Report on *The Impact of Financial Incentives on Task Performance: The Role of Cognitive Abilities and Intrinsic Motivation*, a Dissertation submitted by Ondrej Rydval

By Glenn Harrison

This thesis is, on balance, a valuable contribution to the bridge between psychology and economics. It examines the role that cognitive abilities play in affecting performance in a well-defined task, and compares the relative explanatory power of those abilities and financial incentives. I do not believe the thesis has much to say about intrinsic motivation, for reasons explained below, but that is just a problem with the title. What is accomplished in chapter 3 is very impressive. The overall goal is to move economists beyond the "straw man" of a labor theory of cognition, which is a contrived alternative that nonetheless deserves to be addressed and quickly demolished. That theory is said to claim that if one pays subjects then they will exert more effort and one will observe a monotonic improvement in performance. I believe this idea is firmly in the heads of many economists as a local hypothesis, but I doubt any would entertain it globally. I will return to that issue as I go through each chapter. In any event, even if I disagree somewhat with the overall claims made in the thesis, the terrain is well worth traveling.

At an overall level, the thesis has been *very* poorly structured. Fortunately, I add quickly, this does not affect the final, "bottom line" assessment of the substantive contribution. But it is so striking that it is worth noting. The first two chapters are very loose in comparison to the third chapter. They are short pieces in which some interesting hypothesis is rather casually tested with data that seem unusually ill-suited to the task, and the conclusions stated as if they are "proof" rather than just a useful working out of how one *might* make the hypothesis about latent concepts and processes operational. I will have more comments about these chapters below, but it is a pity that they have equal chapter-status to chapter 3, since that is an original and careful work that would constitute a valuable thesis by itself. I do not know if the thesis was supposed to be structured as "three loosely connected essays and a staple" instead of "three essays on one theme," and I know the value of the former. But sometimes the latter, old-fashioned style, is the best, and this is one of them. Chapter 3 has far too many "fat footnotes" that should have been in the text, and by a fat footnote I mean page-length footnotes if they were the size of the main text. Some of these develop ideas that are quite central to the contribution of the chapter, and in exactly the length that one would expect in a thesis. But it was difficult to read because one was always having to read them to see that detailed matters had been taken care of, and that is not something one normally should have to do when reading a thesis or any academic paper. I encourage the supervisors to think more about how theses are structured in the future, since this is not a model that should be encouraged.

My overall assessment is that the thesis is a valuable contribution. If the chapters were equally weighted in that assessment, it would be different, but they are so dis-similar that it is impossible to do that and be fair. My detailed comments below should make my point clear here.

Chapter 1 examines the data from some experiments by Gneezy & Rustichini *QJE* 2000 (GR). In those experiments 120 Israeli subjects were given a number of questions that test some aspects of general intelligence: we are not told which ones, and there are many in the literature. The between-subjects design varies the reward to a correct answer. In one treatment subjects were paid nothing, in another they were paid 0.1 local currency units per correct answer, in another they were paid 1, and in another they were paid 4. This translates into nothing, USD 3 cents, USD 30 cents and USD 1, more or less. The contribution here is to look, literally with eyeballs, at the individual level data instead of the average data on performance, and to interpret the test in terms of general cognitive capital.

The key step here is to *assume* that the within-treatment variation in performance is all due to general cognitive ability differences in the population, and that between-treatment variation in performance is all due to effort induced by increased incentives. Since the former has a big range, and the latter has a smaller range, it is argued that cognitive abilities are just as important as financial rewards.

There is no attempt at a theoretical model, and one does not have to come up with much theory to see how difficult inferences are from this experiment. Use the capital-labor metaphor of a cognitive production function. But there might be capital inputs that are fixed with respect to output, and some that are variable: hardly a radical notion. So how do we disentangle the effects of the variable capital inputs from the effects of labor (assumed to be variable)? Maybe the higher financial rewards kicked in the fixed capital inputs *for the task as a whole*, but not the variable capital inputs or labor for each item. Or label these differently. If effort consists of "attention to the task" and "sweating over whether a specific answer is right," one can imagine the former as a fixed labor input and the latter as a variable labor input. Both are effort. The thesis seems to assume that labor is always variable, and that is far from obvious without a theoretical framework. I do not accept that "it is all in Camerer and Hogarth" since they are equally casual on this matter.

A number of concerns, some more serious than others:

- It is not at all obvious that these questions add up to what is called "IQ" in the literature. Quite apart from the vast literature on what the term IQ is supposed to measure, most

serious scholars in the area seem to have a more specific set of things that are measured. Knowing GR, I suspect that these questions had almost no serious connection to the formal measures of IQ that are constructed. So I think one should just call this what it is, and not IQ. This is also of some importance because this test is constantly referred to as "general cognitive capital" in the thesis, and it is not clear that it is indeed that.

- There seems to be little discussion of any related literature. Is there really nothing out there? What about McDaniel and Rutström *EE* 2001, cited in chapter 3? The paper seems very "rushed" when it comes to discussing the literature, particularly in comparison to chapter 3 as noted below, which is much more careful.
- The statistical analysis is non-existent. Maybe the differences are obvious to the eyeball of the author, but that is not at all obvious to this reader. The differences seem to be due to a very few outliers at the bottom end, which makes one wonder about the statistical power of some of the inferences. This is particularly true about the between-subject aspects of the design, since one might just have sample variability in terms of observable characteristics.
- The treatment with an "insulting" amount of money per correct answer is viewed as the same task as the others but just with a cardinal variation in the rewards. This strikes me as problematic. Recall the joke about Bernard Shaw asking some lady if she would sleep with him for a million pounds, and her blushing, smiling reaction that she might consider it. Then he asked if she would instead sleep with him for a farthing, and she explained, "sir, what do you take me for?" His reply, "madam, we just established that, now we are negotiating the price." Just so with this task. The very mention of an insulting reward could, arguably, change the task itself in the mind of the subject. This is a slippery slope to go down in formal terms, but it is enough to make one wonder why one would invest serious analytical energy on such experiments.

In summary, my main concern here is with the manner in which this casual review of a casual experiment is presented as more than it is. This is a useful comment on the GR experiment, to set up and perhaps motivate the hypotheses of interest, but not much more than that.

Chapter 2 is a review of some experiments by some behavioral accountants, who examined behavior in an accounting task and collected some variables proxying cognitive capital that might be specific to this task. The task itself is not explained, and has something to do with "accounting controls over the purchasing cycle" (p.12, line -4), whatever they are. The hypothesis being examined, in the overall thesis, is the relative role of *task-specific* cognitive capital and effort that is induced by financial rewards.

Task-specific cognitive capital is measure by two variables, the number of accounting

classes taken and auditing experience. Both are measured as integers: the number of courses, and the number of years experience. They are then collapsed into binary measures of "high" or "low" cognitive capital subjects, in an odd manner. The definition is that high capital subjects have either above-median courses *or* above-median experience. Why throw out the cardinal information in this manner? The particular stratification adopted seems arbitrary, particularly when the raw data is available and allows one to look at quantitative interactions and higher-order non-linearities. (There is a mention on p.16 (line -6) of an alternative stratification, but no results are presented, which is frustrating).

The statistical tests are suspicious, in the sense that they seem to be only partial. The pooled standard deviation of the variable Recall, the only performance measure that is incentivized, is 5.46 (Table 1). So differences of 2 and 2.5 between the treatments would not be statistically significant using old-fashioned *t*-tests. But in some cases they are significant using certain semi-parametric tests. I'm prepared to believe that there *might* be some difference, but this deserves more care, particularly if the stratification into "high capital" and "low capital" subjects (Table 2) is based on the arbitrary binary classification noted above. And there is the issue of small samples after the stratification, making power an issue as well. Parametric tests can often compensate for small samples, albeit at an obvious cost (there is no free inferential lunch).

My general concern here is that a proper conditioning is needed in the statistical analysis, to account for the key cognitive capital variables as well as any other sample composition differences. I was "tolerating" the unconditional analysis of Table 2 based on one arbitrary stratification, as a warm-up to the proper statistical analysis .... but it did not come.

Some minor points, some quite minor:

- The title for section 1 is mis-spelled.
- Note that only RETR was incentivized, and give the other performance measures less attention in discussing Table 1 (on page 12). Some other experiments, such as McDaniel and Rutström, did incentivize the time subjects expended on a task, and this could be an important control here since it is presumably a substitute input in the cognitive production function as they note in their design.
- Footnote 7 makes no sense – what does it mean to say one subject was "unidentifiable"?
- I get suspicious when medians are reported in Table 2 and not standard deviations, as noted earlier, since these are small samples and there could be a lot of noise.

In summary, this analysis seems incomplete. I'm not all that bothered by this, since the task is not clear and it is far from obvious that one has controls for task-specific cognitive capital (and that

the controls one does have are properly used).

Chapter 3 was a delight. I found myself working through the footnotes, and having all my questions answered. The tone of scholarship here seems to be orders of magnitude different and deeper than chapters 1 and 2. Indeed, I am mentally framing this chapter as the thesis, and not just one chapter.

One particularly nice feature of the chapter is the wide battery of "professional strength" cognitive tests presented to subjects. Even more discussion of these would have been useful, as well as the logistical and cost issues in their use. This is something that experimental economists, even behaviorally-inclined ones, know little about.

The literature review provided here (starting on page 24) should have been provided much earlier, and puts to shame the casual citations of chapters 1 and 2. I am not sure I would agree with the sweeping conclusion that the literature cited on page 24 suggests that individual heterogeneity in cognitive capital can explain departures from rationality as claimed (p.24, l. -7), but the review is intriguing.

This chapter also has some "pregnant suggestions" for future research, which is also what I like to see in a deeper work of scholarship. One fine example is footnote 17, proposing that one could allow subjects to buy more cognitive capital, just like some TV game shows allow contestants to "buy a vowel" when guessing some partially displayed word.

The instructions in the appendix are very well constructed, and I reviewed them in detail since there was no information at all provided about the experiments in chapters 1 and 2 (and my motivation for tracking them down was not high, I have to admit). One could easily have had a whole chapter on these and the factors that went into their design, material buried in long footnotes now.

The discussion of why subjects were allowed to bet on their forecasts as well as get rewarded on their accuracy (footnote 36) deserves more explication. This is related to some issues that have arisen in behavioral game theory when subjects are incentivized to state their beliefs about what they expect other players to do. An expanding literature there shows that such treatments can influence how people behave in games, so that there might be some interaction.

One disappointing feature of the design is that the risk aversion measures were hypothetical (page 42). And this fact was buried in a footnote! The literature is very clear about the differences

between hypothetical and non-hypothetical risk attitudes, not least the Holt and Laury *AER* 2002 paper cited for the design of the test. Maybe they are correlated, maybe, but this was a major gap in the design since there are considerable uncertainties in the rewards subjects were incentivized by. And there was clearly a decent budget here, as the discussion of windfall gains (p.42) makes clear.

One particularly puzzling section explained why this design had nothing to do with intrinsic motivation (page 39, in the paragraph starting "Not directly"). Since this is in the thesis title, this could have been explained much more thoroughly, and perhaps with the aid of a theoretical framework, such as in Benabou & Tirole *REStudies* 2003.

The statistical analysis is much more in line with what one expects. A Tobit model is estimated in which there are some right-hand side controls. (This allows me to just skip over some mind-numbing tabulations of correlations shown in Tables 2a and 2b). What about controls for sample differences in terms of observable characteristics, such as sex, age, educational level within the university, or income? The discussion of the need for the Tobit censoring model (footnote 71) is good, but makes the statistical mistake of discussing how many of the subjects were *actually* censored. The issue is rather how many *would* have their *error* distribution censored, even if the expected response or observed response is not censored. That is, a subject with an observed response ϵ away from some boundary would be censored if their error was allowed to be any amount larger than ϵ.

One minor irritant: there are many sloppy citations. It is unprofessional to just cite a paper as "mimeo." and this is done in many instances. Cite the location of one of the authors, if there is no working paper to cite. This is just laziness, and such things should not occur in a doctoral dissertation.

I also like the fact that this chapter seems qualified in the right scholarly manner. None of the sweeping generalizations in conclusion, as in chapters 1 and 2. Part of the reason is that the literature seems to be analyzed more fundamentally. In several places there is a clear recognition that the hypothesis tests are latent to the observed behavior (e.g., page 58, lines 2-6). This helps one draw the right, valid conclusions, and not over-state things. I get the feeling that this chapter was written well after chapters 1 and 2, when many of the conceptual issues had been thought through more clearly, and no attempt was made to go back and extend the earlier chapters.