Exp Econ (2011) 14:349–374
DOI 10.1007/s10683-010-9271-3

Saving behavior and cognitive abilities

T. Parker Ballinger · Eric Hudson · Leonie Karkoviata · Nathaniel T. Wilcox

Received: 8 May 2009 / Accepted: 20 December 2010 / Published online: 20 February 2011

Abstract Experiments on saving behavior reveal substantial heterogeneity of behavior and performance. We show that this heterogeneity is reliable and examine several potential sources of it, including cognitive ability and personality scales. The strongest predictors of both behavior and performance are two cognitive ability measures. We conclude that complete explanations of heterogeneity in dynamic decision making require attention to complexity and individual differences in cognitive constraints.

This research was supported by grants from the National Science Foundation with award numbers SES-0350748 (Ballinger) and SES-0350565 (Wilcox), as well as by the University of Houston Research Council. We are grateful to Randall Engle and Richard Heitz of the Attention and Working Memory Lab at Georgia Tech for providing us with their "automatic operation span" software (and advice from D. Stephen Lindsay that led us to them). Bram Cadsby, Glenn Harrison, Ondrej Rydval, Tomomi Tanaka, Mark Thompson, two anonymous referees and the editor provided help or useful commentary, and Sharon O'Donnell and Rick Wilson provided valuable advice on software programming. None of these people are responsible for any errors that remain.

Electronic supplementary material The online version of this article (doi:10.1007/s10683-010-9271-3) contains supplementary material, which is available to authorized users.

T.P. Ballinger
Department of Economics and Finance, Stephen F. Austin State University, Nacogdoches, TX 75962-3009, USA

E. Hudson
New Mexico Public Defender Department, Las Cruces, NM 88001, USA

L. Karkoviata
College of Business, University of Houston-Downtown, Houston, TX 77002, USA

N.T. Wilcox (✉)
Economic Science Institute, Chapman University, Orange, CA 92866, USA
e-mail: nwilcox@chapman.edu

Keywords Bounded rationality · Heterogeneity · Cognitive abilities · Consumption and saving

JEL Classification C91 · D91 · E21

Behavioral heterogeneity is a major theme in the empirical literature on consumption and saving. Work by Hall and Mishkin (1982) and Campbell and Mankiw (1990) suggested that economies might be composed of a mixture of households—some relatively forward-looking and others relatively myopic. Hey and Dardanoni (1988) first documented behavioral heterogeneity in laboratory saving experiments, and individual differences in attention to the future can organize much of this (Ballinger et al. 2003; Carbone and Hey 2004; Carbone 2006).[1] Here we show that this heterogeneity of saving behavior is large and reliable, and that cognitive ability differences explain both an appreciable amount of performance heterogeneity and heterogeneity of estimated forward-looking attention in consumption and saving.

The term "bounded rationality" is used in many ways. In general, we mean the view that constraints on human cognition explain many deviations from canonical theories of rational reasoning, inference and decision making (Nisbett and Ross 1980). Some economists discuss these constraints with standard economic metaphors, arguing that they give rise to computational costs (e.g. Conlisk 1980; Wilcox 1993a, 1993b; Gabaix and Laibson 2000). Camerer and Hogarth (1999) explain how cognitive abilities may be thought of as "cognitive capital" within a "cognitive capital and labor" production theory of decision making performance.[2]

Cognitive abilities might be specific to particular "domains" of decision making. In that event, knowing that a person's decision making in one domain is constrained by specific cognitive constraints tells us little or nothing about his decision making in another domain. Nevertheless, theory and measurement of domain-general cognitive abilities has been a central topic of psychology for a century. From Spearman (1904) to Carroll (1993), it has been known that performance is positively correlated across a great variety of mental tasks, and there is widespread agreement that some relatively domain-general cognitive abilities exist[3] (though learned domain-specific abilities are a large part of human performance as well). Although any of several cognitive ability measures might be useful measures of general cognitive capital involved in economic decision making, none of the mental tasks considered in this old

[1] Brown et al. (2009) is an exception. But since they use subjects at two highly selective universities, this does not conflict with the notion that heterogeneity of cognitive abilities explains cross-sectional variance in performance.

[2] Psychological models of error/effort tradeoffs anticipate these ideas (Russo 1978; Beach and Mitchell 1978; Payne et al. 1988). Various empirical results support such approaches (Conlisk 1996), at least in part.

[3] Debates continue as to whether there are one or several domain-general abilities (Hunt 1999), just how domain-general any cognitive ability is (see e.g. Shah and Miyake 1996 versus Kane et al. 2004), and whether any relatively general abilities can be augmented by training, experience and/or strategies (e.g. Turley-Ames and Whitfield 2003).

and extensive psychological literature are decision making situations that preoccupy economists.

Frederick (2005) and Benjamin et al. (2006) find that various measures of cognitive ability explain cross-sectional variation in well-known risk and time *preference* phenomena associated with *simple* binary choices between lotteries and dated payments. These results encourage us, but our interest lies with complex economic problem-solving like dynamic consumption and saving under income risk. Relatively simple decision tasks are perhaps ideal for revealing preference differences across subjects and treatments precisely because complexity has been designed out as a causal factor.[4] But relatively simple decision problems may censor the potential ability of the least cognitively constrained subjects.[5] We can design dynamic decision making tasks that are relatively easy to explain to subjects yet still computationally challenging. Note, however, that relatively challenging tasks could have paradoxical effects: They might cause an increased propensity to give up on conscious analysis, or might encourage many subjects to employ similar simplification procedures.[6] Therefore, it is not obvious that increased task complexity and/or difficulty will enhance the explanatory force of differences in cognitive abilities. Our experimental design varies the "difficulty" of income streams in a controlled way both within and between individuals, and the results show a relatively strong effect of cognitive abilities for relatively "easy" income streams but less consistent effects for "hard" ones.

Measures of cognitive abilities are extremely varied. Therefore, our experimental design is sequential: Initial samples examine several promising measures, and the most successful ones are then validated in new samples to protect against pretest bias. Two measures seem particularly useful. First, two analytical subtests of the Beta III test of nonverbal reasoning (Kellogg and Morton 1999) consistently predict performance and behavior in our saving game. This test is neither mathematical nor verbal: It is based purely on reasoning about visual images, and this finding bears on domain-generality. Choice under risk and over time, and indeed complex saving decisions, might be viewed by some as "math tests of a different name:" If so, one might be relatively unimpressed when a math-based measure of cognitive ability predicts such decisions. Similarly, we might regard a verbally-based ability test, such as a verbal SAT score, as simply predicting how well subjects understand verbal and/or written instructions of a laboratory decision problem. In the case of a wholly image-based test like the Beta III, such worries are less compelling: We are perhaps seeing something that truly generalizes across domains.

We also find that a "working memory span" (or WM span) test consistently predicts performance and behavior in our saving game. These tests are regarded as mea-

[4] Of course, this is not obvious: Specific cognitive capital may be primarily adapted to handle field environments with specific complexities and/or institutional features, and may function poorly in lab situations stripped of these (Winkler and Murphy 1973; Dyer and Kagel 1996; Harrison and List 2004).

[5] Gabaix and Laibson (2000) make a related design point. Note, however, that differences in cognitive abilities are known to predict failures of reasoning even in the simplest canonical reasoning tasks (Stanovich and West 2000), as well as simple risk, time and fairness preferences (Frederick 2005; Benjamin et al. 2006).

[6] For instance, Hogarth's (1975) simple decision cost model predicts that decision effort and time will ultimately decrease after complexity rises above some critical level.

suring the capacity for controlled allocation of attention and thought (Conway et al. 2005). A large number of studies over the last two decades strongly suggest that WM span is a robust, domain-general predictor of intelligent performance.[7] WM span tests correlate more strongly with performance in many different reasoning tests (with widely varying surface features) than performance in those same tests correlate with one another (Engle et al. 1999; Engle and Kane 2004): WM span is a leading candidate for a domain-general cognitive capital measure.

Cognitive abilities may be correlated with various preferential, personality and demographic differences that might truly account for heterogeneity in saving behavior. This is why a multivariate analysis that includes other potentially correlated predictors will be more compelling than simple bivariate relationships between economic behavior and cognitive abilities. We do find a robustly significant effect of measured cognitive abilities on saving behavior and performance, even when we control for several other potential predictors (demographic and personality differences) that may be correlated with cognitive abilities. In this respect, our study goes beyond existing correlational studies that mostly rely on bivariate evidence, such as Stanovich and West (2000) and Frederick (2005).

1 Theory and experimental design

1.1 Saving game mechanics and theory

Each *subject* s plays five independent *rounds* r of a saving game. For now we suppress the round subscript r to concentrate on the mechanics of a single round. Each round has $t = 1, 2, \dots, 20$ *periods* and begins with exogenous "starter savings" of $A_1 = 2$ ECUs (experimental currency units) at the start of period 1. An i.i.d. random income $\tilde{y}$, equal to either 0 or 6 ECUs with equal probability, is also received at the start of every period t (including period 1). Let y_t^s be *actual* income received at the start of period t by s. After observing this, *cash-in-hand* $X_t^s = A_t^s + y_t^s$ is available to s for consumption and saving, where $A_t^s = X_{t-1}^s - c_{t-1}^s \geq 0\ \forall t > 1$ is her saving accumulated prior to period t. Subject s then chooses integer-valued consumption $c_t^s \in [0, X_t^s]$, leaving savings equal to $A_{t+1}^s = X_t^s - c_t^s$ for future periods. The consumption choice c_t^s purchases "points" according to the utility function $u(c_t)$ below, which we constructed to meet a specific design goal discussed shortly.[8]

[7] WM span measures may also reflect individual differences in prefrontal cortex function in the brain (Kane and Engle 2002). The prefrontal cortex is implicated in delay of current gratification for a delayed reward, as well as various planning behavior. This seems (at least potentially) particularly relevant to consumption and saving.

[8] Ballinger et al. (2003) used a discrete approximation of a CRRA utility function with $\sigma = 3$ and a consumption floor; this utility function *roughly* resembles a CRRA function with $\sigma \approx 2$ (and a consumption floor). Both coefficients are within ranges of estimates one sees in the survey-based econometric literature on consumption and saving. Experimentalists are used to seeing estimates of $\sigma \approx 0.5$ from their risky choice data, but those estimates are based on choices over relatively small monetary gains. Greater local curvature of the estimated function will be required over large absolute consumption levels to generate similar amounts of local risk aversion.

c_t	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	≥15
$u(c_t)$	0	26	51	65	77	89	101	104	107	110	111	112	113	114	115	116

If the subject's goal is to maximize expected total points accumulated by the end of a round, the problem faced by a subject in each period of a round is

$$\max_{c_t} u(c_t) + E_t \sum_{j=t+1}^{20} u(c_j),$$
$$\text{subject to} \quad A_{j+1} = X_j - c_j \quad \forall j = t, t+1, \ldots, 19, \ \text{given } X_t = A_t + y_t. \qquad (1)$$

This is a 20-period simplified version of what Browning and Lusardi (1996) dubbed the "standard additive model" of life-cycle consumption and saving. With the strict borrowing constraints and utility function specified above, the model yields precautionary motives for saving against future income uncertainty. The simple i.i.d. binomial income process, and the absence of utility discounting or returns on saving, make this simplified version of the standard additive saving game relatively transparent and easy to explain to subjects. Yet it is still computationally difficult: Such problems usually have no analytical solution (see, e.g., Deaton 1992), and *ex ante* optimal consumption policy functions $c_t^o(X_t, 20)$ for problem 1 can only be found by computational methods, using backward iteration and search. This is how we solve for these optimal policy functions.[9] Let $Y = \{y_1, y_2, \ldots, y_{20}\}$ denote any 20-period income sequence or "stream." Applied sequentially to any stream Y, $c_t^o(X_t, 20)$ will generate an *ex post* sum of point purchases over a game, denoted by $U^o(Y, 20)$.

We also define a class of boundedly rational policy functions for a 20-period game. Imagine *ex ante* optimal policy functions $c_p^o(X_p, \tau)$ for versions of problem 1 with final periods or "horizons" $\tau < 20$ and periods $p = 1, 2, \ldots, \tau$ (rather than periods $t = 1, 2, \ldots, 20$). One could apply those policy functions to stream Y of a 20-period game by setting $p = \max\{1, t + \tau - 20\}$ and substituting cash-in-hand X_t for X_p in the policy function $c_p^o(X_p, \tau)$. Let $U^o(Y, \tau)$ denote the *ex post* point total achieved by doing this. This is the point total of a player who optimally plans ahead at most $\tau - 1$ periods and ignores all periods beyond that. This "myopically optimal planning" is less computationally burdensome, and so defines a class of boundedly rational policies that are central to both our experimental design and data analysis.

1.2 Dependent measures: policy quality, policy consistency and performance

We can estimate subjects' *apparent horizons* of optimization τ^s from their observed 20-period vectors of consumption decisions. This is a convenient scalar summary of the structure of subject *behavior* and one way of characterizing their degree of

[9] Since everything is discrete (income process, randomness, subject consumption choices etc.) and finite (e.g. 20-period horizon), no approximation is needed: Computational solutions for the optimal policy functions are exact.

bounded rationality. We use the nonlinear least squares estimation of τ^s described by Ballinger et al. (2003); see the Appendix for a short summary. Each estimate $\hat{\tau}^s$ implies a *root mean squared error* $rmse^s$ of the consumption decisions of subject s around the consumption policy choices implied by $\hat{\tau}^s$, which we interpret as noise or inconsistency in the behavior of subject s. Together, $\hat{\tau}^s$ and $rmse^s$ are a bivariate characterization of subject behavior. These measures may be computed for each subject s assuming policy stability across all rounds r, or they may be computed separately for each round r played by each subject s (for instance, to decompose learning into policy improvement versus consistency improvement), in which case we write $\hat{\tau}_r^s$ and $rmse_r^s$. Ballinger, Palumbo and Wilcox note that there are strongly diminishing marginal gains to having a less myopic policy (higher τ), so we use $\ln(\hat{\tau}_r^s)$ as a dependent measure rather than $\hat{\tau}_r^s$ itself, calling it "policy quality." We call $-rmse_r^s$ "policy consistency" and also use it as a dependent measure.

A univariate summary measure of behavior is also useful. The most natural one is a suitable measure of *performance*. Performance measures, and forecasts of performance based on previous experimental results, play a large role in our experimental design, the manner in which we motivate subjects in the experiment, and in our data analysis as well. To begin, notice that $U^o(Y, 1)$ is the *ex post* point total of the policy $c_1^o(X_t, 1) \equiv X_t$ (given stream Y). This "nil policy" spends all of its cash-in-hand every period (as if the game lasts only one period). Since this policy does nothing to smooth consumption or accumulate precautionary stocks, its *ex post* point total $U^o(Y, 1)$ is an important baseline for measuring performance. Any subject who engages in *minimally* sensible consumption smoothing and precautionary saving should do no worse than$U^o(Y, 1)$.[10] At the other extreme, $U^o(Y, 20)$ is the *ex post* point total of the *ex ante* optimal policy $c_t^o(X_t, 20)$: No policy (and no subject) does *systematically* better than this.[11] Therefore, $U^o(Y, 20) - U^o(Y, 1)$ is the *ex post* point gain (given stream Y) of *ex ante* full rationality over nil rationality. If subject s earns $U^s(Y)$ total points from stream Y, her *actual* point gain over nil rationality is $U^s(Y) - U^o(Y, 1)$. We measure the performance of subject s with stream Y, written $P^s(Y)$, as the ratio of these two differences:

$$P^s(Y) = [U^s(Y) - U^o(Y, 1)]/[U^o(Y, 20) - U^o(Y, 1)]. \tag{2}$$

[10] There are some policies, such as "save and binge" policies, that perform systematically worse than the nil policy. These save everything for $k \geq 1$ periods, then spend everything, and repeat. On the basis of past results, we expected actual point totals below $U^o(Y, 1)$ to be rare. In fact it occurs more than once (in two or more rounds) for about 1 in 20 subjects and, as explained later, we omit such subjects from our statistical analysis.

[11] By "systematically," we mean that *expected* performance of *any* policy is definitionally no greater than the expectation of $U^o(Y, 20)$ *across all streams* Y. But suboptimal policies can by luck outperform the optimal policy for specific streams. The simplest way to see this is to consider a "clairvoyant policy" for some particular stream Y'(it is optimal given perfect foresight of Y'): Obviously such a policy is different from, and must outperform, the ex ante optimal policy applied to stream Y'. A subject could by luck use a policy on stream Y' that resembles the clairvoyant policy for stream Y', and so earn more points than the *ex ante* optimal policy does. We observe a small handful of subjects who earn more points than the optimal policy in some round of the saving game, as one could expect on the basis of these considerations.

1.3 Motivation of subjects

To motivate subjects, we truncate the performance measure (2) to the unit interval,[12] that is $P^{s*}(Y) = \max(0, \min[1, P^s(Y)])$. Let D be a maximum cash payment available for each round. We reward subject s by paying her $P^{s*}(Y) \cdot D$—a fraction of D equal to her truncated performance in the round.[13] Ballinger et al. (2003) observed that inexperienced median subjects plan ahead about two periods (optimal for a 3-period game). Therefore, we use $U^o(Y, 3)$ as a forecast of the *ex post* point total of a median inexperienced subject in our own saving game (given stream Y).[14] We use this prior information to design our game. Replacing $U^s(Y)$ (actual point total of subject s) with $U^o(Y, 3)$ in (2), we forecast performance by

$$PF(Y) = [U^o(Y, 3) - U^o(Y, 1)]/[U^o(Y, 20) - U^o(Y, 1)]. \tag{3}$$

Using computational methods, we chose our income process and the utility function $u(c_t)$ shown earlier to make the median value (across all possible income streams Y) of this "performance forecast" as small as possible, subject to the requirements that the utility function remains concave and that the income process remains a simple i.i.d. Bernoulli process (so that it is transparently explained to subjects). The resulting median performance forecast for the design is about 0.585. We do this to make a reasonable expectation of suboptimal behavior as painful to the subject as we can.[15] If this performance forecast is roughly accurate and D is large, most subjects will have something to learn across rounds of the saving game and an appreciable incentive to learn it. With $D = \$7.00$ per round, the forecast implies that the median inexperienced subject will leave an expected $\$7.00(1 - 0.585) = \$2.90$ "on the table" in the first round played. So, the expected opportunity cost of persisting in the forecast median behavior (not learning anything) across all five rounds would be $5 \cdot \$2.90 = \$14.50$ to the median subject.

Easily stated heuristic policies can perform much better than the median forecast behavior in the saving game. For instance, this heuristic policy captures a single paragraph of advice written by a subject who did rather well:

[12] The performance measure in (2) can occasionally fall outside of the unit interval for reasons elaborated in footnotes 10 and 11.

[13] Call this the *direct money method*. Another way is to pay D with probability $P^{s*}(Y)$, and zero otherwise: This is the well-known *binary lottery method*. The binary lottery method is more complex, and so less transparent to subjects; and its advertised incentive-compatibility is empirically suspect (Millner and Pratt 1992; Selten et al. 1999). We checked for significant differences in our dependent measures between the two methods in our first sample of 48 subjects (24 subjects were paid by each of these methods) and, finding none, used the direct money method in later samples.

[14] Half of our subject pool is identical to that used by Ballinger et al. (2003). The other half consists of students at a somewhat less selective university, so $U^o(Y, 3)$ might perhaps be somewhat above the expectation of $U^s(Y)$ there, but this also suits our purposes.

[15] In fact, this is also one of our main reasons for rewarding subjects according to the performance measure, rather than directly in terms of point totals. Other things equal, the "cost of misbehavior" (Harrison 1989) in an experiment (the opportunity cost of suboptimal behavior) should be as large as an experimenter can manage. Rewarding subjects directly in terms of their point total, rather than the *gain* of their point total over the nil policy as we do here, results in a *much* lower opportunity cost of suboptimal behavior for any given stake D.

Rule 1: If you are in the last period, or if your cash-in-hand is 2 or less in any period, then spend all of your cash-in-hand. If Rule 1 doesn't apply, go to Rule 2.

Rule 2: If you are in any of periods 16 to 19, spend half of your cash-in-hand (round up if half of cash-in-hand is not an integer). If Rules 1 and 2 don't apply, go to Rule 3.

Rule 3: If your cash-in-hand is less than or equal to 12, spend 2. Otherwise, spend 6.

Across all income streams used in our experiment, this policy has an average performance of about 0.915—well above the median performance forecast of 0.585 in our design, based on the results of Ballinger et al. (2003). That is, some subjects can write down relatively simple heuristic policies that are quite good (cf. Thaler 1994) in the sense that they improve substantially on what median subjects actually seem to have done in past experiments.[16]

Subject instructions included a detailed exercise meant to make the period utility function highly salient, and a test of subjects' understanding of the calculation of the truncated performance measure $P^{s*}(Y)$ which would determine their payment.[17] While subjects are paid for their performance in all five rounds, our design treats the first round of the saving game as extra instruction time: An experimenter sits with subjects during the first few periods of the first round to answer questions, and meets each subject at the end of the first round to gauge understanding and prompt for questions (after this, subjects are on their own). So we do not use first round behavior in our data analysis, and will present evidence below showing that this design feature was a sensible precaution.

Recall that the point totals $U^o(Y, 20)$ and $U^o(Y, 1)$ determine performance and payment. Subjects received no technical description of these totals. Rather, instructions described them as the point totals achieved by a good strategy (the "good score") and a poor strategy (the "poor score"), respectively, facing each income stream the subject would face. These two point totals were only revealed to a subject at the conclusion of each round, along with a computation of performance and earnings for that round.

Subjects know the *functional form* of their payoff function while playing each round: They understand that they start earning cash once their point total exceeds $U^o(Y, 1)$ and cease to earn cash should their point total exceed $U^o(Y, 20)$, and that between these extremes points convert to cash at the rate $D/[U^o(Y, 20) - U^o(Y, 1)]$. But they play each round in ignorance of $U^o(Y, 1)$ and $U^o(Y, 20)$—only learning these at the end of each round. Announcing $U^o(Y, 20)$ and $U^o(Y, 1)$ at the beginning of a round would reveal imperfect but potentially useful information about the coming income stream; and knowing if or when point totals exceed $U^o(Y, 20)$ would end extrinsic motivation for the rest of the round (though this rarely happens).

[16] All subjects wrote advice for a hypothetical future subject at the conclusion of their session. This was a pilot examination of written advice undertaken for planning a *future* experiment: Subjects were not rewarded for it, nor did we intend to use it for this study in any systematic way. So except for this specific example, which we use here for illustrative purposes, we do not formally analyze written advice in this paper.

[17] See our Supplementary Material for the subject instructions.

For those steeped in the salience and dominance precepts (Smith 1982) of the field, this is a controversial payoff method, so we need a good reason for using it. Note this feature of the method: There is a time, just after the subject has completed a saving game round with stream Y but before $U^o(Y, 1)$ and $U^o(Y, 20)$ have been revealed, when the subject has experienced the task but does not yet know how well she has done. We needed to pilot and vet a payoff method with this precise feature for future planned research on social learning processes. For boundedly rational agents, it can be hard to know one's own performance in a task. In the planned research, we wish to exogenously vary this self-knowledge as a treatment. In the planned research, subjects write advice for a future subject who will confront the same task, and we will exogenously vary: (1) whether that advice-writer does or does not know their own performance when writing advice; and (2) whether advice receivers do or do not know the advice-writer's performance when deciding how much credence to give to received advice. All of this can only be done using a payoff procedure that can keep subjects "blind" to their own performance (or someone else's performance) unless or until the experimenter wishes to reveal it to them. In the current experiment we find behavior and performance quite similar to what our performance forecast based on Ballinger et al. (2003) predicts. This suggests that our unusual payoff method has no appreciable unintended or adverse motivational consequences.

1.4 The design: specifics of the income streams

To measure learning in a model-free way, with first round behavior omitted by design, we present each subject with the *same income stream twice*, in the *second* and *fifth* rounds. The learning measure is $L^s = P^s_5(Y^s_l) - P^s_2(Y^s_l) : P^s_r(Y)$ is the performance of subject s in round r with income stream Y, and Y^s_l is the stream presented to subject s in both rounds two and five. It is possible that recognition of the round two stream could bias round five performance upward and inflate this measure of learning. However, statistically strong evidence of learning is apparent by round four, which is never a within-subject repetition of an earlier stream. Put differently, the design does not depend solely on the repeated income streams for evidence of learning, and the other evidence points to learning that is of a very similar strength and size.

Since our design deliberately repeats one stream for each subject, our subjects are *not* actually receiving truly random income draws across all periods and rounds. Other things equal, true randomness is desirable, but we believe that the goal of measuring learning in a model-free way is worth this deviation from standard practice—provided that our results resemble prior results in most essential respects, and we later argue that this is true. Since we deliberately repeat one stream for each subject, we should make the streams subjects actually receive approximately representative of the actual distribution of truly random 20-period binomial streams. While doing this, we also exploit the opportunity to create another design feature that allows us to measure the interaction between cognitive ability and the difficulty of the saving game.

To reduce between-subjects variance of the learning measure L^s attributable to streams (rather than subject behavior), we pre-selected the repeated streams Y^s_l so that while they vary across subjects, their "difficulty" is held constant in a specific sense defined in terms of the performance forecast in (3). We draw 10,000 truly random streams Y to approximate the actual distribution of $PF(Y)$, the performance forecast. The 90th, 50th and 10th quantiles of this approximate distribution are $PF_{90} = 0.73$, $PF_{50} = 0.585$ and $PF_{10} = 0.475$, and we regard streams Y with performance forecasts $PF(Y)$ near these three quantiles as relatively easy, moderate and hard streams, respectively.[18] We select three sets of streams with performance forecasts near these three quantiles: The set **E** (easy streams) is 16 streams with $PF(Y)$ near PF_{90}; the set **M** (moderate streams) is 16 streams with $PF(Y)$ near PF_{50}; and the set **H** (hard streams) is 16 streams with $PF(Y)$ near PF_{10}. Additionally, streams *within each set* are chosen so that the distribution of total income $\sum Y = \sum y_t$ across the 16 streams in each set resembles the true distribution of total income across truly random 20-period streams.[19]

Each subject plays five rounds of the saving game. Every subject experiences three rounds with moderate difficulty streams from **M**, one round with a hard stream from **H**, and one round with an easy stream from **E**. The experiences of all subjects are matched in this sense, although they receive streams with varying total income and experience the various "difficulties" in different orders. Each subject receives a single stream Y_l^s from **M** in their second and fifth round to measure learning unambiguously, and a different stream from **M** in either their first, third or fourth round. Additionally they receive a stream from **H** and a stream from **E** in their first, third or fourth round. The six possible orderings of stream difficulties across rounds one, three and four is balanced across subjects. This allows us see whether income stream difficulty (as represented by 10th, 50th and 90th quantiles of the performance forecast $PF(Y)$) conditions any relationship between cognitive ability and behavior.[20]

Across each sequence of 16 subjects and five game rounds, each income stream in **M** is used exactly twice, and each income stream in **H** and **E** is used exactly once. Since there are six distinct orders of income stream difficulty across rounds one, three and four), the minimum number of subjects required to have balanced variation of these orders (that is also a multiple of 16) is 48 total subjects. We call each such balanced group of 48 subjects a "sample," and we have four such samples for a total of 192 subjects. For pooled bivariate analysis of performance levels, as opposed to

[18] The "difficulty" of a stream mostly reflects the number and/or length of runs of "bad luck" (zero income draws) in it, conditional on total income received. Recall that the performance forecast is based on the use of a policy that optimally plans ahead for just two future periods. This policy performs increasingly poorly in the face of runs of three or more zero income draws.

[19] The set size (16 total income streams) was the minimum size needed to roughly approximate the true sampling distribution of total income in sequences of 20 income draws. The distribution of total income across streams in each set is: 1 stream with $\sum Y = 42$; 2 streams with $\sum Y = 48$; 3 streams with $\sum Y = 54$; 4 streams with $\sum Y = 60$; 3 streams with $\sum Y = 66$; 2 streams with $\sum Y = 72$; and 1 stream with $\sum Y = 78$. In our Supplementary Materials, we show how the resulting collection of 48 streams resembles the advertised income process in its sample distributions of total income and sample autocorrelation coefficients. We also show that subject experiences of the five income streams they actually receive give them no more evidence to reject the advertised (binomial, i.i.d.) income process than would true draws from the advertised income process.

[20] More specifically, by deliberately varying just three ordered difficulty levels, we can do this without making an assumption about the functional way (e.g. linear, quadratic, etc.) in which difficulty conditions any relationship between cognitive ability and our dependent measures.

learning, we will use average performance across rounds two through five (omitting the first round for reasons discussed above), $P^s = (P_2^s + P_3^s + P_4^s + P_5^s)/4$, as the overall performance of subject s.[21]

1.5 Cognitive and personality measures

Table 1 lists all cognitive and personality measures used in any of our samples. Our first two samples (96 subjects in all) are used to *select* two promising cognitive ability tests. Then, we retested those promising measures in the third and fourth samples (another 96 subjects) to *validate* their promise. This procedure guards against pretest bias. Three of the cognitive tests—the "Porteus Maze" (Porteus 1965) and two versions of "Raven's Standard Progressive Matrices" (Raven et al. 1998)—were discarded after the second sample because they showed less promise than the two tests described below.[22]

The "Beta III" (Kellogg and Morton 1999) is the third generation of a nonverbal test of cognitive abilities originally developed during the First World War to screen literate and illiterate United States military recruits on a more equal footing. In the first sample, we administered the entire Beta III test. Results there suggested that the sum of its two analytical subtests, "picture absurdities" and "matrix reasoning",[23] are the best predictor of saving performance, so we continued using these two subtests in the third and fourth samples.

The second test is one of those collectively known as "working memory" or WM span tests. Specifically, it is a recently developed computer-administered version of the "operation span" test of Turner and Engle (1989). All WM span tests have a common structure. A sequence of items are presented to the subject, with each presentation separated by a "distractor" task that requires conscious effort.[24] The subject must then reproduce the presented items in sequence. These tests are regarded as measuring the capacity for controlled allocation of attention and thought (Conway

[21] Because there is systematic variation across subjects in the difficulty of income streams presented in the third and fourth rounds, this is not a minimum variance measure of overall performance levels. As will be clear in panel analysis later, though, income stream difficulty (though highly significant) explains a relatively small amount of the variance in performance; so this does not turn out to be an empirically important matter here.

[22] See our Supplementary Material for details on these discarded tests.

[23] The Beta III test is proprietary, so we cannot show examples of its items. In the "picture absurdities" subtest, subjects identify a missing element in grayscale illustrations of people, objects, rooms, scenes, landscapes, etc. In the "matrix reasoning" subtest, subjects choose an icon from an offered collection that best fits a missing entry in a matrix of icons. Both are thus "pattern completion" tests—the first using natural, semantically rich image patterns, the second using abstract, nonmeaningful visual icon patterns. These portions of the Beta III were used under license from Harcourt Assessment. Today, the Beta III is available through Pearson Assessment.

[24] In the operation span test we use, letters are briefly presented in some sequence, interleaved with simple arithmetic equations that are either true or false. Subjects must ascertain the truth value of each equation as it appears, while maintaining the letters in memory. At the conclusion of each sequence of letters, subjects are asked to recall the presented sequence of letters in correct order. This basic task is repeated many times for letter sequences of varying length; overall test scores are based on the number of letters correctly remembered in their correct serial locations after each basic task. For more details, see Conway et al. (2005).

Table 1 Summary of cognitive, motivational and personality measures, and which samples have them

Measure class	Measure		Chief reference works and brief description	Samples		
				1st	2nd	3rd & 4th
Cognitive scales	Beta III		Kellogg and Morton (1999). Tests of nonverbal reasoning and cognitive functions. Used for nearly a century. We use reasoning part only (combination of scores on tests 2 and 5).	X		X
	Porteus maze		Porteus (1965). Maze-threading with pencil. Claimed to measure both planning ability and impulse control.	X		
	Raven SPM+		Raven et al. (1998). Nonverbal pattern induction or "matrix reasoning" test. Widely used in contemporary research on cognitive functioning and intelligence.	X		
	Raven SPM		Same as above, but somewhat abbreviated and simpler version (without the added upper-tail sensitivity of the "+" version)		X	
	Working memory span		Conway et al. (2005). The "operation span" test. The ability to control attention and thought in the face of processing load. A central measurement in contemporary cognitive psychology.		X	X
Intrinsic motivation	Need for cognition		Cacioppo et al. (1996). Item-response-base measure of intrinsic motivation to engage in effortful thought.		X	X
Personality scales	Procrastination		Tuckman (1991). Item-response-based measure of the tendency to procrastinate.		X	X
	Four dimensions of impulsivity	Premeditation	Whiteside and Lynam (2001). Four item-response-based measures of personality characteristics, each thought to either contribute to, or inhibit, impulsive behavior.		X	X
		Sensation-seeking			X	X
		Perseverance				X
		Urgency				X

et al. 2005). A large number of studies over the last two decades strongly suggest that WM span is a robust, domain-general predictor of intelligent performance. WM span tests correlate more strongly with performance in many different reasoning tests with widely varying surface features—such as the Raven family of tests, the Tower of Hanoi puzzle (e.g., McDaniel and Rutström 2001), the Beta III test and so forth—than performance in these same tests correlate with one another (Engle et al. 1999; Engle and Kane 2004). Based on promising results in the second sample, we continued measuring WM span in later samples.

The surveys in the second, third and fourth samples include an item response battery to measure several potentially relevant personality characteristics. Each item is a statement, and a subject decides how much each statement is descriptive of him or herself, selecting one of four responses (completely false, mostly false, mostly true or completely true). These are numerically coded with the integers 1, 2, 3 and 4 to ordinally concord with the characteristic being measured, and those integer codes are then summed across the relevant items to produce various "personality scales." These include a "need for cognition" scale, which is thought to measure intrinsic motivation to perform well in cognitively challenging tasks (Cacioppo et al. 1996) and a "procrastination" scale (Tuckman 1991). Four scales (premeditation, sensation-seeking, urgency and perseverance) thought to be implicated in impulsive behavior (Whiteside and Lynam 2001) are also measured (just the first two of these in the second sample, but all four in the third and fourth samples).[25] The collection of these personality scales is merely a precaution. Cognitive test scores might be correlated with personality factors, so we check whether our conclusions about cognitive abilities change when the scales are included in multivariate analyses.

All subjects were run *one at a time in individual sessions* that began with a survey and cognitive tests, both of which varied across samples. The first and third samples were collected at a large and diverse urban university, while the second and fourth samples were collected at a smaller and less diverse rural university. In our analyses, an experimental site dummy variable controls for the campus (unity for the urban campus and zero for the rural campus). Subjects were recruited by means of campus-wide advertisements in various media, and are all adult undergraduates. All subjects received a flat \$5.00 payment for showing up. Subjects received an additional flat \$10.00 payment for completing the surveys, item-response personality measurements, and cognitive tests (lasting about 60 to 75 minutes) and were then encouraged to take a short break if desired. Saving game instructions were then presented, and subjects completed five rounds of the game (all taking about 40–50 minutes). The available reward D per round was \$7.00, giving a maximum total payment from the five saving games equal to \$35.00.[26]

2 Results

In the analyses to follow, we pool samples containing the same cognitive and personality measures. While no measure is common to all four samples, we do examine

[25] See our Supplementary Material online for more details on each of these personality scales.

[26] The cognitive tests were more time-consuming for the first sample of 48 subjects. These subjects received \$15.00 for completing the tests rather than \$10.00, and had $D = \$8.00$ rather than \$7.00.

an all-samples pooling to establish certain overall facts and treatment effects and examine the demographic variables gender, age and experimental site. The other three poolings are: (1) The first, third and fourth samples, the broadest pooling where the analytical component of the Beta III test was measured; (2) The second, third and fourth samples, the broadest pooling where WM span was measured (as well as need for cognition, procrastination and two of the impulsiveness subscales); and (3) The third and fourth samples, where we have both cognitive measures and all personality scales. Pretesting ends after the second sample: It is important to keep in mind that results based only on the third and fourth samples are the methodologically cleanest results, since pretest bias is present in results that pool observations from the first and second samples.

There were nine subjects (out of 192 total subjects) with *negative* values of the average or overall performance measure P^s and they are a methodological problem. Almost all of them failed to earn any money in two or more rounds of the saving game, so we may have lost control over their motivation. Their round-to-round performance is also highly erratic: Adding these nine subjects to the panel regressions described below *triples* estimated residual variance. Therefore, we omit these nine subjects (roughly 1 in 20 subjects) from all of our statistical analyses below.[27]

2.1 Policy quality, policy consistency and performance

Table 2 shows Pearson correlations (across subjects) between performance measures P_r^s in different rounds r, illustrating two points. First, the correlations are quite strong across rounds two through five (all significant at $p < 0.0001$): Between-subject performance differences are reliable and, therefore, there is plenty of potentially predictable variance of performance across subjects. Second, correlations between first round performance and performance in later rounds is noticeably weaker (though all significant at $p < 0.01$). First round performance is also noticeably more variable than in later rounds. This suggests that subjects are still developing an understanding of saving game mechanics during the first round, as we anticipated. So, as originally planned, we do not use the first round data for any further analysis.

Overall performance P^s is the average of performance P_r^s across four rounds, so we can use Cronbach's alpha (Cronbach 1951) to estimate the reliability of P^s: It is easily calculated from the correlations in Table 2, and equals 0.87 here.[28] Under the assumptions of classical test theory, Cronbach's alpha is an unbiased estimate of the maximum possible R^2 one may obtain in a regression of P^s on *any* set of predictors.[29] This is a useful result because it allows us to see how well performance can be decomposed as a linear function of policy quality and policy consistency. The

[27] Several of these unexpectedly poor-performing subjects used something very like a "save-and-binge" strategy that performs worse than the nil policy (see supra footnote 10). Others apparently had a *minimum* target level of assets (buffer stock on its head) that they maintained even when this was plainly (and painfully) dominated, for instance spending zero in periods where their cash-on-hand had fallen to their apparent (positive) target level of assets.

[28] Let $\bar{\rho}$ be the mean of $N(N-1)/2$ bivariate Pearson correlations between N items summed to give an overall measure; then Cronbach's alpha for that measure is equal to $N\bar{\rho}/[1+(N-1)\bar{\rho}]$.

[29] The key "assumption of classical test theory" is that the items summed to give the overall measure are one-dimensional in the factor-analytic sense that items measure the same underlying ability. Thinking

Table 2 Pearson correlations between performance P_{sr} in different rounds r

	Round 2	Round 3	Round 4	Round 5
Round 1	0.366	0.280	0.250	0.235
Round 2	–	0.589	0.619	0.661
Round 3	–	–	0.624	0.597
Round 4	–	–	–	0.674

Notes: Results are for all four samples, subjects = 183. All correlations are highly significant

R^2 from a bivariate regression of overall performance P^s on policy quality $\ln(\hat{\tau}^s)$ and policy consistency $-rmse^s$ is 0.74. The ratio of the $R^2 = 0.74$ to the Cronbach's alpha of 0.87 is 0.85: Thus policy quality and policy consistency, as defined here in terms of estimated horizons of optimization and residual sums of squares, together account for about 85% of the reliable and explainable variance in overall subject performance.

We conclude that performance may be thought of as approximately decomposable into policy quality and policy consistency as measured here. Put differently, the univariate behavioral measure (performance) and the bivariate behavioral characterization (policy quality and policy consistency) are different, but approximately equivalent, low-dimensional characterizations of each subject's high-dimensional (20-period vector) consumption decisions. What the bivariate characterization allows is a finer-grained analysis of predictors of heterogeneity. We will see that one cognitive ability measure seems to mostly explain variations in policy quality, while another seems to mostly explain variations in policy consistency. However, we think it is simplest and most natural to begin by analyzing the univariate dependent measure, which is performance.

2.2 Performance

Mean performance P_r^s in rounds $r = 2, 3, 4$ and 5 is 0.59, 0.60, 0.65 and 0.64, respectively. The mean and standard deviation of overall performance P^s are 0.62 and 0.22, respectively; its minimum and maximum are 0.060 and 0.996. Thus there is great variability in subject performance, as expected on the basis of past work (Ballinger et al. 2003) and as hoped for our purposes. The mean of the learning measure L^s (change in performance between the second and fifth rounds) is 0.052. This is significantly positive ($p = 0.0009$ by a one sample t-test), but it is a small effect (about a

of each saving game round as a test item, the first principal component of the four round performance measures (for rounds 2 to 5) accounts for 72% of their collective variance, so this assumption seems reasonable. Write $P^s = \bar{P}^s + \varepsilon$, where $\bar{P}^s$ is a true nonstochastic but unobserved ability of subject s in these saving games, and ε is i.i.d. error in measuring this, due to the randomness of income streams and/or subject execution. Under the classical assumptions, Cronbach's alpha is an estimate of the correlation *alpha* between two identical but separate measurements of P^s (here $\bar{P}^s$ is unchanged but there are new draws of ε). It is then easy to show that $alpha \equiv \mathrm{Var}(\bar{P}^s)/\mathrm{Var}(P^s)$. This is the expected explainable sum of squares divided by the expected total sum of squares, since $\bar{P}^s$ is the nonstochastic, repeatable and predictable part of P^s. Hence, Cronbach's alpha estimates the maximum possible R^2 in a regression of P^s on any set of regressors.

quarter of the standard deviation of P^s) and L^s is almost as variable across subjects (standard deviation of 0.21) as are average performance levels. Learning is significant by round 4 (which is not a within-subject repetition of an earlier stream), so memory of the round 2 income stream is an insufficient explanation of observed learning. For instance, the mean is of $P_4^s - P_2^s$ across all subjects s is 0.059 (significantly positive by a one-sample t-test, $p = 0.0003$). Confining attention to the 62 subjects who received a moderate difficulty stream in round 4 (recall that only one out of three round 4 streams are of moderate difficulty while all round 2 streams are), the mean $P_4^s - P_2^s$ is 0.063 (significantly positive by a one-sample t-test, $p = 0.027$).

Table 3 shows bivariate relationships between average performanceP^s, demographic variables and our measures of cognitive abilities and personality scales, in

Table 3 Spearman rank correlations between average performance levels and demographic, cognitive and personality variables in various poolings of samples

Variable class	Variable	All samples	Samples 1, 3 & 4	Samples 2, 3 & 4	Samples 3 & 4
		Subjects = 183	Subjects = 139	Subjects = 135	Subjects = 91
Demographic variables	Female	−0.14* $p = 0.059$	−0.14* $p = 0.088$	−0.19** $p = 0.024$	−0.22** $p = 0.033$
	Age	0.031 $p = 0.67$	−0.0058 $p = 0.12$	0.084 $p = 0.33$	0.073 $p = 0.49$
	Urban Campus	0.089 $p = 0.23$	0.12 $p = 0.16$	0.13 $p = 0.12$	0.17^{*} $p = 0.099$
Cognitive scales	Beta III analytical	–	0.28^{***} $p = 0.0009$	–	0.24^{**} $p = 0.021$
	WM span	–	–	0.31^{***} $p = 0.0002$	0.22^{**} $p = 0.033$
Personality scales	Need for cognition	–	–	0.16^{*} $p = 0.061$	0.11 $p = 0.29$
	Procrastination	–	–	0.075 $p = 0.39$	0.021 $p = 0.84$
	Premeditation	–	–	−0.094 $p = 0.28$	−0.12 $p = 0.26$
	Sensation-seeking	–	–	0.19^{**} $p = 0.031$	0.14 $p = 0.19$
	Perseverance	–	–	–	−0.13 $p = 0.23$
	Urgency	–	–	–	−0.060 $p = 0.57$

Notes: *, **, and *** indicate significance at $\alpha = 10\%$, 5% and 1%, respectively the form of the nonparametric Spearman (rank) correlation. In all sample poolings, women perform significantly worse than men, but this significance vanishes in most of our multivariate analyses (to come shortly). Under the broadest pooling containing most of the personality scales (the second, third and fourth samples), need for cognition and sensation-seeking are significantly related to performance. The positive relationship between need for cognition and performance is expected if intrinsic motivation contributes to performance. However, the positive correlation of performance with sensation-seeking is not the expected one if sensation-seeking contributes to impulsiveness and impulsiveness is a negative influence on saving performance. Note, however, that these significant correlations vanish in the third and fourth samples alone. Moreover, neither need for cognition nor sensation-seeking are significant in the multivariate analyses that follow. Finally, the cognitive scales (the analytical part of the Beta III, and WM span) are significant in every pooling where they are available. As will be clear shortly, this significance survives in multivariate analyses, somewhat shifting the burden of proof to those who would maintain that cognitive ability differences are merely instruments for some other difference.[30]

Table 4 shows Spearman correlations between performance and cognitive ability, broken down by the income stream difficulty treatment. The relationship between

Table 4 Spearman rank correlations between average performance levels and cognitive measures, by income stream difficulty

Income stream difficulty	Cognitive scale	Samples 1, 3 & 4 Subjects = 139	Samples 2, 3 & 4 Subjects = 135	Samples 3 & 4 Subjects = 91
Easy $PF(Y) \approx 0.73$	Beta III analytical	0.20^{*} $p = 0.053$	–	0.25^{*} $p = 0.053$
	WM span	–	0.36^{***} $p = 0.0004$	0.30^{**} $p = 0.018$
Moderate $PF(Y) \approx 0.585$	Beta III analytical	0.22^{***} $p < .0001$	–	0.20^{***} $p = 0.0021$
	WM span	–	0.27^{***} $p < .0001$	0.17^{***} $p = 0.0065$
Hard $PF(Y) \approx 0.475$	Beta III analytical	0.32^{***} $p = 0.0020$	–	0.15 $p = 0.26$
	WM span	–	0.24^{**} $p = 0.025$	0.16 $p = 0.21$

Notes: *, **, and *** indicate significance at $\alpha = 10\%$, 5% and 1%, respectively

[30] As noted earlier, income stream difficulty in rounds 3 and 4 varies across subjects, so this is an experimentally induced part of the total variance of P^{s}. However, a similar bivariate analysis in which income stream difficulty effects are first partialled out of P^{s} is essentially identical in all respects to what we discuss here.

cognitive ability and performance seems robust across the samples only for easy and moderate difficulty income streams. For the hard income streams, the positive relationship observed elsewhere in this table is not significant in the 3rd and 4th samples (though it is in broader poolings). Camerer and Hogarth (1999) view cognitive capital as procedural knowledge and suggest that particularly hard problems may challenge the knowledge of even the most able subjects. This may be why variations in our subjects' cognitive abilities robustly explain variations in performance with easy and moderate income streams, but not so robustly with the hard income streams.

We now turn to multivariate analyses. For this purpose, we treat the experimental data as a panel with four repeated measurements of performance (in each of rounds 2 through 5). We use a random effects estimator to account for any reliable between-subject variance of performance levels that remains after controlling for observed differences between the subjects (demographic, cognitive and personality differences) and treatment variation within and between subjects (the difficulty of, and total income in, the current income stream and the previous income stream). Besides accounting for the lack of independence between the repeated measurements on each subject, this also allows us to say how much of the potentially predictable variance in performance is explained by the various measures, treatments and learning. All standard errors and tests are computed using the heteroscedasticity-robust "sandwich" estimator.[31] Degrees of freedom are adjusted downward for tests concerning all effects that vary strictly between subjects, such as demographic, cognitive and personality measures.

Table 5 shows most of the results of the panel analyses. The parameterization of the model and linear transformations of its regressors make the intercept interpretable as the mean round 2 performance of the average male subject at the rural campus, when facing a moderate difficulty income stream with total income of 60.[32] Recall that the forecast performance for such income streams, on the basis of previous work (Ballinger et al. 2003), is about 0.585: The estimated intercepts are very close to this. This is remarkable since this experiment differs from that of Ballinger, Palumbo and Wilcox in several ways: (a) subject population; (b) motivational mechanism; (c) utility function; (d) length of the saving game; and (e) a very different software interface and instructional protocol. In spite of these differences, the one-parameter "apparent horizon" model with $\tau = 3$, estimated on the basis of the earlier experiment, predicts mean inexperienced (second round) subject performance in this new experiment quite well. We also used a motivational scheme in which subjects do not learn the "good score" and "poor score" until each round concludes, so that marginal incentives are uncertain during play. This does not seem to have a deleterious effect on behavior and performance.

[31] We find no evidence of significant autocorrelation of the residuals from these models. There is some weak evidence of heteroscedasticity conditioned on the cognitive ability measures though (as one might expect, lower variance with higher cognitive ability), so the precaution of the sandwich estimator seems appropriate.

[32] Specifically, variables are transformed in the following ways. Performance forecasts *PF* (higher means an "easier" income stream) are differenced from 0.585—the approximate difficulty of all moderate difficulty income streams. Total income $\sum Y$ is differenced from 60, its average sample value; and age, and all cognitive and personality scales, are standardized using the sample mean and sample standard deviation calculated within each pooling of samples.

Table 5 Random effects panel regressions of performance P_r^s on trials, treatments and subject characteristics in various combinations of samples: estimates and significance tests

Regressor class	Regressor	All samples Subjects = 183		Samples 1, 3 & 4 Subjects = 139		Samples 2, 3 & 4 Subjects = 135		Samples 3 & 4 Subjects = 91	
		Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests
Intercept	Intercept	0.60^{***} (0.032)	–	0.56^{***} (0.039)	–	0.64^{***} (0.032)	–	0.61^{***} (0.045)	–
Learning	Round 3 difference	0.0063 (0.018)	Yes (***)	0.021 (0.021)	Yes (***)	−0.0039 (0.020)	Yes (*)	0.014 (0.025)	Yes (***)
	Round 4 difference	0.055^{***} (0.016)		0.083^{***} (0.019)		0.037^{**} (0.018)		0.070^{***} (0.022)	
	Round 5 difference	0.052^{***} (0.015)		0.067^{***} (0.018)		0.035^{**} (0.018)		0.052^{**} (0.023)	
Treatments (properties of current and once-lagged income streams)	$PF(Y_r)$ (***)	0.38^{***} (0.086)	Yes (***)	0.41^{***} (0.097)	Yes (***)	0.35^{***} (0.097)	Yes (***)	0.39^{***} (0.11)	Yes (**)
	$PF(Y_{r-1})$	0.014 (0.065)		0.030 (0.076)		−0.0065 (0.074)		0.0060 (0.093)	
	$\sum Y_r$	0.000029 (0.00082)		−0.00046 (0.00093)		0.00036 (0.00092)		−0.00019 (0.0011)	
	$\sum Y_{r-1}$	0.00079 (0.00075)		0.00099 (0.00083)		0.00092 (0.00092)		0.0013 (0.0011)	
Demographic variables	Female	−0.053 (0.033)	No	−0.037 (0.036)	No	-0.081^{**} (0.039)	No	−0.077 (0.050)	No
	Age	0.014 (0.014)		−0.016 (0.012)		0.018 (0.016)		−0.0052 (0.015)	
	Urban campus	0.024 (0.033)		0.047 (0.037)		0.0049 (0.044)		0.033 (0.052)	
Cognitive scales	Beta III analytical	–	–	0.075^{***} (0.018)	–	–	–	0.052^{**} (0.021)	Yes (***)
	WM span	–		–		0.079^{***} (0.017)		0.045^{**} (0.020)	

Moreover, the performance forecast $PF(Y_r)$ is a highly significant predictor of performance, indicating that our scheme for rating income stream difficulty has some aggregate value. Still, if the performance forecast was an unbiased predictor, its coefficient would be unity; and this hypothesis is easily rejected. We believe this occurs (in part) because the performance forecast assumes a common $\tau = 3$ apparent horizon model for all subjects. As a result, its predictive value is probably attenuated by policy heterogeneity across subjects.

Table 5 (*Continued*)

Regressor class	Regressor	All samples Subjects = 183		Samples 1, 3 & 4 Subjects = 139		Samples 2, 3 & 4 Subjects = 135		Samples 3 & 4 Subjects = 91	
		Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests
Personality scales	Need for cognition	–	–	–	–	0.026 (0.016)	No	0.0082 (0.021)	No
	Procras-tination	–		–		0.025 (0.019)		−0.0063 (0.027)	
	Premed-itation	–		–		−0.0031 (0.022)		−0.0016 (0.028)	
	Sensation-seeking	–		–		0.028 (0.023)		0.011 (0.029)	
	Perseverance	–		–		–		−0.040* (0.023)	
	Urgency	–		–		–		−0.020 (0.023)	

Notes: *, **, and *** indicate significance at $\alpha = 10\%, 5\%$ and 1%, respectively

When a subject completes a saving game round with a relatively hard income stream (or one with relatively low total income), it is possible that he learns more from that experience, or otherwise exercises more caution, in the next round than does a subject who faced a relatively easy income stream (or relatively high total income). Yet we find no evidence of this. Once-lagged performance forecasts $PF(Y_{r-1})$ and total incomes $\sum Y_{r-1}$ are insignificant in all regressions. Put differently, there is no evidence that performance depends on treatment history (order effects). Finally, as we hoped would be true given the way we measure performance, current round total income $\sum Y_r$ has no significant effect on current round performance either.

The significant gender effect in the bivariate analyses vanishes in the multivariate analysis and, as a group, the demographic variables (gender, age and site) are jointly insignificant. We note that this is not because women and men in these samples differ systematically in their measured cognitive abilities: There is no significant difference in the WM span or Beta III scores of women and men in any pooling of the samples. Additionally, as mentioned earlier, the significance of personality scales mostly vanishes in the multivariate analysis. While perseverance is marginally significant in the third and fourth samples where it is measured, its negative sign is the opposite of what one would expect if perseverance inhibits impulsiveness and impulsiveness is a negative influence on saving performance.

Finally, it is very clear that, wherever they are found, the cognitive measures are highly significant predictors of performance. Even in the small pooling (just the third and fourth samples) where both measures are available and hence in the regression together, they are both individually significant (and so not mutually redundant, even though positively correlated).

In terms of explained variance, the Beta III and WM span alone explain about 19% of the potentially predictable between-subjects variance (estimated as the variance of random effects in the model with no demographic, cognitive or personality measures). This is not as large as one might like, but it is by no means trivial: Their significance is practical as well as statistical.

2.3 Policy quality and policy consistency

Policy quality is the natural logarithm $\ln(\hat{\tau}_r^s)$ of the nonlinear least squares estimate $\hat{\tau}_r^s$, the estimated horizon of optimization of subject s in round r; while policy consistency is the negative of the root mean squared error $-rmse_r^s$ from these nonlinear least squares regressions. Table 6 shows the results of panel regressions (identical in all respects to those performed for Table 5) of these two dependent measures, using just the data from the third and fourth samples. The first column of Table 6 simply repeats the last column of Table 5, so that the panel regressions of performance, policy quality and policy consistency may be compared with ease.

There are two interesting findings in Table 6. The first concerns learning. Recall that there is strongly significant evidence of learning (though the effect is small) in that performance rises significantly between the second and fifth rounds. This does *not* appear to be because subjects are learning a better policy: The center column of Table 6 shows no evidence of a significant rise in apparent horizons of optimization across rounds. Instead, subjects learn to apply a roughly unchanging policy more consistently: The right column of Table 6 shows that policy consistency improves significantly across rounds. In this experiment, performance improvements across rounds are mostly due to a reduction of noise rather than the improvement of structural policy quality. The second interesting finding is that the two cognitive ability measures appear to explain different sources of cross-sectional variation in performance. Working memory span significantly explains variations in policy quality but not policy consistency. The reverse is true for the analytical components of the Beta III, which significantly explain cross-sectional variations in policy consistency but not policy quality.

3 Discussion and conclusions

Cognitive abilities are the best predictors of saving performance in our game. This agrees with the evidence reviewed by Stanovich and West (2000), showing that substantial variance of most (though not all) classic experimental failures of reasoning are explained by measures of cognitive abilities. It also agrees with recent suggestive results of Rydval and Ortmann (2004), Frederick (2005) and Benjamin et al. (2006). Multivariate controls for sex, age, experimental site and various personality scales (intrinsic motivation, procrastination and impulsiveness) do not eliminate the significance of cognitive abilities. The computational complexity of saving problems, interacting with heterogeneous cognitive constraints of savers, is a likely determinant of differences in saving behavior across individuals.

On a more general note, measures of cognitive abilities are likely to find many uses in experiments. We discuss WM span since we are relatively familiar with it and

Table 6 Random effects panel regressions of apparent horizons and policy consistency on trials, treatments and subject characteristics in validation samples (samples 3 and 4): estimates and significance tests

Dependent variable →		Performance P_r^s		Policy quality, $\ln(\hat{\tau}_r^s)$		Policy consistency, $-rmse_r^s$	
Regressor class	Regressor	Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests
Intercept	Intercept	0.61^{***} (0.045)	–	1.38 (0.12)	–	-1.19^{***} (0.077)	–
Learning	Round 3 difference	0.014 (0.025)	Yes (***)	−0.041 (0.067)	No	0.085 (0.054)	Yes (***)
	Round 4 difference	0.070^{***} (0.022)		−0.026 (0.067)		0.18^{***} (0.050)	
	Round 5 difference	0.052^{**} (0.023)		−0.082 (0.069)		0.17^{***} (0.052)	
Treatments (properties of current and once-lagged income streams)	$PF(Y_r)$	0.39^{***} (0.11)	Yes (**)	-0.63^{*} (0.33)	Yes (***)	0.057 (0.26)	No
	$PF(Y_{r-1})$	0.0060 (0.093)		−0.032 (0.27)		−0.12 (0.19)	
	$\sum Y_r$	−0.00019 (0.0011)		0.0099^{***} (0.0025)		−0.0025 (0.0020)	
	$\sum Y_{r-1}$	0.0013 (0.0011)		0.00014 (0.0024)		0.0015 (0.0024)	
Demographic variables	Female	−0.077 (0.050)	No	−0.18 (0.12)	No	−0.050 (0.065)	No
	Age	−0.0052 (0.015)		−0.036 (0.039)		0.020 (0.023)	
	Urban campus	0.033 (0.052)		−0.0082 (0.13)		0.089 (0.067)	
Cognitive scales	Beta III analytical	0.052^{**} (0.021)	Yes (***)	0.054 (0.056)	Yes (**)	0.074^{***} (0.026)	Yes (***)
	WM span	0.045^{**} (0.020)		0.11^{**} (0.052)		0.015 (0.025)	

because of its ubiquity in contemporary psychological research on cognitive functioning. Some argue that WM span taps a domain-general capacity for controlled attention (Kane et al. 2004). Psychologists use managerial metaphors when discussing it, speaking of its "supervisory" or "executive" functions (Engle 2002): It predicts in-

Table 6 (*Continued*)

Dependent variable →		Performance P_r^s		Policy quality, $\ln(\hat{\tau}_r^s)$		Policy consistency, $-rmse_r^s$	
Regressor class	Regressor	Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests	Estimate (std. err.)	Joint tests
Personality scales	Need for cognition	0.0082 (0.021)	No	−0.0025 (0.012)	No	0.0028 (0.0040)	No
	Procrastination	−0.0063 (0.027)		−0.00056 (0.013)		−0.00061 (0.0073)	
	Premeditation	−0.0016 (0.028)		0.0055 (0.015)		−0.0094 (0.0072)	
	Sensation-seeking	0.011 (0.029)		0.0017 (0.0087)		−0.0026 (0.0047)	
	Perseverance	−0.040* (0.023)		−0.010 (0.016)		−0.0032 (0.0083)	
	Urgency	−0.020 (0.023)		−0.0091 (0.012)		0.0018 (0.0071)	

Notes: *, **, and *** indicate significance at $\alpha = 10\%$, 5% and 1%, respectively

telligent performance and problem-solving ability. Shah and Miyake (1999) remark that much complex cognition "involve[s] multiple steps with intermediate results that need to be kept in mind temporarily to accomplish the task at hand successfully" and identify the capacity to handle this, while processing information, as working memory.

Seen in this manner, decision performance may be mediated in basic and deep ways by working memory capacity. For instance, the apparent failure of asset integration (Kahneman and Tversky 1979), an important normative requirement in dynamic decisions under risk, might be mediated by working memory, since asset integration (and generally, normatively desirable "broad decision bracketing" of all kinds) requires simultaneous attention to outcomes of old decisions and features of current and future ones (Read et al. 1999). WM capacity might predict depth of reasoning in k-step reasoning and cognitive hierarchy theories of game play (Nagel 1995; Stahl and Wilson 1995; Camerer et al. 2004). In fact, WM capacity might interact with game complexity to predict how individual subjects *change* their depth of reasoning across games with increasingly complex structural features.

WM capacity helps us ignore distractions and focus on what is important. Because of this, WM span may play a mediating role between effortful, conscious processes and automatic processes in "dual process" theories of the mind (Feldman-Barrett et al. 2004). For behavioral theorists who stress the role of outputs of automatic processes, such as immediate affective reactions to alternatives and events, differences in WM span may therefore be important in determining their ultimate impact on behavior. For instance, consider immediate emotional reactions to "unfair" actions in some repeated game with random rematching to new partners in each period. It might be sensible to inhibit the force of those emotions on immediate future actions since (due to rematching) the subject is about to meet a new partner. “Low spans” may find this more difficult than “high spans” if their more severe attention resource constraints make it more difficult to inhibit emotional reactions. WM capacity might mediate the frequency of retaliatory actions (as well as other actions based on automatic emotional responses) in games.

The discussion above is meant to be illustrative rather than exhaustive. There are many phenomena of interest to behavioral and experimental economists that may be mediated in an important way by executive control of attention. If so, WM span measures could be an extremely useful “cognitive capital” measure for research in those areas. We believe the present study illustrates its promise.

Appendix: Estimation of apparent horizons of optimization and policy consistency

The nonlinear least squares estimation of horizons of optimization is taken from Ballinger et al. (2003). Let $Y_{rt}^{s} = \{y_{r1}^{s}, y_{r2}^{s}, \ldots, y_{rt}^{s}\}$, X_{rt}^{s} and c_{rt}^{s} denote the actual income history, observed cash-in-hand and consumption choice (respectively) of subject s in period t of round r. We use computational methods (Deaton 1992) to find $c_{p}^{o}(X_{p}, \tau)$, the optimal policy for problem 2, for all $\tau \leq 20$. Let $X_{rt}^{o}(Y_{rt}^{s}, \tau)$ be the cash-in-hand that policy $c_{p}^{o}(X_{p}, \tau)$ holds in period t, given that it is applied to the income sequence Y_{rt}^{s} with $p = \max\{1, t + \tau - 20\}$. Then the nonlinear least squares estimator of $\hat{\tau}_{r}^{s}$, the subject’s “apparent horizon (of optimization) in round r,” is

$$\hat{\tau}_{r}^{s} \equiv \underset{\tau\in\{1,2,\ldots,20\}}{\arg\min} \left(\sum_{t=1}^{T} [e_{rt}^{s}(\tau)]^{2} \right) \Big/ 20$$

$$\text{where } e_{rt}^{s}(\tau) = c_{rt}^{s} - c_{p}^{o}[X_{rt}^{o}(Y_{rt}^{s}, \tau)] \text{ and } p = \max\{1, t + \tau - 20\}.$$

Then, $rmse_{r}^{s}$ is simply the square root of the value of this objective function evaluated at $\hat{\tau}_{r}^{s}$. To estimate apparent horizons $\hat{\tau}^{s}$ (assuming a constant policy across rounds 2 to 5, that is no learning across rounds), the sum in the minimand is taken across all four rounds and divided by 80 rather than 20: $rmse^{s}$ is the square root of the value of this objective function evaluated at $\hat{\tau}^{s}$.

The shifted period index $p = \max\{1, t + \tau - 20\}$ of the policy function in the minimization accounts for the fact that policy functions with $\tau < 20$ are “myopic.” For instance, with $\tau = 3$, the policy function behaves as if it is always beginning a three-period game ($p = 1$) for all true game periods $t = 1$ to 17. Thereafter, in true periods $t = 18$, 19 and 20, p is 1, 2 and 3 (respectively): The myopic policy function with $\tau = 3$ is only truly optimal in these last periods of the true game.

References

Ballinger, T. P., Palumbo, M. G., & Wilcox, N. T. (2003). Precautionary saving and social learning across generations: An experiment. *Economic Journal*, *113*, 920–947.

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, *3*, 439–449.

Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2006). Who is "behavioral"? Cognitive ability and anomalous preferences. Harvard University working paper.

Brown, A. L., Chua, Z. E., & Camerer, C. F. (2009). Learning and visceral temptation in dynamic saving experiments. *Quarterly Journal of Economics*, *124*, 197–231.

Browning, M., & Lusardi, A. (1996). Household saving: Micro theories and micro facts. *Journal of Economic Literature*, *34*, 1797–1855.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: the life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253.

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: a review and capital-labor-production framework. *Journal of Risk and Uncertainty*, *19*(1–3), 7–42.

Camerer, C. F., Ho, T.-H., & Chong, K. (2004). A cognitive hierarchy model of behavior in games. *Quarterly Journal of Economics*, *119*(3), 861–898.

Campbell, J. Y., & Mankiw, N. G. (1990). Permanent income, current income and consumption. *Journal of Business and Economic Statistics*, *8*, 265–279.

Carbone, E. (2006). Understanding intertemporal choices. *Applied Economics*, *38*, 889–898.

Carbone, E., & Hey, J. D. (2004). The effect of unemployment on consumption: an experimental analysis. *Economic Journal*, *114*, 660–683.

Carroll, J. B. (1993). *Human cognitive abilities: a survey of factor-analytic studies*. Cambridge: Cambridge University Press.

Conlisk, J. (1980). Costly optimization versus cheap imitators. *Journal of Economic Behavior and Organization*, *1*, 275–293.

Conlisk, J. (1996). Why bounded rationality? *Journal of Economic Literature*, *34*, 669–700.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: a methodological review and user's guide. *Psychonomic Bulletin and Review*, *12*, 769–786.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Deaton, A. S. (1992). *Understanding consumption*. London: Oxford University Press.

Dyer, D., & Kagel, J. H. (1996). Bidding in common value auctions: how the commercial construction industry corrects for the winner's curse. *Management Science*, *42*, 1463–1475.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science*, *11*, 19–23.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 44, pp. 145–199). Amsterdam: Elsevier.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: a latent variable approach. *Journal of Experimental Psychology: General*, *128*, 309–331.

Feldman-Barrett, L., Tugade, M. M., & Engle, R. W. (2004). Individual differences in working memory capacity and dual-process theories of the mind. *Psychological Bulletin*, *130*, 553–573.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25–42.

Gabaix, X., & Laibson, D. (2000). A boundedly rational decision algorithm. *American Economic Review*, *90*, 433–438.

Hall, R. E., & Mishkin, F. S. (1982). The sensitivity of consumption to transitory income: estimates from panel data on households. *Econometrica*, *50*, 461–481.

Harrison, G. (1989). Theory and misbehavior of first-price auctions. *American Economic Review*, *79*, 749–762.

Harrison, G., & List, J. A. (2004). Field experiments. *Journal of Economic Literature*, *42*(4), 1009–1055.

Hey, J., & Dardanoni, V. (1988). Optimal consumption under uncertainty: an experimental investigation. *Economic Journal*, *98*(390), 105–116 (supplement).

Hogarth, R. (1975). Decision time as a function of task complexity. In D. Wendt & C. Vlek (Eds.), *Utility, probability and human decision making* (pp. 321–338). Dordrecht: Reidel.

Hunt, E. (1999). Intelligence and human resources: past, present and future. In P. L. Ackerman, P. Kyllunon, & R. Roberts (Eds.), *Learning and individual differences: process, trait and content determinants* (pp. 3–30). Washington: American Psychological Association.

Kahneman, D., Tversky, A. (1979). Prospect theory: an analysis of decision under risk. *Econometrica*, *47*, 263–291.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: an individual-differences perspective. *Psychonomic Bulletin and Review*, *9*, 637–671.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working-memory capacity: a latent-variable approach to verbal and visuo-spatial memory span and reasoning. *Journal of Experimental Psychology: General*, *133*(2), 189–217.

Kellogg, C. E., & Morton, N. W. (1999). *Beta III manual*. San Antonio: The Psychological Corporation.

McDaniel, T. M., & Rutström, E. E. (2001). Decision making costs and problem solving performance. *Experimental Economics*, *4*, 145–161.

Millner, E., & Pratt, M. (1992). A test of risk inducement: is inducement of risk-neutrality neutral? Virginia Commonwealth University Department of Economics working paper.

Nagel, R. (1995). Unraveling in guessing games: an experimental study. *American Economic Review*, *85*(5), 1313–1326.

Nisbett, R. E., & Ross, L. D. (1980). *Human inference: strategies and shortcomings of social judgment*. Prentice-Hall: Englewood Cliffs.

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

Porteus, S. D. (1965). *Porteus maze tests: fifty years' application*. Palo Alto: Pacific Book Publishers.

Raven, J., Raven, J. C., & Court, J. H. (1998). *Manual for Raven's progressive matrices and vocabulary scales*. San Antonio: The Psychological Corporation.

Read, D., Loewenstein, G., & Rabin, M. (1999). Choice bracketing. *Journal of Risk and Uncertainty*, *19*, 171–197.

Russo, J. E. (1978). Comments on behavioral and economic approaches to studying market behavior. In A. A. Mitchell (Ed.), *The effect of information on consumer and market behavior* (pp. 65–74). Chicago: American Marketing Association.

Rydval, O., & Ortmann, A. (2004). How financial incentives and cognitive abilities affect task performance in laboratory settings: an illustration. *Economics Letters*, *85*(3), 315–320.

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology*, *15*, 201–293.

Selten, R., Sadrieh, A., & Abbink, K. (1999). Money does not induce risk neutral behavior, but binary lotteries do even worse. *Theory and Decision*, *46*, 211–249.

Shah, P., & Miyake, A. (1996). The separability of working memory resources for spatial thinking and language processing: an individual differences approach. *Journal of Experimental Psychology: General*, *125*, 4–27.

Shah, P., & Miyake, A. (1999). Models of working memory: an introduction. In A. Miyake & P. Shah (Eds.), *Models of working memory: mechanisms of active maintenance and executive control* (pp. 1–26). Cambridge: Cambridge University Press.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *American Economic Review*, *72*, 923–955.

Stahl, D., & Wilson, P. (1995). On player's models of other players: theory and experimental evidence. *Games and Economic Behavior*, *7*, 218–254.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: implications for the rationality debate? *Behavioral and Brain Sciences*, *23*, 645–665.

Thaler, R. H. (1994). Psychology and savings policies. *American Economic Review*, *84*, 186–192.

Tuckman, B. W. (1991). The development and concurrent validity of the procrastination scale. *Educational and Psychological Measurement*, *51*, 473–480.

Turley-Ames, K. J., & Whitfield, M. M. (2003). Strategy training and working memory task performance. *Journal of Memory and Language*, *49*, 446–468.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task-dependent? *Journal of Memory and Language*, *28*, 127–154.

Wilcox, N. (1993a). Lottery choice: incentives, complexity and decision time. *Economic Journal*, *103*, 1397–1417.

Wilcox, N. (1993b). On a lottery pricing anomaly: time tells the tale. *Journal of Risk and Uncertainty*, *7*, 311–324.

Whiteside, S. P., & Lynam, D. R. (2001). The five factor model and impulsivity: using a structural model of personality to understand impulsivity. *Personality and Individual Differences*, *30*, 669–689.

Winkler, R. L., & Murphy, A. M. (1973). Experiments in the laboratory and the real world. *Organizational Behavior and Human Performance*, *20*, 252–270.