Journal of Risk and Uncertainty, 19:1–3; 7–42 (1999)

The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework

COLIN F. CAMERER camerer@hss.caltech.edu

Rea and Lela G. Axline Professor of Business Economics, Division of Humanities and Social Sciences 228-77, California Institute of Technology, Pasadena, CA 91125

ROBIN M. HOGARTH

Wallace W. Booth Professor of Behavioral Science, Graduate School of Business, University of Chicago, Chicago, IL 60637

Abstract

We review 74 experiments with no, low, or high performance-based financial incentives. The modal result is no effect on mean performance (though variance is usually reduced by higher payment). Higher incentive does improve performance often, typically judgment tasks that are responsive to better effort. Incentives also reduce "presentation" effects (e.g., generosity and risk-seeking). Incentive effects are comparable to effects of other variables, particularly "cognitive capital" and task "production" demands, and interact with those variables, so a narrow-minded focus on incentives alone is misguided. We also note that *no* replicated study has made rationality violations disappear purely by raising incentives.

Key words: Experimental economics, rationality, bounded rationality, judgment, incentives, experimental methodology

JEL Classification: B41, D80

I. Introduction

The predicted effect of financial incentives on human behavior is a sharp theoretical dividing line between economics and other social sciences, particularly psychology. The difference is manifested in alternative conventions for running experiments. Economists presume that experimental subjects do not work for free and work harder, more persistently, and more effectively, if they earn more money for better performance. Psychologists believe that intrinsic motivation is usually high enough to produce steady effort even in the absence of financial rewards; and while more money might induce more effort, the effort does not always improve performance, especially if good performance requires subjects to induce spontaneously a principle of rational choice or judgment, like Bayes' rule.

The effect of incentives is clearly important for experimental methodology. In addition, varying incentives can tell us something about human thinking and behavior which should interest all social scientists, and may be important for judging the effects of incentives in naturally-occurring settings (e.g., compensation in firms, or public responses to taxation).

Ultimately, the effect of incentives is an empirical question. Indeed, it is an empirical question which has been partly answered, because many studies have explored the effect of varying levels of incentive in many different tasks. In this paper we summarize the results of 74 studies comparing behavior of experimental subjects who were paid zero, low or high financial performance-based incentives.

The studies show that the effects of incentives are mixed and complicated. The extreme positions, that incentives make no difference at all, or always eliminate persistent irrationalities, are false. Organizing debate around those positions or using them to make editorial judgments is harmful and should stop.

The presence and amount of financial incentive *does* seem to affect average performance in many tasks, particularly judgment tasks where effort responds to incentives (as measured independently by, for example, response times and pupil dilation) and where increased effort improves performance. Prototypical tasks of this sort are memory or recall tasks (in which paying attention helps), probability matching and multicue probability learning (in which keeping careful track of past trials improves predictions), and clerical tasks (e.g., coding words or building things) which are so mundane that monetary reward induces persistent diligence when intrinsic motivation wanes. In many tasks incentives do not matter, presumably because there is sufficient intrinsic motivation to perform well, or additional effort does not matter because the task is too hard or has a flat payoff frontier. In other tasks incentives can actually hurt, if increased incentives cause people to overlearn a heuristic (in problem-solving "insight" tasks), to overreact to feedback (in some prediction tasks) to exert "too much effort" when a low-effort habit would suffice (choking in sports) or when arousal caused by incentives raises self-consciousness (test-taking anxiety in education).

In the kinds of tasks economists are most interested in, like trading in markets, bargaining in games and choosing among risky gambles, the overwhelming finding is that increased incentives do not change average behavior substantively (although the variance of responses often decreases). When behavior does change, incentives can be interpreted as shifting behavior away from an overly socially-desirable presentation of oneself to a more realistic one: When incentives are low subjects say they would be more risk-preferring and generous than they actually are when incentives are increased.

II. Capital, labor, and production

Take a subject's point of view. An experiment is a cognitive activity for which subjects volunteer (usually), somewhere between playing "charades" with friends at

a party and doing a neighbor's taxes for extra pocket money. Subjects come to the experiment with *knowledge* and *goals*. Earning more money is presumably one goal. Subjects surely have other goals as well: They may be intrinsically motivated to perform well, may want to appear intelligent by making quick decisions, sometimes try to amuse other subjects or fulfill the experimenter's implicit "demands," and may want to exhibit socially desirable behavior (like generosity and risk-taking).

In economic terms, we can think of a subject's goals as an objective function he or she is trying to maximize. Knowledge is "cognitive capital." The requirements of the task, which we call "production," are also important for determining performance. Psychologists ask: How well can subjects with particular knowledge, in a specific task, achieve their goals? Equivalently, economists ask: How well can subjects maximize their objective function, given available capital, and a particular production function?

Previous discussions in experimental economics have focussed almost exclusively on the objectives of minimizing effort cost and maximizing monetary reward, because economists instinctively assume thinking as a costly activity. For example, in Smith and Walker's (1993) "labor theory," subjects respond to increased incentive by expending more cognitive effort, which is presumed to reduce variance around responses. The simplest kind of labor theory rests on two intuitions: (1) Mental effort is like physical effort—people dislike both, and will do more of both if you pay them more; and (2) effort improves performance because, like scholars, subjects have access to a wide range of all-purpose analytical tools to solve experimental problems. This simple view ignores two important factors—intrinsic motivation (some people *like* mental effort, and those people disproportionately volunteer for experiments!); and the match between the analytical skills possess and the demands of the tasks they face. Effort only improves performance if the match is good. This latter omission is remedied by introducing the concepts of capital and production into the labor theory.

Capital

Cognitive psychologists distinguish "declarative knowledge"—facts about the world—from "procedural knowledge"—a repertoire of skills, rules and strategies for using declarative knowledge to solve problems. Knowing that Pasadena is northeast of Hollywood is declarative knowledge; knowing how to read a map of Los Angeles is procedural knowledge.

Experimenters are usually interested in the procedural knowledge of subjects, not the declarative knowledge. (In a sense, good instruction-writing ensures that all subjects have the declarative knowledge to understand how their decisions affect their performance.) We take procedural knowledge and "cognitive capital" to be roughly the same. 'Pieces' of capital are a variety of tricks or approaches to solving an experimental task, like the many specialized tools on a carpenter's tool belt or a

cook's knowledge of foods, kitchen utensils, and recipes. In economics experiments, cognitive capital includes heuristics like anchoring on a probability of .5 and adjusting, rules of thumb like cutoffs for rejecting ultimatum offers, analytical formulas or algorithms, personal skills or traits (e.g., unusual ability to concentrate, terrific short-term memory, perceptual skill, high 'need for achievement'), domain-specific procedures ("always wait until the end of the period to buy") and so forth.

An important feature of capital is how it is acquired. In the time horizon of a laboratory experiment, subjects probably acquire capital through learning-by-doing rather than from learning-by-thinking. As Smith (1991) wrote,

> Many years of experimental research have made it plain that real people do not solve decision problems by thinking about them in the way we do as economic theorists. Only academics learn primarily by reading and thinking. Those who run the world, and support us financially, tend to learn by watching, listening and doing (p. 12).

Furthermore, useful cognitive capital probably builds up slowly, over days of mental fermentation or years of education rather than in the short-run of an experiment (1–3 hours). (Cognitive psychologists say it takes 10 years or 10,000 hours of practice to become expert at difficult tasks; see e.g., Ericcson and Smith, 1991). However, incentives surely *do* play an important role in inducing long-run capital formation.

Production

A task's production requirements are the kinds of capital necessary to achieve good performance. Some tasks, like clerical ones, require simple attention and diligence. (Trading in markets might be like this, but includes perhaps patience and memory.) Other tasks, like probability judgments, might use analytical skill or domain-specific knowledge. Complicated games might require special analytical devices like backward induction.

Adding capital and production to the labor theory has several general implications.

First, capital variables—like educational background, general intelligence, and experience with a task—can have effects that are as strong as the effect of financial incentives (and interact with incentives). If experimenters manipulate incentives because of a prior belief that incentive effects are large, they should spend more time measuring and manipulating capital variables as well.

Second, asking how well capital is suited to the production task at hand is important because poorly-capitalized subjects may perform *worse* when incentives are stronger (just as running too fast, without proper stretching or coaching, can injure muscles).

Third, the nature of the production task matters because some tasks are simply easier (or have "floor effects")—that is, almost every subject has some kind of capital which enables good production—while others are hard ("ceiling effects"; few subjects have the necessary capital).

Fourth, elements of experimental design can affect the production function and alter performance systematically. Simplicity of instructions, stimulus display, opportunities for communication, and so forth, can affect performance and may also interact with incentives. (For example, tasks that are designed to be engaging may increase attention, lowering the "cost" of effort or raising intrinsic motivation, and reducing the marginal effect of higher financial incentives.)

Fifth, considering capital and production together implies that in some tasks, people should sacrifice short-run performance to learn better decision rules—to acquire cognitive capital useful for production—which raises a fresh empirical question of whether people sacrifice earning for learning optimally (see Merlo and Schotter, 1999).

III. A review of studies

The concepts of capital, labor, and production were introduced to provide a loose framework within which empirical effects of incentives can be understood.

The empirical heart of our paper is an informal review of 74 studies comparing behavior of experimental subjects who were not paid, or were paid low or high financial incentives, according to their performance. The studies are those we knew of and which came to our attention, so the sampling is nonrandom. However, we also sampled every article which varied financial incentives published in the *American Economic Review*, *Econometrica*, *Journal of Political Economy*, and *Quarterly Journal of Economics* from 1990–98. More careful surveys of studies were done by Bonner et al. (1996), Hertwig and Ortmann (1998), and Jenkins et al. (1998); we compare our conclusions with theirs below. Because of the opportunistic sampling we used, the reader is entitled to regard the paper as an informed essay or collection of conjectures, which may or may not prove true after a more careful meta-analysis of studies (and further research).

Studies were included if they satisfied two rules: (i) Incentive levels were reported and varied substantially within the study; and (ii) the study reported enough detail of the level of incentive and size of any performance effects to enable us to classify the effects of incentive. Thus, studies were excluded if they lacked a within-study control group or underreported details of incentive effects. As far as we could tell, subjects always knew the payoff functions they faced.

Studies satisfying the control and reporting criteria (i) and (ii) are summarized in Table 1. The studies are classified in several groups—incentives help mean performance, incentives hurt mean performance, incentives have no effect on mean performance, incentives affect behavior but behavior cannot be judged by a performance standard, and incentive effects are confounded with effects of other

Table 1. Review of experiments measuring effects on financial incentives on performance

Author (year)	Task	Incentives	Effect of higher incentives
		INCENTIVES IMPROVE MEAN PERFORMANCE	
Ashton (1990, groups 1–2)	Predicting company bond ratings (four categories) from three numerical measures of financial performance	0 vs. L ($120.96 each for top 2 of 51 S's)	Higher number of correct ratings (4.64 vs. 5.58); lower variance (3.57 vs. 1.74); feedback, written justification raised number correct too (5.55, 5.31).
Atkinson (1958)	Arithmetic, drawing tasks	L ($5.10), H ($10.20) for high score in groups of N = 20, 3, 2, or top 3 of 4	Better performance (48.37 vs. 51.96, $p < .01$ L vs. H); inverted U-shaped effect of probability of winning (48.03, 51.39, 53.21, 49.18); high "need for achievement" S's do better.
Awasthi & Pratt (1990)	Judgment problems: conjunction, sample size, and sunk cost	0 vs. L ($2.42)	Slight decline in error rate (.46 vs. .41), reduced error more for S's high in "perceptual differentiation" .44 vs. .21; more time spent in L condition (4.2 vs. 5.7 min)
Camerer, Ho & Weigelt (1997)	Dominance-solvable "beauty contest games" measuring levels of iterated of dominance	L ($1/round) vs. H ($4/round)	Slightly closer to Nash equilibrium
Castellan (1969)	Probability matching: sequential betting on independent draws of events (S's should always bet on the most likely event E, which has p(E) chance)	L (3.96 c/trial), H (39.6 c/trial)	Shift toward maximizing trials 81–180 (58% vs. 61%, p(E) = .6; 83% vs. 87%, $p < .01$, p(E) = .775)
Cooper et al. (in press)	Signaling games with output-quota ratchet effects; Chinese students/managers subjects	L (30 yuan/S) vs. H (150 yuan/S); (30 yuan = $3.75 at official FX; mgrs earn 100 yuan/day)	Higher frequency of strategic (pooling) choices (50% vs. 60%), no difference in frequency of planner "mistakes" (66% vs. 69%); effect diminished by experience, mimicked by instruction context
Drago & Heywood (1989)	Choice of decision number e in piece-rate and rank-order labor; incentive treatment is "flatness" of expected payoff function	L (8.81 c gain from e = 0 to e* = 37) vs. H (84.4 c gain)	Mean closer to prediction of 37 (48.7 vs. 37.2, round 12), lower variance (964 vs. 51, round 12).
Glucksberg (1962)	Easy problem-solving (with a helpful visual clue) and recognition of familiar words	0 vs. L (fastest 25% S's $23.58 each, fastest S $94.34)	Problem-solving: faster (5.0 vs. 3.7 min), more solutions (26 vs. 30); word recognition: faster (47.0 vs. 34.0 sec)
Grether (1980, 1992 exps 1–2)	Probability judgments of events and choice of most-likely events, based on sample information (Some use of scoring rules)	0 vs. L ($10 for correct choice)	Similar non-Bayesian patterns, but incentive S's less far from Bayesian; fewer erroneous responses (12% vs. 4%).
Harrison (1994)	Choices of gambles to test EU theory ("Allais paradox")	0 vs. L (EV = $.55–$6.45)	Small reduction in Allais paradox (35% vs.15%, conditions AP0-AP1), statistically marginal (p = .14 two-tailed z-test)
Hogarth et al. (1991)	Prediction of stochastic outcome from two cues	0 vs. L (1.16 c/point)	Higher accuracy when penalty function was "lenient" (small weight on squared error in evaluation function), means 358 vs. 288 (experiment 1)

Jamal & Sunder (1991)	Trading in commodity double auctions (treatments are incentives, trading experience of subjects, and large/small number of traders)	0 vs. L (1.16–2.32 c/point, $9.28–13.91/session)	Sharper convergence to predicted equilibrium price with incentives ($p = .003$), solely in markets with small number of inexperienced S's.
Kahneman & Peavler (1969)	Remembering digit-noun pairs	L (3.96 c) vs. H (19.8 c)	Better memory (18% vs. 55%); high incentive increased pupil dilation
Libby & Lipe (1992)	Recall and recognition of items on a list of accounting controls	0 vs. L (11.6 c/item + bonus for top 5 S's)	Better recall (9.8 vs. 120 items), no difference in recognition (15.8 vs. 16.3); more effort (1105 vs. 1281 sec)
Riedel, Nebeker & Cooper (1988)	Transferring data from hand-written questionnaires to scannable forms	0 vs. bonuses ($1.44 times 1, 2, 3, 4, 5 for exceeding 5.75/hour)	Better performance in bonus groups vs. 0 (more quantity, fewer errors, $p < .001$). No difference among levels of bonuses.
Salthouse, Rogan & Prill (1984)	Recall of digits and letters in a "divided attention" task: Two sequences, digits or letters. Total incentive 4 c/trial, incentives for each of the two sequences were (x, 4-x), x from 0 to 4.	0 vs. L (5.95 c/digit or letter)	Better recall for high-incentive sequences (20% for 0, 80% for 4), experiment 1
Scott, Farh & Podsakoff (1988)	Assembly of frame-mat puzzles	0 vs. L ($.079/assembly)	More work done (18.5 vs. 22.3, $p < .001$) (O group paid L wage but were y "surprised" and told this only after doing the task)
Siegel, Siegel & Andrews (1964)	Probability matching	0 vs. L (22.99 c if right), H (± 22.99 c)	Shift toward maximizing (70%, 77%, 93%)
Smith (1962)	Trading in double auctions	0 vs. L (23.58 c payment/trade)	Sharper convergence to competitive equilibrium prices and quantities
Smith (1965)	Trading in double auctions with excess supply (competitive equilibrium gives sellers zero surplus)	pL(4 of 27 S's get paid each period) vs. L ($54.71 surplus per period + 22.61 c per trade)	Sharper convergence to competitive equilibrium (mean deviations $2.26 vs. $2.13 period 1, 63.3 c vs. 9.04 c period 4), smaller variance in prices ($10.85 vs. 67.8 c in period 4)
Smith & Walker (1993)	Bidding in first-price private-value auctions	Five levels: $.58/auction times 0, 1, 5, 10, 20	More risk-aversion, lower variance of bids around linear (constant relative risk-aversion) bid function; effect of 1 session of bidding experience equal to incentive level of 20
Wright & Anderson (1989)	Probability judgment after being given a random "anchor"; dependent variable is difference between high- and low-anchor probabilities	0 vs. L ($371.44 total for top 45 of 77 S's)	Less effect of anchoring (.235 vs. .160); lower standard deviation (in 14 of 18 context-anchor level comparisons)
	INCENTIVES HURT MEAN PERFORMANCE		
Arkes, Dawes & Christensen (1986)	Predicting student honors from grades. (S's given formula with 70% accuracy)	0 vs. L ($.14/trial) vs. L′ ($6.99 for best of 16)	Lower accuracy (70% vs. 64% vs. 66%, control group); incentive S's used formula less, did worse
Ashton (1990)	Predicting company bond ratings (four categories); subjects given decision aid (bond rating score)	0 vs. L ($120.96 each for top 2 of 51 S's)	Lower number of correct ratings (6.38 vs. 6.04); higher variance (1.85 vs. 3.35)

Table 1. (Continued)

Author (year)	Task	Incentives	Effect of higher incentives
		INCENTIVES HURT MEAN PERFORMANCE	
Friedman (1998)	Deciding whether to switch chosen "door" in "three-door" problem (switching is Bayesian)	L (+$.40/+$.10) vs. H (+$1/−$.50)	*Less* switching at higher incentives (43.9 vs. 48.7, p = .00 to .10 in probit regressions)
Glucksberg (1962)	Difficult "insight" problem-solving (Duncker candle problem) and recognition of familiar ("church") and unfamiliar ("vignette") words	0 vs. L (fastest 25% S's $23.58 each, fastest S $94.34)	Problem-solving: fewer solutions (22 vs. 16), slower (7.4 vs. 11.1 min); word recognition: faster recognition for familiar words (47 vs. 34 sec), slower for unfamiliar words (151.9 vs. 199.8)
Grether & Plott (1979)	Choice-pricing preference reversals over money gambles	0 vs. L (EV from $2.79−7.97)	Higher rate of reversals for P-bet choices (55.9% vs. 69.7%, p = .05) (experiment 1)
Hogarth et al. (1991)	Prediction of stochastic outcome from two cues	0 vs. L (1.16 c/point)	Lower accuracy when penalty function was "exacting" (high weight on squared error in evaluation function), means 319 vs. 301 (experiment 1)
McGraw & McCullers (1979)	Set-breaking problem (Luchins water jug): nine similar problems, followed by 10th different one; dependent variable is performance on 10th	0 vs. L ($.10 + $2.07 if all answers correct)	Slower solution time on 10th problem (181 vs. 289 sec) (difference not due to extra checking time, but to slower identification of "set-breaking" solution)
Miller & Estes (1961)	Identification of visual stimuli (two faces with different eyebrows)	0 vs. L ($0.048/trial) H ($2.38/trial)	More errors in L and H than 0 (21%, 32%, 34%); no difference in response times; S's were 9-year old boys
Schwartz (1982)	Learning rules governing which sequences of lever presses are rewarded	0 vs. L ($.016/success) vs. M ($1.62 for rule discovery) vs. H (L and M)	Negative effect of trial-by-trial payoff from L, H (63% of rules discovered vs. 80% for 0, 83% M), for pretrained subjects only; no effect for inexperienced subjects (95% rules discovered); cf. Merlo and Schotter (in press).
		INCENTIVES DO NOT AFFECT MEAN BEHAVIOR	
Bohm (1994)	Choice-pricing (Vickrey-auction buying price) preference reversals over future money payments (1072 Swedish kroner in 3 mo vs. 1290 SEK in 15 mo)	0 vs. L (1/10 chance of getting preferred choice, or 10 S's in Vickrey auction, high bidder wins)	Small, insignificant reduction in overall percentage of preference reversals (choosing one claim but bidding more for another), 19% vs. 29% (Table 1, finance students only); bigger reduction in reversal for those who choose 3-month payment, 15% vs. 63%
Bolle (1990)	Ultimatum bargaining: offer take-it-or-leave-it part of amount X (prediction = offer of $.01, acceptance of any positive amount)	L (2.42 DM), H (24.2 DM), pL (p = 1/10 of 24.2 DM), pH (p = 1/10 of 242 DM)	No difference in mean offers (41%, 36%, 41%, 45%) or lowest amount accepted (38%, 33%, 28%, 32%) (note: (pL,L) and (pH,H) have some expected payoffs but (pL,H) were more similar)

Bull, Schotter & Weigelt (1987)	Choices of decision numbers (simulating effort) in rank-order labor "tournaments"	L (0–$1.96/trial), H (0–$7.89/trial)	Decision numbers "did not differ" (fn 8, no details reported)
Camerer (1987)	Trading in double auctions for risky assets where "representativeness" judgments can bias asset prices	L ($.475/asset) vs. H ($2.38/asset)	No difference in amount of bias (.09 vs. .14 bias in 1-red samples, .10 vs. .01 in 2-red samples)
Camerer (1989)	Choices of gambles	0 vs. L (expected value $6.30)	No significant differences in risk-aversion or test-retest reliability
Camerer (1990, p. 315)	Ultimatum bargaining	pL (X = $12.17) pH (X = $121.70) (p = 1/39)	No difference in offers (39% vs. 38%) or lowest amount accepted (21% vs. 15%)
Cox & Grether (1996)	Preference reversals: discrepancies between gamble choices and valuation (established using Becker-DeGroot-Marschak procedure or sealed-bid auction, or English clock descending-price auction)	0 vs. L (.5 of H) vs. H (mean $60.57 for 1-1/2 hr)	No difference between rates of predicted and unpredicted reversals using BDM (60%, 73%, 46%, period 1), small difference in second-price auction (37%, 76%, 73%), opposite difference in English clock auction (93%, 79%, 47%). repetition eliminates predicted PR in second-price auction (0, 29%, 27% in round 5), English clock auction only in L,H conditions. L and H S's more risk-averse (46% vs. 54%). No difference in intransitivity (16% 0 vs. 11% for L-H)
Craik & Tulving (1975)	Learning to remember words	L (2.76c) M (8.29c), H (16.6c)	No effect on amount of accurate recall (65%, 64%, 66%, experiment 10)
Fehr & Tougareva (1996)	Choices of wages and efforts in experimental labor market	L (up to $1/period) vs. H (up to $10/period)	No effect on average wage, worker effort, or slope of effort-wage relation (.0074 vs. .0072). Average subject income $17/month (Russians)
Fiorina & Plott (1978)	5-person committee choices of a two-dimensional point (cf. Kormendi and Plott, 1980)	L (2.3–11.6 c/unit) H ($2.32–$6.96/unit)	Marginally significant reduction in deviation of averaged data from the core (2.5 units vs. 1, p = .08 and .11 on different dimensions by Epps-Singleton test); reduction in % of points outside near-core region (p = .06); less variance (20 vs. 7)
Fouraker & Siegel (1963)	Duopoly and triopoly quantity choices (Cournot)	L vs. H ($37.27, 23.29, 9.32 bonus to top 3)	No difference in mean or variance of profits
Forsythe et al. (1994)	Ultimatum bargaining	0 vs. L ($5.36) vs. H ($10.73)	No difference in offers or lowest amount accepted; less cross-session variance; mean offers 40%, 45%, 47%
Guth, Schmitt-berger & Schwarze (1982)	Ultimatum bargaining	L (1.62 DM) H (16.2 DM)	No difference in offers or lowest amount accepted
Hey (1982, 1987)	Search: decisions about which price to accept from a sequence of prices	0 vs. L (expected £8.15)	No significant effect on amount of optimal stopping (25% vs. 33%) or apparent search rules
Hoffman, McCabe & Smith (1996a)	Ultimatum bargaining	L ($10.73/pair) vs. H ($107.27/pair)	No significant difference (contest/exchange, mean offer 31% vs. 28%, mean rejected offer 20% vs. 18%; random, mean offer 44% vs. 44%, mean rejected offer 35% vs. 40%)
Irwin et al. (in press)	Bids for $3 ticket elicited by Becker-DeGroot-Marschak method with different penalities for suboptimality	L (1 c for $1 error) vs. H (20 c for $1 error)	No significant effect on mean deviation from truthful bidding ($.62 vs. $.50) (experiment 2, full information)

Table 1. (Continued)

Author (year)	Task	Incentives	Effect of higher incentives
	INCENTIVES DO NOT AFFECT MEAN PERFORMANCE		
Kahneman, Peavler & Onuska (1968)	Mental arithmetic: remembering four-digit strings and adding 0 or 1 to each digit	L (8.26 c) H (41.3 c)	No effect on accuracy (88% vs. 82%); increased pupil dilation in easier add-0 condition
Loomes & Taylor (1992)	Choices over 3 gamble pairs (are there regret-induced intransitive cycles?)	0 vs. L (EV = £4.22)	No difference, 21.6% cycles vs. 18.5% cycles
McKelvey & Palfrey (1992)	Choices in multi-stage "centipede games"	L ($.20-$) vs. H ($60−$24)	No significant difference (.06 vs. .15 equilibrium taking at first node)
Neelin, Sonnenschein & Spiegel (1988)	Sequential bargaining: subjects alternate offers dividing a "shrinking pie" of size X across five rounds	L (X = $6.56) vs. H (X = $19.69)	No difference in mean percentage of X offered (34% vs. 34%) or mean offer rejected (26% vs. 30%)
Nilsson (1987)	Recall and recognition of words	0 vs. L ($13.58 for best S, n = 10)	No difference in recall (35% vs. 33%) or recognition (58% vs. 55%); incentive S's self-reported working harder.
Roth et al. (1991)	Ultimatum bargaining	L ($11.60/pair) vs. H ($34.79/pair)	No difference in ultimatum games (median 48−50% in rounds 1, 10 for both L,H); small, insignificant difference in "market" games with 9 proposers competing (median 58% vs. 78%)
Samuelson & Bazerman (1985)	Buyers bidding against an informed seller ("acquire-a-company" problem)	0 vs. L (+$7.19 to −$14.39)	No effect on median, modal bid (= 50, version 3, compare Figs. 3, 10a).
Siegal & Fouraker (1960)	Buyer-seller bargaining over price-quantity pairs (incentive is differential between Pareto-optimal and adjacent-quantity outcome)	L (48.5 c−77.65 c difference) vs. H ($2.91)	No significant difference in mean profit (266.92 c vs. 43.68 c). much lower variance (2426 c vs. 92.21 c)
Straub & Murnighan (1995)	Ultimatum offers and acceptance thresholds, complete and partial information (responders do not know pie size)	pL ($10.47 pie) times 1,3,5,8,10 (Prob of playing p = 2/1813)	No significant difference in mean offers (31% to 26% for multiplier 1 to 10) or mean acceptance thresholds (19% vs. 20%)
Wallsten, Budescu & Zwick (1993)	Probability judgments using numerical or verbal expressions	0 vs. L (total $20 bonus for top 4 of 7 S's)	No significant difference in accuracy (measured by an incentive-compatible spherical scoring rule); some difference in positive (P) and negative (N) conditions (P, N guarantee > 0, < 0 payoffs for stating probability .5)
Weber et al. (1997)	Trading risky assets in double auctions with different endowment conditions (long, neutral, short)	0 vs. L (EV .22 DM/unit) vs. H (EV 2.20 DM/unit)	"No difference in market prices" (p. 17)

INCENTIVES AFFECT BEHAVIOR, BUT NO PERFORMANCE STANDARD

Battalio, Jiranyakul & Kagel (1990)	Choice of gambles		More risk-averse
Beattie & Loomes (1997)	Choice of gambles	0 vs. pL (1/4 chance) vs. L (EV £3 to £7.81)	No difference (q's 1–3); no difference in "common ratio effect"; more risk-aversion (q 4; 36%, 22%, 8%)
Binswanger (1980)	Choice of gambles (by poor Indian farmers)	0 vs. L (EV 86.54 rupees = .2% of average S's wealth)	More risk-aversion at higher stakes, .86, 8.65, 86.54 rupees; no difference in mean risk-aversion, hypothetical vs. real 86.54 rupees (p. 398); hypothetical choices more dispersed
Cubitt, Starmer & Sugden (1998)	Choice of gambles	0 vs. L (EV from £2.5 to 12)	More risk-averse (50% vs. 60%, groups 3.1–3.2)
Edwards (1953)	Choice of gambles (p, $X) with pX held constant	0 vs. H (pX = 53c, 0, − 53c)	More risk-seeking. Larger deviations from EV and EU maximization for bets with pX = 53c or 0; fewer intransitivities
Forsythe et al. (1994)	Dictator "games" (one person divides $10.73 between self and other)	0 vs. L ($5.36) vs. H ($10.73)	More self-interested offers (50% offer half and 15% offer 0 vs. 20% offer half and 35% offer 0); means 48%, 28%, 23%
Grether & Plott (1979)	Choice of gambles	0 vs. L (EV from $2.79 to 7.97)	More risk-seeking ($p < .01$) (experiment 1)
Cummings, Harrison & Rutström (1995)	Choice of whether to buy consumer products	0 vs. L (hypothetical vs. actual purchase)	Fewer purchases (9% vs. 31%).
Hogarth & Einhorn (1990)	Choice of gambles	L (EV = $.12) vs. H (EV = $12.10)	More risk-averse (40% vs. 73% gains, 24% vs. 36% losses) (experiment 3)
Irwin, McClelland & Schulze (1992)	Vickrey auctions for insurance against risky losses	0 vs. L (.011, − $45.06)	More risk-averse: median bids higher, fewer zero bids and very high bids
Kachelmeier & Shehata (1992)	Choice of gambles (Canadian and Chinese students: 67.58 yuan = monthly income)	L (1.13 yuan) vs. H (11.26 yuan)	More risk-averse (certainty-equivalents higher than expected value for L incentive); both L and H overweight low winning probabilities
List & Shogren (1998)	Valuations of received Christmas gifts	0 vs. L (4th-price Vickrey auction for 244 gifts)	Higher valuations in actual auction ($96 vs. $137)
Sefton (1992)	Dictator games	0 vs. pL (1/4 chance of playing) vs. L ($5.63/pair)	More self-interested offers in L condition, means $2.15, $2.06, $1.23 (0, pL the same, both significantly different from L)
Schoemaker (1990)	Choice of gambles	0 vs. pL (7/242 chance of playing EV = $60.48, − 60.48)	Slightly more risk-averse (75% vs. 77% gains, 23% vs 34% losses, $p = .20$)
Slonim & Roth (1998)	Ultimatum bargaining	L ($1.90), M ($9.70), H ($48.4)	Rejection rates of *percentage* offers lower with increased stakes (44%, 19%, 13% for offers 25–40%), p (M vs. L) = .04, p (H vs. L) = .002; bargaining in Slovak crowns (60, 300, 1500)
Slovic (1969)	Choice of gambles	0 vs. L (EV $7.12)	More risk-averse ($p < .01$)

Table 1. (Continued)

Author (year)	Task	Incentives	Effect of higher incentives
	INCENTIVE EFFECTS ARE CONFOUNDED WITH EFFECTS OF OTHER TREATMENTS		
Bahrick (1954)	Learning names of geometric forms; peripheral learning of colors is worse (incentive paid for form learning only); 0 subjects told not to try hard	0 vs. L (max $8.55)	Faster learning of forms (16.9 trials vs. 19.6); worse learning of colors (6.1 vs. 7.6)
Baumeister (1984)	Physical game of hand-eye coordination: moving two rods so a ball rolls between them, then dropping the ball into a slot (confound between incentive and stated performance goal)	0 vs. L ($1.49/trial)	Worse scores in first trial (33.6 vs. 28.3), same scores in second trial (34.1 vs. 33.2); no variances reported.
Eger & Dickhaut (1982)	Posterior probability judgment; searching for evidence of "conservatism" in Bayesian updating (confound between incentive and elicitation technique; no-incentive S's give odds, incentive S's pick "payoffs tables" which E bets against)	0 vs. L ($64.66, $40.41, $24.25 for top 3 S's in each group of 9–10)	Reduced judgment error (conservatism), measured by slope of log odds against log Bayesian odds (.63 vs. 1.04); less error in accounting vs. abstract context
Fouraker & Siegel (1963 exps 3–5)	Buyer-seller bargaining over price and quantity (confound between incentive and experience: H trial was 21st of 21 trials)	L vs. H	No difference in mean prices or quantities; variance 1–4 times smaller
Kroll et al. (1988)	Investing in risky assets (confound between higher incentive and risk-aversion)	L vs. H	More risk-averse, closer to optimal portfolio, work harder
Phillips & Edwards (1966)	Posterior probability judgment (confound between incentive and use of two proper and one improper scoring rules)	0 vs. L ($.44 max/trial)	Lower Bayesian errors (30% lower); lower cross-S variance (.005 vs. .012)
Slovic & MacPhillarny (1974)	Multicue prediction with missing values (e.g., admitting students who each have two test scores, with only one test in common) (confound between incentive and trial-by-trial feedback, missing cue distribution information)	0 vs. L (max $12.18)	No difference in fraction of S's weighting common test more heavily (70%, 77%)
Wright & Aboull-Ezz (1988)	Judgments of probability distribution of GMAT, age, and salary of MBA students (confound between incentive and use and explanation of scoring rule)	0 vs. L ($157.48 total for 10 best S's, n = 51)	Lower mean squared error (.007 vs. .004); lower cross-S variance (half in L)

treatments. The table reports the authors, task, and incentive level in each study. The rightmost column summarizes the effects of the incentive levels given in the third column. The table was constructed by the first author and every entry was checked by a research assistant.

An example will show how to read the table. In the Awasthi and Pratt (1990) study there were two performance-based incentive levels, denoted 0 and L. Zero means choices were hypothetical so there was no performance-based incentive (usually subjects were paid a few dollars for participating). L means the stakes were low (they were paid $2.42 for answering each judgment problem correctly); H denotes higher stakes. For comparability, all payments were inflated to 1997 dollars using the GDP deflator.

Note that L and H denote lower and higher levels within a study, not absolute levels. The absolute levels are generally reported as well. This reporting convention does make it difficult to compare across studies, since the L level in one study may, in absolute terms, be higher than the H level in another study. However, this convention does make it possible to tell whether, in general, raising stakes from L to H improves performance (regardless of what those levels are).

The table reports that the fraction of subjects making errors was 46% in the 0 condition and 41% in the L condition, so higher incentives reduced error slightly. The fourth column also notes that L subjects took more time to complete the task (5.7 minutes instead of 4.2), and the reduction in error caused by higher incentive, from 44% to 21%, was greatest for subjects who were high in "perceptual differentiation" (measured by a psychological test).

Rather than reviewing each study, we will describe some regularities in the several categories of results, which are summarized in Table 2.

When incentives help

There are many studies in which higher incentives do improve mean performance. Table 2 suggests that incentives appear to help most frequently in judgment and decision tasks (they also sometimes hinder performance in this class of tasks). They improve recall of remembered items, reduce the effect of anchoring bias on judgment, improve some kinds of judgments or predictions, improve the ability to solve easy problems, and also sharpen incentives to make zero-profit trades in auctions or do piece-rate clerical work.

An example is Libby and Lipe (1992), who studied recall and recognition of 28 internal firm controls which accountants might look for when auditing a firm (e.g., "spoiled checks are mutilated and kept on file"). Subjects then had to recall as many of the controls as they could (in the "recall" task) or recognize controls seen earlier, on a new list which included some spurious controls (in the "recognition" task). Some subjects were paid a flat fee ($2) for participating (the 0 condition) and others earned 10 (11.6 cents in 1997) cents for each item correctly recalled or recognized, along with a $5 bonus for each of the top five subjects. Incentives

Table 2. The number of studies exhibiting various incentive effects

Type of task	Helps	Has no effect	Hurts	Has an effect, but no performance standard
Judgments and decisions				
Probability judgment	3	2		
Binary choice (including "three door" problem)	2		1	
Multivariate prediction	2		4	
Problem solving	2		2	
Item recognition/recall	3	3	1	
Clerical (drawing, data transfer, assembly)	3			
Games and markets				
Dominance-solvable games	1			
Tournaments	1	1		
Signaling games	1			
Sequential bargaining		2		
Ultimatum games		6		1 (fewer rejections of fixed-% offers at higher stakes)
Trust games (labor markets, centipede)		2		
Auctions: double	3	1		
Auctions: private value	1			1 (Vickrey for gifts, higher valuations)
Auctions: common value		1		
Spatial voting		1		
Duopoly, triopoly		1		
Individual choices				
Dictator tasks				2 more self-interested
Risky choices		3		8 more risk-averse, 2 more risk-seeking
Non-EU choice patterns	1	1		
Preference reversals		2	1	
Consumer purchases				1 fewer actual purchases
Search (wages)		1		

caused subjects to work harder (about 3 minutes longer). Incentives also caused subjects to recall more items correctly (12.0 vs. 9.8) but did not improve recognition much (16.3 vs. 15.8). Libby and Lipe suggest that incentives do induce more effort, but effort helps a lot in recalling memories, and only helps a little in recognizing an item seen previously. Their study is a glimpse of how incentive effects can depend dramatically on the kind of task a person performs.

Kahneman and Peavler's (1969) study is notable because it measures a physical manifestation of the effort induced by higher incentives—pupil dilation. Their subjects learned a series of eight digit-noun pairs from an audiotape (e.g., "3-frogs").

Then subjects were told a digit (e.g., 3) and asked to say which noun had been paired with that digit. For some digits, subjects had a low incentive to guess the noun correctly (1 cent) and others had a high incentive (5 cents). When subjects were told the incentive level on each trial, their pupils dilated (they grew wider in diameter). When incentives were high dilation was larger (pupils changed in diameter from 3.99 millimeters to 4.04) than when incentives were low (3.97 to 3.98). The difference in the amount of dilation in the low and high incentive conditions is tiny but highly significant ($t = 3.2$, $p < .01$). High-incentive subjects also got more nouns correct (55%) than low-incentive subjects (18%).

A simple count of studies in which incentives affect average behavior (versus those in which incentives don't matter) shows that a disproportionate number of effects result from raising the level of incentives from 0 (i.e., subjects choose hypothetically and are paid no performance-based incentive) to a low level L. Raising incentives from some modest level L to a higher level H is more likely to have no effect. This suggests that while adding some incentive to otherwise-hypothetical choices often matters, experiments which then multiply stakes by 2, 4, or 20 do not produce similar boosts in performance. It is too early to call for an end to such (expensive!) experiments but the results in Table 1 suggest little reason to think the effects of very large incentives will be substantial.

When incentives hurt

In a few tasks, incentives appear to actually hurt. All of these are judgment or decision tasks. Many of the studies establishing these negative effects are likely to be controversial, and the effects are often unclear for various methodological reasons. (Researchers itching to study incentives empirically might start by trying to replicate some of these results.)

A striking example is Arkes, Dawes and Christensen (1986). Their subjects were told grades for each of 20 students and were asked to predict whether the students won honors. In one condition, students were given a simple formula for predicting honors from grades, which was right 70% of the time. (Students were told how accurate the formula was, and were warned that outpredicting the formula is difficult.) No-incentive subjects generally used the formula and got 66% right. Incentivized subjects, paid \$.10/trial (\$.19 in 1997\$), tended to abandon the formula and actually got fewer right (63%). While their effort was not measured directly, one can interpret the incentivized subjects' abandonment of the simple formula as an exertion of effort; but their extra effort hurt performance, rather than improved it.

Ashton (1990, groups 5–6) got the same result in a similar setting, prediction of bond ratings. This phenomenon is related to the fact that experts in many domains —law, medicine, graduate admissions, psychiatry—make worse predictions than simple formulas based on observable, quantitative predictors (see Dawes, Faust and Meehl, 1989, for a review of nearly a hundred field studies). In these domains

formulas require little effort and predict well. Increased incentives cause people to exert more effort, adding their own judgment to the formula (or ignoring it), leading to predictions which are often worse. In terms of capital and production, these sorts of judgment tasks require simple calculations focussing on only a few cues. When "too much capital" is used, it backfires.

Hogarth et al. (1991) found that when subjects were stiffly penalized for forecasting inaccurately in a two-variable "multicue learning" task, the effect of incentives was to encourage more experimentation which lowered overall performance. In two studies on 'insight' problems like the Luchins water-jug task, Glucksberg (1962) and McGraw and McCullers (1979) found that subjects were slower to have the insightful experience which gave a correct answer if they were paid. Since these problems require subjects to 'break set' and think unorthodoxly to find the answer, the negative effects of incentives means highly-incentivized subjects may be exerting more effort, but more effort blinds them to the surprising answer.

Incentives might also hurt when added incentives make people self-conscious about an activity which should be automatic (though no studies in Table 1 use these tasks). The phenomenon appears as "choking" in sports (professional basketball players sink significantly fewer free-throw shots in high-pressure playoff games than in regular-season games; see Camerer, 1998), and test-taking anxiety in education (see Baumeister, 1984), and can be traced to Yerkes and Dodson (1908).

When incentives make no difference

The most common result is that incentives did not affect mean performance. These include studies on market trading, bargaining, and some studies of risky choices.

Incentives appear to not matter when the marginal monetary return to increased effort is low. Effort returns will be low when it is either very easy to do well, or very hard to improve performance (known in psychology as "floor" and "ceiling" effects). For example, in bargaining, Camerer (1990), Forsythe et al. (1994), Guth, Schmittberger and Schwarze (1982), Neelin, Sonnenschein and Spiegel (1988), and Siegel and Fouraker (1960) found no substantial differences in average behavior. Think of bargaining behavior as a simultaneous expression of a person's degree of self-interest (or oppositely, an expression of fairness or altruism) and a person's understanding of their bargaining power in a particular situation. In making alternating offers for division of a "pie" that shrinks with each rejected offer, for example, people may make nonequilibrium offers because they are not purely self-interested, or because they cannot compute the equilibrium offer. Incentives probably make little difference in these experiments because they do not substantially alter either the degree of self-interest or a subject's understanding. The game-theoretic solutions to these games are either so transparent (a "floor," in the case of ultimatum bargaining) or so difficult to figure out (a "ceiling" for sequen-

tial bargaining requiring backward induction) that only specific training will induce equilibrium offers (in the case of multi-stage bargaining).

Floor and ceiling effects are common in other tasks where incentives make little difference. For example, Kahneman, Peavler and Onuska (1968) studied pupil dilation and performance in a task where subjects heard four-digit strings, then repeated back the string, adding either 0 or 1 to each number. They found that pupils dilated more when incentives were higher—a sign that subjects were working harder—but there was no increase in accuracy because subjects got 88% right even with low incentives (i.e., performance was close to a ceiling at 100% accuracy). Samuelson and Bazerman (1985) found the opposite in a study of bids in a notoriously difficult "acquire-a-company" problem. Bidding for real money did not improve performance (but did raise the variance) because discovering the optimal bid is extremely difficult.

It is worth noting that in many experiments, financial incentives might appear to have little effect because subjects are intrinsically motivated to perform well, so money adds little extra motivation. When subjects volunteer, for instance, they surely self-select for high intrinsic motivation. In extrapolating results to nonvolunteer populations, like students who are essentially forced to participate for course credit or survey respondents approached in malls or called at home, one should be careful to generalize from the results of experiments in which subjects volunteer.

In many of the studies where incentives did not affect mean performance, added incentives *did* reduce variation (Grether, 1981, noticed this fact early on). For example, Fiorina and Plott (1978) studied five-person committees choosing a point in a two-dimensional policy space. Each subject earned an amount of money which depended on how close the committee's point was to the point they preferred. Subjects in the committees earned 1–5 cents (low incentive) or $1–3 (high incentive) for every unit that the committee's point was closer to their preferred point. High incentives did not change the mean deviation from the core point predicted by cooperative game theory very much, but did reduce variance around the core point dramatically. Similarly, Irwin et al. (in press) found that higher incentives in the Becker-DeGroot-Marschak method for eliciting valuations did not affect the mean value elicited, but did reduce the standard deviation by half.

When incentives affect behavior, but there is no performance standard

There are quite a few studies in which incentives do affect behavior, but there is no normative standard for optimal behavior so one cannot pass judgment on whether incentives "improved" performance per se. About half these studies involve choices among gambles. In three studies incentives had no effect on risk attitudes. When there was an effect, with one exception (Edwards, 1953), the effect of actually playing gambles was to make subjects more risk-averse (see also Weber, Shafir and Blais, 1998, for a meta-analysis with the same conclusion). In studies with "dictator games"—players dictate an allocation of a fixed sum between themselves and

another subject—subjects usually kept substantially more when choices were real rather than hypothetical. Finally, there are a large number of studies comparing hypothetical choices to buy everyday products with actual choices. Only one study is included in our sample (Cummings, Harrison and Rutstrom, 1995; but see Harrison and Rutstrom (in press) for a review of forty studies, mostly in environmental valuation). In their study, subjects were asked whether they would buy a juicer, chocolate, or a calculator. About three times as many subjects said they would buy, as actually did (31% vs. 9%). Overreporting purchase intention is quite familiar in marketing studies, and in political science (people overreport both intentions to vote, and whether they actually did vote).

A related example is probability matching in binary learning experiments. In these experiments, in each of many trials subjects bet on which of two lights (say, red or green) will light up. Suppose the red light comes on 60% of the time, and each trial is independent (though subjects usually don't know that). Then the profit-maximizing strategy is to always bet red, but subjects typically choose red between 60% and 100% of the time, roughly matching the relative frequency of choosing red with the probability of red. When incentives are raised, subjects move toward the profit-maximizing prediction, choosing red more often (Siegel, Siegel and Andrews, 1964; Castellan, 1969). This behavior can be explained by a model in which subjects find the task boring (it is!) and therefore get utility from varying their response, or get added utility from winning a bet on the less likely underdog color (green). As incentives are raised, subjects consume less variation and earn more profit, accepting some boredom in exchange for more money (see Smith and Walker, 1993).

In all these cases, we can interpret subjects as having some nonfinancial goal—to appear risk-taking (gambles) or generous (dictator games), to please the experimenter by intending to buy something (purchase experiments), or avoid the boredom of making the same choice hundreds of times (probability matching)—which is partially displaced by profit-maximization when incentives are increased. This kind of incentive effect is fundamentally different from the effect of incentives in inspiring greater effort, clearer thinking, and better performance.

When incentives are confounded with other treatments

Table 1 includes a few studies which confounded incentives with another treatment variable so that it is impossible to tell whether financial incentive, or the confounded variable, caused a change in performance. In some cases confounds are deliberate; for example, in exploratory designs on market experiments, investigators often adjust "exchange rates" for converting points to money, and confound those changes with simultaneous changes in parameters. Table 1 reports only cases where confounds appear to be unnoticed. We cannot draw conclusions from these studies, but we include them for completeness and to caution experimentalists who are interested in studying incentive effects about the need for proper control. For

example, Wright and Aboull-Ezz (1988) had students judge probability distributions of GMAT scores, age, and starting salaries of recent MBAs. Students in the incentive condition were paid according to an incentive-compatible scoring rule. No-incentive subjects were not told about the scoring rule. The incentivized subjects did have lower absolute errors in probability than the no-incentive subjects (.04 vs. .07), but the difference could be due to the scoring rule rather than to financial incentives per se. (To break the confound, a control group which are given scoring-rule feedback about their judgments but not given any incentive for accuracy, and a control group which is incentivized but given no scoring rule, could be compared to the first two groups.)

In Kroll, Levy and Rapoport's (1988) study of portfolio allocation, increased incentives may have increased subjects' risk-aversion, which may explain why the high-incentive subjects chose portfolios which are closer to optimal. (The optimal portfolio contained a healthy proportion of the least risky asset.) This example is particularly disturbing because their study is prominently published and has been cited as evidence that higher incentives produce better performance.

What others have said

Our paper is closely related to four others. (Very interested readers should read all four.) Smith and Walker (1993) present a formal "labor-theoretic" framework, and argue from a sample of 31 studies that increased incentives tightens the distribution of errors around the theoretical optimum. While increased incentives do seem to reliably reduce variance, we argue that the effects of incentives are perhaps more complicated than that, and add capital and production (informally) to the central solo role that effort plays in their framework.

Many of our basic conclusions were arrived at independently by Bonner, Young and Hastie (1996), who conducted a more thorough review of a wider array of research. Their review classifies results according to five types of incentive schemes —flat rates (no performance-based incentive), piece rates, variable rates (stochastic piece rates), quota systems, and tournaments.

They find little systematic difference among these types of incentive. They find frequent positive effects in domains where little skill is required and effort improves performance—pain endurance, vigilance or detection (e.g., spotting typos), and clerical or production tasks. They find weaker evidence for positive effects in memory and judgment or choice tasks, and essentially no positive effects in problem-solving. Bonner, Young and Hastie also highlight the important role of skill (or capital, in our terms), calling it "the most important, yet neglected moderator of the effects of incentives on performance" (p. 40).

Hertwig and Ortmann (in press) include a small discussion of incentive effects in a paper contrasting experimental practices in economics and psychology (cf. Camerer, 1996). Their paper includes a census of available studies (10 in number) from 1987–97 of the *Journal of Behavioral Decision Making*, and uses a standard

meta-analytic measure of effect size (eta) to permit comparison across studies. They conclude that increased incentives almost always have a modest effect, and call for "learning more about the specific conditions under which payoffs improve, do not matter to, or impair task performance, and investigating how payoffs (and opportunity costs) affect decision strategies and information processing."

Jenkins et al. (1998) sampled all studies in several applied psychology journals from 1975–96 which reported detailed individual-level effects of monetary incentives (with control groups). They found 47 studies and combined the results in a formal meta-analysis. Forty-one studies measured the effect of increased pay on output ("performance quantity"), generally in mundane clerical tasks such as assembling erector sets or coding items. Most studies found significant increases in output from higher incentive. Only six studies measured the quality of performance, and the effects of increased incentive in those studies are quite weak. They also found that the level of intrinsic motivation in the task did not seem to affect the size of the incentive effect, and that simple laboratory studies *understated* incentive effects, relative to richer laboratory simulations or field studies.

Applying the capital-labor-production metaphor

The capital-labor-production metaphor points naturally to several features of cognitive capital and production requirements which, in turn, suggest interesting new classes of experiments. (By contrast, the pure labor theory suggests only that raising incentives may produce different distributions of errors.) We mention four categories: capital-labor substitution, capital formation, task design, and capital transfer.

Capital-labor substitution. Capital and labor are substitutes in most physical production processes. Similarly, cognitive capital and effortful thinking are productive substitutes in some tasks. An example is the stagecoach problem: Find the least-cost series of nodes which connect an initial node to a destination. People can solve problems in this class labor-intensively, by enumerating all possible paths and choosing the lowest-cost one. If they know the dynamic programming principle (i.e., they have that principle in their stock of cognitive capital) they can substitute capital for labor by working backward from the destination. A high level of capital and little labor will produce an answer as cheaply and accurately as a low level of capital and lots of labor.

A familiar, general example of capital substituting for labor is experience of subjects. Several studies compare the effects on performance of experience with financial incentives. For example, Jamal and Sunder (1991) find that both experience and financial incentive increase convergence to competitive equilibrium in experimental commodity markets, and experience has a more statistically reliable effect. Smith and Walker (1993) estimate that the effect of one session of experience on the convergence of first-price auction bids around the (risk-neutral)

Nash bidding function is about the same as the effect of multiplying a base incentive by ten. Cooper et al. (in press) were the first to suggest (based on their observations) that higher pay may substitute for learning in games where learning effects are large. Notice that this insight cuts both ways: It implies that paying subjects more may enable experimenters to induce faster learning (or better thinking), speeding up the rate at which subjects master tasks and permitting more complex designs. But it also implies that even poorly-motivated subjects may learn to perform well with enough learning opportunity. In any case, a more thorough exploration of experience versus incentives, going beyond the bounds of this paper, would certainly be useful.

Another example of capital-labor substitution is the effect of giving contextual labels to subjects' choices. Contextual labels enable subjects to activate domain-specific heuristics or choice rules (e.g., Sniezek, 1986). For example, logic problems like the "Wason 4-card problem," which require subjects to recognize that P → Q is logically equivalent to not-Q → not-P, are much easier for subjects when placed in a familiar, practical context (Cheng and Holyoak, 1985), particularly one which correspond to detection of cheating (Cosmides, 1985). In economics experiments, Eger and Dickhaut (1982) report that accounting students did substantially better in a probability judgment task (roughly equal to the improvement from higher incentive) when abstract labels were replaced with an accounting context. Cooper et al. (in press) did a study of signaling games with 'ratchet effects,' in which a productive firm manager who reports high output is penalized by having an output quota ratcheted upward in the future. Using Chinese subjects (some of whom were firm managers), they found that when contextual labels described the game actions as production, quotas, etc., subjects learned some features of the pooling equilibrium more rapidly. Natural labels are largely unexplored by experimental economists, mostly out of fear that natural language creates a non-monetary utility for making choices which loosens control over incentives (e.g., fewer subjects might choose "defect" in the prisoner's dilemma than would choose a strategy blandly labelled "D" or "strategy 2"). Natural labelling certainly does run this risk, but it might also enable subjects to use cognitive capital, reducing response error and speeding up learning.

Capital formation. The capital metaphor suggests that nonfinancial determinants of capital formation might be interesting to study. Three examples are between-session "learning," communication, and instruction.

Experimental economists suspect that something important occurs between experimental sessions: Subjects "digest" their experimental experience, perhaps talk to other subjects, and articulate what they did and saw to friends who did not participate. Much of this learning may be "implicit," meaning that subjects are learning things they are not aware of (which is a well-documented phenomenon in cognitive psychology, e.g., Reber, 1989). This capital formation takes place entirely outside the lab, and is therefore beyond the control and measurement of the experimenter, but some features of the process could be measured (e.g., by

unobtrusively observing or recording subjects as they discuss an experiment during a planned break between sessions).

In most experiments, communication is restricted on the grounds that it is unrealistic, may influence social values, or weakens control over a subject's information. But if learning from others (and from 'teaching' others) are ways of building capital, and one is interested in capital-labor determinants of performance, then communication becomes a particularly interesting variable. For example, allowing subjects to work in teams would, for some tasks, be an interesting treatment variable.

Experimental instructions are unquestionably an important influence on capital formation. Experimental economists usually try to write extremely simple and clear instructions as a kind of optimal task design (see below). In some cases, however, simply instructing people about decision rules—supplying capital—is one way to measure whether those rules are used instinctively. For example, Camerer et al. (1993) were interested in whether subjects used backward induction in bargaining. One way to answer this question is to instruct some subjects about backward induction and see whether they behave differently than uninstructed subjects. They do. The difference is evidence that the backward induction analytical device was not part of uninstructed subjects' 'capital' (but could be easily acquired through simple instruction).

Task design: tailoring production requirements to capital. Instructions typically describe the details of the mapping from a subjects' choices to her payoff, without suggesting preferable strategies, because the subjects' ability to discover optimal strategies is usually the focus of inquiry. But since instructions convey production requirements to subjects, they can also influence whether subjects are able to use their capital to produce effectively. Instructions are often written with something like this kind of task design in mind. Computer displays are designed so that important information is prominently displayed and visible (minimizing attention requirements) and history is retrievable from a menu (minimizing memory requirements). Subjects are sometimes given tables enabling them to compute mapping from actions to payoffs, to simplify calculations they may not be able to do perfectly. Many experimenters do such studies, fiddling with instructions until they are "clear." For example, Smith and Walker (1993) write:

> In a new experimental situation, if the experimenter finds that decision error is biased enough to contradict the theory, then the first thing to question is the experimental instructions and procedures. Can they be simplified? (p. 10)

They write that simplifying instructions "may help to reduce decision cost." In our framework, instructions can convey production requirements more clearly, minimizing the additional capital needed to perform well.

Capital transfer. The usefulness of cognitive capital in different productive tasks is an important empirical question. Put in psychological terms, how well does training in one task transfer to another?

There are many reasons to think transfer is low. Just as carpenters, chefs, and golfers use many specialized tools rather than a few all-purpose ones, evidence from cognitive psychology suggests that a lot of knowledge comes in the form of memory for domain-specific facts or decision rules customized to situations (in cognitive science this is sometimes called "modularity"). Experts tend to have lots of knowledge about facts in some domain, but the rules they infer from those facts are not easily generalized (e.g., Camerer and Johnson, 1991). Chess experts, for examples, have large 'vocabularies' of positions from famous games, and know what move to play from each position, but the high-level rules they induce from their knowledge ("defend the center," "protect your king") do not generalize well to other domains.

More generally, there is little evidence that well-educated subjects perform experimental tasks much differently than less-educated ones (see Ball and Cech, 1996). In addition, subjects trained to use a heuristic which is optimal in problems with certain surface features often fail to apply the same heuristic when faced with new problems that are structurally-identical but have different surface features. For example, Kagel and Levin (1986) found that subjects gradually reduced their bids in repeated three-person common-value auctions, so they learned to mostly avoid the "winner's curse." Then the number of bidders was changed to six. If subjects had learned the structural reason for the winner's curse—choosing the highest bid tends to select the most optimistic common-value estimate—they would reduce their bids when the number of bidders rises, but instead they raised their bids. The data suggest that what subjects learned in the three-bidder case (their cognitive capital) was customized to that situation, and did not transfer well to the six-bidder case.

A final thought: Further research on the capital-labor theory would benefit greatly from having more types of data about decision processes than experimental economists usually collect. Smith and Walker (1993) articulate a bias against studying cognitive processes which many economists share:

> One can think of z as the decision cost or effort (concentration, attention, thinking, monitoring, reporting, acting) which the subject applies to the task presented by the experimenter. Like quarks in particle physics *we may have no direct measures of z*, but we look for traces of its effects on the choice of y . . . by manipulation of the experimental procedures that affect z and thus y. [Emphasis ours]

We disagree because one *can* measure decision effort (z) more directly. Studies have done precisely this using looking-up patterns (Camerer et al. 1993), response times (Wilcox, 1993), measures of recall (which proxy for the amount of decision

effort expended in the first place), verbal protocols, pupil dilation (e.g., Kahneman and Peavler, 1969), heart rate or galvanic skin response (e.g., Dickhaut et al. 1997) and so forth.

IV. Stylized facts and provocative conjectures

The results compiled in Table 1 can be summarized as stylized facts or provocative conjectures.

1. Most studies do not show a clear improvement in mean performance. The most common result is no effect on mean performance (see also Bonner, Young and Hastie, 1996, Tables 3–4). Of course, the failure to find a significant performance effect of incentive may be due to low statistical power (which is difficult to judge without making power calculations for each study). Aggregating a series of insignificant effects in a proper meta-analysis adds power and could establish collective significance where simply counting studies, as we have done, would not.

Nonetheless, it is widely believed among economists—perhaps even more so among non-experimentalists—that paying subjects will necessarily increase their effort and their performance. The underpinning of this hypothesis was carefully articulated by Vernon Smith (1976), who wrote (p. 277):

> ...it is often possible in simple-task experiments to get satisfactory results without monetary rewards by using instructions to induce value by role-playing behavior (i.e., 'think of yourself as making a profit of such and such when...')..but such game values are likely to be weak, erratic, and easily dominated by transactions costs, and subjects may be readily satiated with 'point' profits.

The last sentence summarizes the case against using hypothetical rewards, and in favor of using money: Money is thought to be stronger in force, more reliable, and less satiable than hypothetical rewards. The extent to which any given reward mediums—money, points, grades, public announcement of scores—have these features is an empirical question. Smith was convinced about the special motivational properties of money after observing double auctions which failed to converge sharply unless subjects were paid, especially for low-profit marginal trades (Smith, 1962). But the claim that nonfinancial rewards are weak and satiable in other tasks has not been as firmly established. It may be that in double auctions, which require substantial training sessions and many periods of stationary "Groundhog Day" replication, subjects tend to get especially tired or bored, and money keeps their attention from flagging better than other rewards. However, this is not a strong argument for always using money in tasks where fatigue and boredom are less likely to set in.

The faith economists have in financial incentives is important because it influences all stages of experimental methodology, reporting, citation, and debate. For example, a search of the *American Economic Review* from 1970–97 did not turn up a single published experimental study in which subjects were not paid according to performance. Authors believe that referees will automatically reject a study which uses only hypothetical-payment data (and the authors are probably correct!). Furthermore, seminar participants invariably criticize experimental evidence of violations of rationality principles by conjecturing that if enough incentive were offered the violations would disappear, ignorant of the fact that this conjecture has generally proved false. For example, Aumann (1990) wrote:

> It is sometimes asserted that game theory is not "descriptive" of the "real world," that people don't really behave according to game-theoretic prescriptions. To back up such assertions, some workers have conducted experiments using poorly motivated subjects, subjects who do not understand what they are about and are paid off by pittances; as if such experiments represented the real world (p. xi).

This passage implies that subjects who are motivated by more than "pittances" *will* be described by game theory, even if lower-paid subjects do not. In fact, there is simply no laboratory evidence for this claim, and plenty of evidence against it.

Since our review shows that payment does not *always* matter, we suggest a revised three-part standard for judging results: Critics can insist that researchers use substantial incentives for tasks which have shown substantial incentive effects in previous studies; authors can argue for not using incentives if previous studies have established little effect; and in cases where previous studies are ambiguous, authors must run at least one real-payment condition. (The latter requirement would also add to the body of literature establishing incentive effects, which is hardly conclusive at this point.)

2. When incentives do affect performance, they often reduce the variance of responses (see Smith and Walker, 1993). Incentives often reduce variance by reducing the number of extreme outliers, probably caused by thoughtless, unmotivated subjects. Lower variance is important for three reasons:

First, the fact that incentives lower variance might provide an important clue about how incentives affect attention and reasoning, and consequently performance.

Second, if incentives reduce variation in responses, they improve statistical power and help experimenters test predictions more effectively. Used for this purpose, increased incentive is simply a way of producing higher-quality data and doing better science (like buying purer chemicals or less reactive beakers to do better chemistry). Of course, other methods might work the same magic more cheaply. Trimmed means and robust statistical methods also reduce the influence of outliers. Higher-power tests (e.g., Forsythe et al., 1994), and power-optimized

experimental designs (El-Gamal and Palfrey, in press; Müller and Ponce De Leon, 1996), increase the quality of inferences drawn from noisy data. Experimenters who use incentives purely to reduce dispersion should adopt these other techniques as well.

Third, variance reduction can change group outcomes dramatically in some tasks, when aggregate behavior is especially sensitive to decisions by outlying individuals. Creative tasks (like R & D), in which discovery of a correct answer by one person implies a group discovery, order-statistic coordination games (e.g., Van Huyck, Battalio and Beil, 1990) and asset markets in which behavior depends sensitively on common knowledge of rationality (e.g., Smith, Suchanek and Williams 1988) are examples: One unusual person might cause the group to behave unusually. If high incentives reduce individual variance they may reduce variance in group behavior even more dramatically; in those cases incentives will have a particularly strong treatment effect which should probably not be ignored.

3. Incentive effects are comparable in magnitude to other kinds of treatment effects; and incentives may be substitutes for, or complements with, other treatments. The capital-labor-production theory emphasizes that while incentives do have effects, the effects are often comparable in magnitude to the effects of capital and production variables. In a striking example, Baker and Kirsch (1991) studied pain endurance of female students who held their hands in cold water for 4–8 minutes. In an incentive condition the subjects earned $2 for lasting four minutes and $1 for each additional minute of pain they could stand. In a coping condition they were instructed in how to deal with pain. Incentives did induce the students to withstand more pain, but learning to cope increased their pain endurance as well. Coping skill is a capital variable with a positive effect comparable to the effect of incentives.

Capital and task variables may also be substitutes or complements with incentives. For example, many experimenters suspect that experience is a substitute for incentive. For example, Jamal and Sunder (1991) found that incentives reduced the variance of prices in commodity double-auctions with inexperienced subjects, but had little effect with experienced subjects. A reasonable guess is that the effect on mean performance and reduced variance from one session of experimental experience is roughly equivalent to the effect of doubling or tripling incentives. Some studies show a more dramatic experience effect. Smith and Walker (1993) estimate that one session of experience reduces the dispersion of bids around a Nash equilibrium bidding function about as much as a twenty-fold increase in incentives. McKelvey and Ordeshook (1988) report experience effects which are about equal to the effect of a hundred-fold increase in incentive in Fiorina and Plott (1978). The substitutability of experience effects and incentive effects suggests that the implicit requirement in experimental economics that subjects be paid according to performance could be replaced with a requirement that experimenters who do not pay subjects performance incentives should at least report some data from experienced subjects (which many experimenters do anyway).

Feedback is likely to be a complement with incentives because it is hard to imagine that incentives alone, without feedback about the quality of previous decisions, would have much effect; and the effect of feedback is likely to be stronger in the presence of incentives.

Incentives may interact with treatments in other ways too. Awasthi and Pratt (1991) found that subjects of a certain kind (high in "perceptual differentiation," one measure of intelligence) reduced their error rate by half with higher incentives, while other subjects did not improve at all. Glucksberg (1962) found that incentives helped performance on easy problems but hurt performance on hard problems. Schwartz (1982) found that high incentives reduced performance only for subjects who had been pretrained (and, in his interpretation, had learned a 'stereotypical' response). Atkinson (1958) found that subjects performed better if they had a high measured "need for achievement" (a proxy for intrinsic motivation). Our point is not that these types of individual differences among people or among tasks should be the main focus of economics experiments. But economists who vary incentive conditions because they presume incentives are a highly predictive variable should also pay attention to task and personal variables.

4. In tasks with no performance standard, incentives seem to induce substitution away from socially desirable or pleasurable behavior. In tasks like allocation of money (dictator games), choosing among risky gambles, and perhaps others, it appears that subjects act more generously and risk-preferring when payments are hypothetical. If they behave this way because generosity and risk-taking are seen as socially desirable, and social desirability depends to some extent on subject-experimenter interaction, then incentives may be especially useful for minimizing these kinds of "demand effects" (cf. Hoffman, McCabe and Smith, 1996b). Also, if one is interested in differences among individuals (or groups) in social preference or risk-taking, then calibrating these "tastes" by varying incentive may be a particularly effective way to use incentives (e.g., Andreoni and Miller, 1997), and a different use than to induce careful thought.

We end this list with a provocative conjecture:

5. There is no replicated study in which a theory of rational choice was rejected at low stakes in favor of a well-specified behavioral alternative, and accepted at high stakes. The complaint that subjects were insufficiently motivated often arises when a principle of rational choice—transitivity, dominance, game-theoretic equilibrium, or perhaps self-interest—appears to be violated in favor of an alternative, more psychologically plausible, hypothesis. Critics and referees very commonly assert that if the stakes were just high enough the rationality rejection would disappear. While several studies have tried to make rationality violations disappear—in utility theory paradoxes, ultimatum bargaining, and voting experiments—none have succeeded in clearly overturning anomalies.

Because the intellectual stakes are so high when interesting anomalies are discovered, a limited number of replications aimed at testing their robustness (to

stakes, experience, etc.) are probably still worthwhile. However, since *all* established anomalies have survived these kinds of hostile attacks, uninformed critics should quit talking as if simply raising the stakes would make effects disappear. So far, that hasn't proved true; and nothing in any sensible understanding of human psychology suggests that it would.

V. Conclusion

We reviewed 74 experimental papers in which the level of financial performance-based incentive given to subjects was varied. Our primary interest is in advancing the simmering debate in experimental methodology about when subjects should be paid, and why.

The data show that incentives sometimes improve performance, but often don't. This unsurprising conclusion implies that we should immediately push beyond debating the caricatured positions that incentives always help or never help. Adopting either position, or pretending that others do, is empirically misguided and scientifically counterproductive. In our view, the data show that higher levels of incentives have the largest effects in judgment and decision tasks. Incentives improve performance in easy tasks that are effort-responsive, like judgment, prediction, problem-solving, recalling items from memory, or clerical tasks. Incentives sometimes hurt when problems are too difficult or when simple intuition or habit provides an optimal answer and thinking harder makes things worse. In games, auctions, and risky choices the most typical result is that incentives do not affect mean performance, but incentives often reduce variance in responses. In situations where there is no clear standard of performance, incentives often cause subjects to move away from favorable 'self-presentation' behavior toward more realistic choices. (For example, when they are actually paid, subjects who dictate allocations of money to others are less generous and subjects choosing among gambles take less risk.)

One way to comprehend these results is a "capital-labor-production theory" of cognition (extending Smith and Walker, 1993). The capital-labor-production framework assumes that the 'labor' or mental effort subjects exert depends upon their intrinsic motivation and financial incentives. But the effect of extra effort on performance also depends on their level of cognitive 'capital'—know-how, heuristics, analytical skills, previous experience in the task, and so forth—and its productive value for a specified task. Capital and labor can substitute: For example, a few experiments suggest that one session of experimental experience has an effect roughly comparable to (at least) tripling incentives.

Capital-labor-production theory provides a language for describing why incentives matter in some tasks but not in others. Tasks which are easy require little capital, so subjects can perform well with little motivation and paying extra will not help much. Tasks which are hard require too much capital (which cannot be formed in the short run of an experiment), so the effect of labor on performance

can be low (or negative). Obviously, spelling out the details of the capital-labor theory is a big project for another day. The main point is that to the extent incentive effects are worth studying, the effects of capital-relevant treatment variables are worth studying too.

An obvious direction for future research is to ask about these effects in natural settings, such as inside firms. Firms casually experiment with mixtures of incentive schemes all the time and often have an implicit theory about the interplay of incentive, human capital, and task demands. There is ample field evidence that incentives do alter behavior in ways predicted by theory, but there is less evidence that firms offer the contracts they are predicted to (see Prendergast, in press, for an authoritative review). The experimental data suggest that for easy or hard jobs, and intrinsically motivated workers, marginal changes in incentives will not improve performance much. However, for boring jobs, unmotivated workers, or tasks in which variance is bad, incentives are likely to have positive effects. Of course, these generalizations abstract from phenomena which are likely to loom larger in firms than in the lab—for example, social comparison among workers to the wages of others, dynamic "ratchet" effects in motivating effects of incentives, and so forth. Another lesson from the lab is that the effects of incentive on performance are comparable in magnitude (and often less than) the effects of experience, individual differences, task difficulty, and so on. Firms might improve performance by redesigning tasks to suit human capital as much as they can improve performance by raising incentives.

Our review also suggests some revisions to experimental method. Currently it is essentially impossible to report experimental research in economics journals if subjects have not been financially motivated. We think this prohibition should be replaced by a three-part standard: (i) Referees who would reject a paper purely on the grounds that subjects were not paid must cite a preponderance of previous literature establishing that incentives affect behavior meaningfully, in a task similar to that studied in the paper under consideration. (ii) Authors could defend the practice of collecting data from unpaid subjects by pointing to previous research showing that financial incentives did not matter in their task. (iii) For the many tasks where the data are mixed, authors should be encouraged (or perhaps required) to run different incentive conditions. (The latter requirement would build up a database of systematic observations rapidly—in a sense, it would spread the economic "tax" of finding out whether incentives do matter equally to all experimentalists.) These rules should help point the debate where it should head—away from differences in implicit models of subject behavior and towards data.

An open question is what results from the laboratory tell us about incentives in naturally-occurring environments (e.g., wages in firms, taxation and subsidy for public choices). Our view is that experiments measure only short-run effects, essentially holding capital fixed. The fact that incentives often do not induce different (or better) performance in the lab may understate the effect of incentives in natural settings, particularly if agents faced with incentive changes have a chance to build up capital—take classes, seek advice, or practice. In principle,

different sorts of experiments could be conducted in which subjects return repeatedly, or have a chance to invest in capital as part of their experimental choices, to allow for long-run effects, and experimenters interested in extrapolating to the outside world might consider running such experiments.

Finally, we cannot end a casual review of this sort without several caveats. Our sampling of studies and classification of incentive levels, and effects, should certainly be done more carefully. Besides the usual problems of meta-analysis, comparing incentive effects across different experiments would benefit from putting all incentives on a single scale (say, 1997 dollars per choice) and tying response rates to incentive levels, perhaps with some kind of general stochastic choice function.

There are many other questions about uses of financial incentives in experiments which our review does not address.

The lottery ticket procedure: There is some debate about whether paying subjects in functions of units of probability (the "binary lottery" procedure) induces controlled risk tastes reliably. The procedure should work in theory, if subjects reduce compound lotteries and maximize their chance of winning a fixed prize, but it does not work in practice (e.g., Selten et al., 1995), or at best, works only for the minority of subjects who obey reduction when directly tested (Prasnikar, 1998).

Losses: Because it is generally difficult to impose losses or punishments on subjects for bureaucratic reasons—university committees that approve protocols involving human subjects strongly object to it—we do not know how earning money and losing money differ.

Paying a fraction of subjects: Another question we cannot answer is whether paying one out of N subjects a larger stake, or paying subjects for one out of N high-stakes choices, provides as much incentive as paying a lower stake for each choice. Some of the studies we reviewed do use these random-payment schemes and it appears that these are roughly equivalent, at least for simple choices (paying one out of N may even be more motivating, if subjects overweigh their chances of being selected). However, more careful exploration would be useful.

Tournaments: Finally, some experimenters use "tournament" incentives in which the returns to performance are convex in performance or status-based (e.g., only the top few performers receive large prizes). In theory, tournament incentives should induce more status-seeking and risk-taking and hence, do not lead subjects to incentive-compatibly maximize expected profit (which is why economists generally eschew them). Whether tournaments actually do have those unintended effects has not been carefully investigated.

Acknowledgments

Very helpful comments were received from Baruch Fischhoff, Reid Hastie, John Kagel, Daniel Kahneman, George Loewenstein, Rob MacCoun, Chuck Manski, Richard Thaler, two anonymous referees, and many participants in the

NSF/Berkeley Econometrics Lab conference on elicitation of preferences, July/August 1997. Angela Hung provided meticulous research assistance.

References

Andreoni, James and John H. Miller. (1997). "Giving According to GARP: An Experimental Study of Rationality and Altruism," University of Wisconsin Department of Economics Working Paper, October.

Arkes, Hal R., Robyn M. Dawes, and Caryn Christensen. (1986). "Factors Influencing the Use of a Decision Rule in a Probabilistic Task," *Organizational Behavior and Human Decision Processes* 37, 93–110.

Ashton, Robert H. (1990). "Pressure and Performance in Accounting Decision Settings: Paradoxical Effects of Incentives, Feedback, and Justification," *Journal of Accounting Research* 28, 148–180.

Atkinson, John W. (1958). "Towards Experimental Analysis of Human Motivation in Terms of Motives, Expectancies, and Incentives." In John W. Atkinson (ed.), *Motives in Fantasy*, *Action*, *and Society*. New York: Van Nostrand.

Aumann, Robert. (1990). "Foreword." In Alvin E. Roth and Marilda A. Oliveira Sotomayor (ed.), *Two-Sided Matching*: *A Study in Game-Theoretic Modeling and Analysis*, p. xi. Cambridge, UK: Cambridge University Press.

Awasthi, Vidya and Jamie Pratt. (1990). "The Effects of Monetary Incentives on Effort and Decision Performance: the Role of Cognitive Characteristics," *The Accounting Review* 65, 797–811.

Bahrick, Harry P. (1954). "Incidental Learning under Incentive Conditions," *Journal of Experimental Psychology* 47, 170–172.

Baker, S. L. and I. Kirsch. (1991). "Cognitive Mediators of Pain Perception and Tolerance," *Journal of Personality and Social Psychology*, 61, 504–510.

Ball, Sheryl B. and Paula-Ann Cech. (1996). "Subject Pool Choice and Treatment Effects in Economic Laboratory Research." In R. Mark Isaac (ed.), *Research in Experimental Economics* Vol. 6, pp. 139–292. Greenwich, CT: JAI Press.

Battalio, Raymond C., John H. Kagel, and Komain Jiranyakul. (1990). "Testing between Alternative Models of Choice under Uncertainty," *Journal of Risk and Uncertainty* 3, 25–50.

Baumeister, Roy F. (1984). "Choking Under Pressure: Self Consciousness and Paradoxical Effects of Incentives on Skillful Performance," *Journal of Personality and Social Psychology* 46, 610–620.

Beattie, Jane and Graham Loomes. (1997). "The Impact of Incentives upon Risky Choice Experiments," *Journal of Risk and Uncertainty* 14, 155–168.

Binswinger, Hans P. (1980). "Attitudes Toward Risk: Experimental Measurement in Rural India," *American Journal of Agricultural Economics* 62, 395–407.

Bohm, Peter. (1994). "Time Preference and Preference Reversal among Experienced Subjects: The Effects of Real Payments," *The Economic Journal* 104, 1370–1378.

Bolle, Friedel. (1990). "High Reward Experiments without High Expenditure for the Experimenter?" *Journal of Economic Psychology* 11, 157–167.

Bonner, Sarah E. S., Mark Young, and Reid Hastie. (1996). "Financial Incentives and Performance in Laboratory Tasks: The Effects of Task Type and Incentive Scheme Type," Unpublished manuscript, University of Southern California Department of Accounting.

Bull, Clive, Andrew Schotter, and Keith Weigelt. (1987). "Tournaments and Piece Rates: An Experimental Study," *Journal of Political Economy* 95, 1–33.

Camerer, Colin F. (1987). "Do Biases in Probability Judgment Matter in Markets? Experimental Evidence," *American Economic Review* 77, 981–997.

Camerer, Colin F. (1989). "An Experimental Test of Several Generalized Utility Theories," *Journal of Risk and Uncertainty* 2, 61–104.

Camerer, Colin F. (1990). "Behavioral Game Theory." In R. Hogarth (ed.), *Insights in Decision Making*: *Theory and Applications*. Chicago: University of Chicago Press, 1990, pp. 311–336.

Camerer, Colin F. (1996). "Rules for Experimenting in Psychology and Economics, and Why They Differ." In W. Guth and E. Van Damme (eds.), *Essays in Honor of Reinhard Selten*. New York: Springer-Verlag.

Camerer, Colin F. (1998). "Behavioral Economics and Nonrational Decision Making in Organizations." In J. Halpern and B. Sutton (eds.), *Decision Making in Organizations*. Ithaca, NY: Cornell University Press.

Camerer, Colin F. (in press). "Prospect Theory in the Wild: Evidence from the Field." In D. Kahneman and A. Tversky (eds.), *Choices*, *Values*, *and Frames*.

Camerer, Colin F. and Eric Johnson. (1991). "The Process-Performance Paradox in Expert Judgment: Why Do Experts Know So Much and Predict So Badly?" In A. Ericsson and J. Smith (eds.), *Toward a General Theory of Expertise*: *Prospects and Limits*, pp. 195–217. Cambridge, UK: Cambridge University Press.

Camerer, Colin F., Eric Johnson, Talia Rymon, and Sankar Sen. (1993). "Cognition and Framing in Sequential Bargaining." In K. Binmore, A. Kirman, and P. Tani (eds.), *Frontiers of Game Theory*, Cambridge, MA: MIT Press.

Camerer, Colin F., Teck Ho, and Keith Weigelt. (1997). Unpublished data.

Camerer, Colin F. and Keith Weigelt. (1988). "Experimental Tests of a Sequential Equilibrium Reputation Model," *Econometrica* 56, 1–36.

Castellan, N. John. (1969). "Effect of Change of Payoff in Probability Learning," *Journal of Experimental Psychology* 79, 178–182.

Cheng, Patricia and Keith Holyoak. (1985). "Pragmatic Reasoning Schemas," *Cognitive Psychology* 17, 391–416.

Conlisk, John (1989). "Three Variants on the Allais Example." *American Economic Review* 79, 392–407.

Cooper, David J., John H. Kagel, Wei Lo, and Qingliang Gu. (in press). "An Experimental Study of the Ratchet Effect: The Impact of Incentives, Context, and Subject Sophistication on Behavior," *American Economic Review*.

Cosmides, Leda. (1985). "The Logic of Social Exchange: Has Natural Selection Shaped How Humans Reason? Studies with the Wason Selection Task," *Cognition* 31, 187–276.

Cox, James C. and David M. Grether. (1996). "The Preference Reversal Phenomenon: Response Mode, Markets, and Incentives," *Economic Theory* 7, 381–405.

Craik, Fergus I. M. and Endel Tulving. (1975). "Depth of Processing and the Retention of Words in Episodic Memory," *Journal of Experimental Psychology*: *General* 104, 268–294.

Cubitt, Robin P, Chris Starmer, and Robert Sugden. (1998). "On the Validity of the Random Lottery Incentive System," *Experimental Economics* 1, 115–132.

Cummings, Ronald G., Glenn W. Harrison, and E. Elisabet Rutstrom. (1995). "Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive-Compatible?" *American Economic Review* 85, 260–266.

Dawes, R. M., D. Faust and P. E. Meehl. (1989). "Clinical versus Actuarial Judgment," *Science* 243, 1668–1674.

Dickhaut, John, Kip Smith, Kevin McCabe, Nicole Peck, and Vijay Rajan. (1997). "The Emotional and Mental Effort Dynamics of the English Auction," University of Minnesota Working Paper, Presented at ESA Meeting, September.

Drago, Robert and John S. Heywood. (1989). "Tournaments, Piece Rates, and the Shape of Payoff Function," *Journal of Political Economy* 97, 992–998.

Edwards, Ward. (1953). "Probability Preferences in Gambling," *American Journal of Psychology* 66, 349–364.

Edwards, Ward. (1961). "Probability Learning in 1000 Trials," *Journal of Experimental Psychology* 62, 385–394.

Eger, Carol and John Dickhaut. (1982). "An Examination of the Conservative Information Processing Bias in an Accounting Framework," *Journal of Accounting Research* 20, 711–723.

Eisenberger, R. and J. Cameron. (1996). "Detrimental Effects of Rewards: Reality or Myth?" *American Psychologist*, 51, 1153–1166.

El-Gamal, Mahmoud and Thomas R. Palfrey. (1996). "Economical Experiments: Bayesian Efficient Experimental Design." *International Journal of Game Theory*, 25, 476–495.

Ericsson, K. Anders and Jacqui Smith. (eds). (1991). *Toward a General Theory of Expertise*: *Prospects and Limits*. Cambridge, UK: Cambridge University Press.

Fehr, Ernst and Elena Tougareva. (1996). "Do High Monetary Stakes Remove Reciprocal Fairness? Experimental Evidence from Russia," University of Zurich Working Paper.

Fiorina, Morris P. and Charles R. Plott. (1978). "Committee Decisions under Majority Rule: An Experimental Study," *American Political Science Review* 72, 575–598.

Forsythe, Robert, Joel L. Horowitz, N. E. Savin, and Martin Sefton. (1994). "Fairness in Simple Bargaining Experiments," *Games and Economic Behavior* 6, 347–369.

Fouraker, Lawrence and Sidney Siegel. (1963). *Bargaining and Group Decision Making*. New York: McGraw-Hill.

Friedman, Daniel. (1998). "Monty Hall's Three Doors: Construction and Deconstruction of a Choice Anomaly," *American Economic Review* 88, 933–946.

Glucksburg, Sam. (1962). "The Influence of Strength and Drive on Functional Fixedness and Perceptual Recognition," *Journal of Experimental Psychology* 63, 36–41.

Grether, David. M. (1980). "Bayes' Rule as a Descriptive Model: The Representativeness Heuristic," *Quarterly Journal of Economics* 95, 537–557.

Grether, D. M. (1981). "Financial Incentive Effects and Individual Decision Making," California Institute of Technology Working Paper No. 401.

Grether, D. M. (1990). "Testing Bayes Rule and the Representativeness Heuristic: Some Experimental Evidence," *Journal of Economic Behavior and Organization* 17, 31–57.

Grether, David M. and Charles R. Plott. (1979). "Economic Theory of Choice and the Preference Reversal Phenomenon," *American Economic Review* 69, 623–638.

Güth, Werner, R. Schmittberger, and B. Schwarze. (1982). "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization* 3, 367–388.

Harless, David W. and Colin F. Camerer. (1994). "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica* 62, 1251–1290.

Harrison, Glenn W. (1994). "Expected Utility Theory and the Experimentalists," *Empirical Economics* 19, 223–253.

Harrison, Glenn W. and E. Elisabet Rutstrom. (in press). "Experimental Evidence of Hypothetical Bias in Value Elicitation Methods." In C. R. Plott and V. L. Smith (eds.), *Handbook of Experimental Economics Results*.

Hertwig, Ralph and Andreas Ortmann. (in press). "Experimental Practices in Economics: A Methodological Challenge for Psychologists," Behavioral and Brain Sciences.

Hey, John D. (1982). "Search for Rules of Search," *Journal of Economic Behavior and Organization* 3, 65–81.

Hey, John D. (1987). "Still Searching," *Journal of Economic Behavior and Organization* 8, 137–144.

Hoffman, Elizabeth, Kevin McCabe, and Vernon Smith. (1996a). "On Expectations and Monetary Stakes in Ultimatum Games," *International Journal of Game Theory* 25, 289–301.

Hoffman, Elizabeth, Kevin McCabe, and Vernon L. Smith. (1996b). "Social Distance and Other-Regarding Behavior in Dictator Games," *American Economic Review* 86, 653–660.

Hogarth, Robin M. and Hillel J. Einhorn. (1990). "Venture Theory: A Model of Decision Weights," *Management Science* 36, 780–803.

Hogarth, Robin M., Brian J. Gibbs, Craig R. M. McKenzie, and Margaret A. Marquis. (1991). "Learning from Feedback: Exactingness and Incentives," *Journal of Experimental Psychology*: *Learning*, *Memory and Cognition* 17, 734–752.

Irwin, Julie R., Gary H. McClelland, and William D. Schulze. (1992). "Hypothetical and Real Consequences in Experimental Auctions for Insurance against Low-Probability Risks," *Journal of Behavioral Decision Making* 5, 107–116.

Irwin, Julie, Michael McKee, Gary McClelland, William Schulze, and Elizabeth Norden. (in press). "Payoff Dominance vs. Cognitive Transparency in Decision Making," *Economic Inquiry*.

Jamal, Karim and Shyam Sunder. (1991). "Money vs. Gaming: Effects of Salient Monetary Payments in Double Oral Auctions," *Organizational Behavior and Human Decision Processes* 49, 151–166.

Jenkins, G. Douglas, Jr., Atul Mitra, Nina Gupta, and Jason D. Shaw. (1998). "Are Financial Incentives Related to Performance? A Meta-Analytic Review of Empirical Research," *Journal of Applied Psychology* 83, 777–787.

Johannesson, Magnus, Bengt Liljas, and Per-Olov Johansson. (1998). "An Experimental Comparison of Dichotomous Choice Contingent Valuation Questions and Real Purchase Decisions," *Applied Economics*, 30, 643–647.

Kachelmeier, Steven J. and Mohamed Shehata. (1992). "Examining Risk Preferences under High Monetary Incentives: Experimental Evidence from the People's Republic of China," *American Economic Review* 82, 1120–1141.

Kagel, John H. and Dan Levin. (1986). "The Winner's Curse and Public Information in Common Value Auctions," *American Economic Review* 76, 894–920.

Kahneman, Daniel and W. Scott Peavler. (1969). "Incentive Effects and Pupillary Changes in Association Learning," *Journal of Experimental Psychology* 79, 312–318.

Kahneman, Daniel, W. Scott Peavler, and Linda Onuska. (1968). "Effects of Verbalization and Incentive on the Pupil Response to Mental Activity," *Canadian Journal of Psychology* 22, 186–196.

Kroll, Y., H. Levy, and A. Rapoport. (1988). "Experimental Tests of the Separation Theorem and the Capital Asset Pricing Model," *American Economic Review* 78, 500–519.

Kroll, Y., H. Levy, and A. Rapaport. (1988). "Experimental Tests of the Mean-Variance Model for Portfolio Selection," *Organizational Behavior and Human Decision Processes* 42, 388–410.

Lepper, Mark R., David Greene, and Richard E. Nisbett. (1973). "Undermining Childrens' Intrinsic Interest in with Extrinsic Reward: A Test of the 'Overjustification' Hypothesis," *Journal of Personality and Social Psychology* 28, 129–137.

Libby, Robert and Marlys G. Lipe. (1992). "Incentives, Effort, and the Cognitive Processes Involved in Accounting-Related Judgments," *Journal of Accounting Research* 30, 249–273.

List, John A. and Jason F. Shogren. (1998). "The Deadweight Loss of Christmas: Comment," *American Economic Review* 88, 1350–1355.

Loomes, Graham and Caron Taylor. (1992). "Non-Transitive Preferences over Gains and Losses," *Economic Journal* 102, 357–365.

McGraw, Kenneth O. and John C. McCullers. (1979). "Evidence of a Detrimental Effect of Extrinsic Incentives on Breaking a Mental Set," *Journal of Experimental Social Psychology* 15, 285–294.

McKelvey, Richard and Ordeshook, Peter. (1988) "A Decade of Experimental Research on Spatial Models of Elections and Committees." In M. J. Hinich and J. Enelow (eds.), *Government*, *Democracy*, *and Social Choice*. Cambridge, MA: Cambridge University Press.

McKelvey, Richard and Thomas Palfrey. (1992). "An Experimental Study of the Centipede Game," *Econometrica* 60, 803–836.

Merlo, Antonio and Andrew Schotter. (1999). "A Surprise-Quiz View of Learning in Economic Experiments," *Games and Economic Behavior* 28, 25–54.

Miller, Louise B. and Betsy W. Estes. (1961). "Monetary Reward and Motivation in Discrimination Learning," *Journal of Experimental Psychology* 61, 501–504.

Müller, W. G. and A. M. C. Ponce de Leon. (1996). "Optimal Design of an Experiment in Economics," *Economic Journal* 106, 122–127.

Neelin, Janet [now Currie], Hugo Sonnenschein, and Matthew Spiegel. (1988). "A Further Test of Noncooperative Bargaining Theory: Comment," *American Economic Review* 78, 824–836.

Nilsson, Lars-Goran. (1987). "Motivated Memory: Dissociation between Performance Data and Subjective Reports," *Psychological Research* 49, 183–188.

Phillips, Lawrence D. and Ward Edwards. (1966). "Conservatism in a Simple Probability Inference Task," *Journal of Experimental Psychology* 72, 346–354.

Prasnikar, Vesna. (1998). "How Well Does Utility Maximization Approximate Subjects' Behavior? An Experimental Study," University of Pittsburgh Department of Economics, December.

Prendergast, Canice. (in press). "The Provision of Incentives in Firms," *Journal of Economic Literature*.

Reber, Arthur S. (1989). "Implicit Learning and Tacit Knowledge," *Journal of Experimental Psychology*: *General* 118, 219–235.

Riedel, James A., Delbert M. Nebeker, and Barrie L. Cooper. (1988). "The Influence of Monetary Incentive on Goal Choice, Goal Commitment, and Task Performance," *Organizational Behavior and Human Decision Processes* 42, 155–180.

Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir. (1991). "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An Experimental Study," *American Economic Review* 81, 1068–1095.

Salthouse, Timothy A., Janice D. Rogan, and Kenneth A. Prill. (1984). "Division of Attention: Age Differences on a Visually Presented Memory Task," *Memory and Cognition* 12, 613–620.

Samuelson, William F. and Max H. Bazerman. (1985). "The Winner's Curse in Bilateral Negotiations," *Research in Experimental Economics* 3, 105–137.

Schoemaker, Paul J. H. (1990). "Are Risk Attitudes Related across Domains and Response Modes?" *Management Science* 36, 1451–1463.

Schwartz, Barry. (1982). "Reinforcement-Induces Behavioral Stereotypy: How Not to Teach People to Discover Rules," *Journal of Experimental Psychology*: *General* 111, 23–59.

Scott, W. E., Jing-Lih Farh, and Philip M. Podsakoff. (1988). "The Effects of 'Intrinsic' and 'Extrinsic' Reinforcement Contingencies on Task Behavior," *Organizational Behavior and Human Decision Processes* 41, 405–425.

Sefton, Martin. (1992). "Incentives in Simple Bargaining Games," *Journal of Economic Psychology* 13, 263–276.

Selten, Reinhard, A. Sadrieh, and Klaus Abbink. (1995). "Money Does Not Induce Risk Neutral Behavior, but Binary Lotteries Do Even Worse," University of Bonn Working Paper No. B-343.

Siegel, Sidney and Lawrence Fouraker. (1960). *Bargaining and Group Decision Making*: *Experiments in Bilateral Monopoly*. New York: McGraw-Hill.

Siegel, Sidney, Alberta Siegel, and Julia Andrews. (1964). *Choice*, *Strategy*, *and Utility*. New York: McGraw-Hill.

Slonim, Robert and Alvin E. Roth. (1998). "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic," *Econometrica*, 65, 569–596.

Slovic, Paul. (1969). "Differential Effects of Real versus Hypothetical Payoffs on Choices among Gambles," *Journal of Experimental Psychology* 80, 434–437.

Slovic, Paul and Douglas MacPhillamy. (1974). "Dimensional Commensurability and Cue Utilization in Comparative Judgment," *Organizational Behavior and Human Performance* 11, 172–194.

Smith, Vernon L. (1962). "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy* 70, 111–137.

Smith, Vernon L. (1965). "Experimental Auction Markets and the Walrasian Hypothesis," *Journal of Political Economy* 387–393.

Smith, Vernon L. (1976). "Experimental Economics: Induced Value Theory," *American Economic Review* 66, 274–279.

Smith, Vernon L. (1991). "Experimental Economics: Behavioral Lessons for Microeconomic Theory and Policy," 1990 Nancy Schwartz Lecture, KGSM, Northwestern University.

Smith, Vernon L., Gerry Suchanek, and Arlington Williams. (1988). "Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets," *Econometrica* 56, 1119–1151.

Smith, Vernon L. and James M. Walker. (1993). "Rewards, Experience and Decision Costs in First Price Auctions," *Economic Inquiry* 31, 237–244.

Sniezek, Janet A. (1986). "The Role of Variable Labels in Cue Probability Learning Tasks," *Organizational Behavior and Human Decision Processes* 38, 141–161.

Straub, Paul G. and J. Keith Murnighan. (1995). "An Experimental Investigation of Ultimatum Games: Information, Fairness, Expectations, and Lowest Acceptable Offers," *Journal of Economic Behavior and Organization* 27, 345–364.

Tversky, Amos and Daniel Kahneman. (1992). "Advances in Prospect Theory: Cumulative Representation of Uncertainty," *Journal of Risk and Uncertainty* 5, 297–323.

Van Huyck, John, Raymond Battalio, and Richard Beil. (1990). "Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure," *American Economic Review* 80, 234–248.

Wallsten, Thomas S., David V. Budescu, and Rami Zwick. (1993). "Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments," *Management Science* 39, 176–190.

Weber, Elke, Sharoni Shafir, and Ann-Renee Blais. (1998). "Predicting Risk-Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation," Ohio State University Department of Psychology Working Paper.

Weber, Martin, Graham Loomes, Hans-Jurgen Keppe, and Gabriela Meyer-Delius. (in press). "The Impact of Endowment Framing on Market Prices—An Experimental Analysis," *Journal of Economic Behavior and Organization.*

Wilcox, Nathaniel. (1993). "Lottery Choice: Incentives, Complexity, and Decision Time," *Economic Journal* 103, 1397–1417.

Wright, William F. and Mohamed E. Aboul-Ezz. (1988). "Effects of Extrinsic Incentives on the Quality of Frequency Assessments," *Organizational Behavior and Human Decision Processes* 41, 143–152.

Wright, William F. and Urton Anderson. (1989). "Effects of Situation Familiarity and Financial Incentives on Use of the Anchoring and Adjustment Heuristic for Probability Assessment," *Organizational Behavior and Human Decision Processes* 44, 68–82.

Yerkes, R. M. and J. D. Dodson. (1908). "The Relation of Strength of Stimulus to Rapidity of Habit-Formation," *Journal of Comparative and Neurological Psychology* 18, 459–482.