 1 Changes in the adherence to protocol with the order of consultation



Source: Full sample: 939 consultations observed by researchers for 95 clinicians in 39 facilities in Arusha Municipal District, Arumeru District and Monduli districts in 2002 and 2003 (Leonard and Masatu, 2005). Sub-sample: 320 patient exit interviews representing patients at 11 facilities in Arusha Municipal District (Leonard and Masatu, 2006). Results represent a semi-parametric smoothed regression following Dinardo and Tobais (2001), controlling for illness, patient and item characteristics. The mean level of compliance for unobserved doctors is adjusted downwards by 8 percentage points, because multiple doctor facilities exhibit higher average quality than single-doctor facilities and unobserved doctors are only present at the better facilities. The mean level of the full sample is adjusted upward by 2 percentage points because the full sample includes low quality rural facilities. No adjustment is made to the scale of the series.

Table 1 Patterns of quality over the order of consultations

	Dependent Variable: is the diagnostic input provided?		
	as reported by the patient		as reported by the observer
	doctor never observed	doctor observed	doctor observed
consultation order			
not present (t| t≤0)	-0.012 (0.010)	0.001 (0.012)	
present (t|t>0)	-0.035 (0.031)	-0.046 (0.007)*	-0.015 (0.002)*
present (yes/no)	0.136 (0.181)	0.479 (0.096)*	
item fixed effects	included		
doctor random effects	included		
# of possible items	1782	3343	18777
# of consultations	109	211	939

Each regression is a random effects probit regression on whether the patient recalled that the item was provided in their consultation (0/1) or whether the observer reported that it was provided (0/1) with random effects at the doctor-level and dummy variables included for each possible protocol item.

** indicates significance at the 5% level.