CSAE WPS/2010-11

Dictator games in the lab and in nature: External validity tested and investigated in Ugandan primary schools

by

Abigail Barr[†] and Andrew Zeitlin[†]

05 May 2010

Abstract: This paper tests the external validity of a simple Dictator Game as a laboratory analogue for a naturally occurring policy-relevant decision-making context. In Uganda, where teacher absenteeism is a problem, primary school teachers' allocations to parents in a Dictator Game are positively but weakly correlated with their time allocations to teaching and, so, negatively correlated with their absenteeism. Guided by a simple theoretical model, we find that the correlation can be improved by allowing for (a) variations in behavioural reference points across teachers and schools and (b) the positive effect of some School Management Committees on teacher attendance.

JEL Classications: C91, D64, I29, O15, O17. **Keywords:** Public service, Education, Experiments, Africa, External validity, Methodology.

The authors would like to thank Dr. Frederick Mugisha, Lawrence Bategeka, Madina Guloba, Nicholas Kilimani, Winnie Nabiddo, and the enumeration team from the Uganda Bureau of Statistics, seminar and conference participants at the CSAE annual conference 2010, University of Oxford, The Thurgau Experimental Economics Meeting 2010, the University of Nottingham, and the Centre for Experimental Social Sciences, Nuffield College, Oxford.

This document is an output from research funding by the UK Department for International Development (DFID) as part of the iiG, a research programme to study how to improve institutions for pro-poor growth in Africa and South-Asia. The views expressed are not necessarily those of DFID.

† CSAE and Department of Economics, University of Oxford.

Dictator games in the lab and in nature: External validity tested and investigated in Ugandan primary schools

1. Introduction

In experimental economics, concerns about external validity are often played down or dismissed as, owing to the control they afford the researcher over factors that are and are not salient to a hypothesis, laboratory experiments provide the most rigorous testing ground for theory (e.g., Plott, 1982; Falk & Heckman, 2009; Crosen & Gächter, 2010). This is a strong argument and is probably why laboratory-based experimentation is on the increase in the social sciences in general and in economics in particular. However, that a theory is supported in the laboratory does not necessarily imply that it tells us something about the world outside the laboratory (Schram, 2009; Bardsley, 2010, chapter 5). This may be of no concern to pure theorists, but applied theorists, especially those is interested in informing policy, ought to be interested in whether a theory is applicable to the outside world and, hence, in the external validity of the experiments used in its testing.

Further, not all experiments are designed to test formal theory. Some are designed to produce stylized facts, or exhibits, about the way humans behave that may explain certain observed phenomena and could usefully become the foci of future theories (Sugden, 2009). Others are designed to explore policy-relevant conjectures that, even though they are not founded on formal theory, are worthy of careful investigation. In these endeavours external validity is considerably more important (Schram, 2009). The argument is straightforward, if an experiment is going to be informative about the way the world works it must simulate the workings of the world. Internal validity remains important (Bardsley et al., 2010). However, if it is achieved only at considerable expense in terms of external validity the overall value of the endeavour becomes questionable.

The factors that compromise the external validity of economic experiments have been characterized in a variety of ways (see Bardsley et al. (2010) for a detailed discussion). In essence, three types of factor are a cause for concern: omissions, contaminations, and the artificiality of alteration. Omissions are factors in the naturally occurring decision-making context or system that are excluded from the experiment. That factors can be excluded or held constant in experiments when, in nature, they are present and vary is, of course, the great strength of experiments. However, especially where factors that are omitted or held-constant in the lab may, in nature, interact with the factors under investigation in the lab, it is reasonable to question external validity (Guala, 2002; Bardsley et al., 2010). The experimental response to this is to run further experiments designed to explore the interactions (Starmer, 1999), an approach referred to by Guala (2002) as "checking by exhaustion". The problem of contamination refers to the factors that are present in the lab but not in nature. These include "Hawthorne effects", "demand effects" and "experimenter effects" in general (Adair, 1984). Without deception, these are difficult to address experimentally. Finally, the artificiality of alteration critique argues that the relational spaces in the lab and in nature are fundamentally different, so much so that, while an experiment could in principle be described as an analogue for the decision-making context it is designed to elucidate, it will always be artificial (Greenwood, 1982).[1] This is a conceptual rather than an empirical critique and can only be addressed by demonstrating that an experiment is indeed a good analogue for the natural decision-making context or system. Guala (2001, 2002) proposes what might be referred to as an omnibus test for external validity that addresses all of these concerns. It simply involves checking to see whether data generated in the laboratory correlates in expected ways with data from nature.

Owing to experimental social scientists becoming more interested in engaging non-student subjects and researchers previously involved in survey-based empirical work starting to use lab-type experiments to generate proxy measures for preferences and behavioural tendencies, the literature now contains a number of correlations fitting Guala's criterion. Focusing on other regarding preferences, correlations worthy of note include the following: Carpenter and Myers (2007) found that Volunteer Fire Department members' behaviour in a donation experiment correlated with the time they dedicated to training with co-volunteers; Benz and Meier (2008) found correlations between charitable giving by individuals and their behaviour in donation experiments; Karlan (2005) found that Peruvians who were more trustworthy in a laboratory experiment were more likely to repay loans; Baran, Sapienza, Zingales (2009) found a correlation between MBA students' reciprocating behaviour in a laboratory experiment and their responsiveness to requests for donations by their business schools at the end of their programme of study; de Oliveira et al (2008) found that Texans' behaviour in a laboratory public good experiment correlated with their involvement in volunteer work; and Fehr and Leibbrandt (2008) found that fishermen who exhibit a higher propensity for cooperation in a laboratory public good experiment are less likely to overfish in their everyday lives.

This paper contributes to this literature in two ways. First, it tests by correlation the external validity of one of the workhorses of experimental economics, namely the Dictator Game (DG), as a laboratory analogue for a specific, naturally occurring, and policy-relevant decision-making context.[2] Second and more significantly, it follows the prescription of Levitt and List (2007), by developing a theoretical model that explains why the correlation is unlikely to be perfect and guides an analysis into how the correlation changes as and when factors excluded from the lab vary in nature. This analysis generates a number of insights into the role that the implemented DG specifically and laboratory experiments in general can play in policy-relevant research.

The remainder of the paper is arranged as follows. Section 2 introduces the policy issue. Section 3 explains why the DG was selected as the laboratory analogue for the status quo in nature and sets out the rest of the research design. Section 4 presents the simple theoretical model that is used as a framework for the analysis and interpretation of the data. Section 5 presents the data and the results of the analysis aimed, first, at testing by correlation the external validity of the

[1] Used here, relational space refers not only to who is relating to whom but also how people relate to other phenomena and to how those relationships are perceived (Bardsley et al., 2010).

[2] The Dictator game was designed by Kahneman, Knetsch, and Thaler (1986) and, first, conducted in a laboratory in incentivized form by Forsythe, Harowitz, Savin, and Sefton (1994). Not strictly a game, it has nevertheless proved very useful as a benchmark against which to compare behaviour in the Ultimatum Game, trust games, third party punishment games, and others. Further, in empirical development economics and anthropology, it is now being used to generate proxy measures for altruism to be used alongside survey data, either as a control or as an explanatory variable of principle interest.

experiment and, second, at investigating several factors that may be perturbing the correlation. Finally, Section 6 discusses the results and concludes.

2. Public Servant Accountability in Developing Countries

Throughout the developing world governments are failing in the provision of public services. Poorly functioning formal institutions are most often cited as the cause of the failure, with corruption in the form of bribery and expropriation, low public servant morale, and high absenteeism being seen as symptoms of the general malady. Institutional reform is presented as the solution, often with specific attention being paid to the role of civil society in holding both politicians and public servants to account. This area of investigation appears well suited to the application of lab-type experimental methods – it is all about interpersonal interactions, the rules of the game, and the extent to which those rules are enforced. A few lab and lab-type experiments addressing these issues have been conducted (Alatas et al., 2009; Barr et al., 2009; Barr and Serra, 2009; Cameron et al, 2010). However, while these experiments are often complemented by interesting and thought provoking correlations between behaviour in the lab and survey data, the correlations rarely if ever constitute convincing evidence of external validity.

In this paper we focus on the public provision of primary school education in Uganda. In a recent study, Chaudhury et al. (2006) conducted unannounced spot checks and found 27 percent of Ugandan primary school teachers absent from their places of work. Ugandan local governments tend to be severely under-resourced and, especially in the case of remote primary schools, ill-equipped to monitor and discipline poorly performing teachers and school managers. Even the most extreme instances of teacher absenteeism tend to be punished only by transfer to an even more remote school (to which they may never turn up, while continuing to draw a salary).[3] Parent-teacher associations (PTAs) have become less effective since the advent of Universal Primary Education in 1997, at which point the government took away PTAs' fund-raising ability (see Miguel and Gugerty, 2005, for related evidence from Kenya). School Management Committees (SMCs), officially sanctioned governing bodies comprising parents, teachers, and local government representatives, lack formal powers to hire, fire, or fine, and, since they cannot require parents to contribute and have no other means by which to raise money, they cannot incentivize teachers with bonus schemes.

The lab-type experiment presented in this paper is part of a larger investigation into the potential for local communities to hold publicly funded Ugandan primary schools and teachers to account. The latter involves a randomized policy intervention aimed at improving the capacity of SMCs. The lab-type experiment was conducted at the time of the baseline survey. Its inclusion in the research design was motivated by Chaudhury et al.'s (2006: 93) comment that "[g]iven the rarity of disciplinary action for repeated absence, the mystery for economists may not be why absence from work is so high, but why anyone shows up at all." Intrinsic motivations may be the explanation. However, if this is the case, the literature indicating that, in the presence of intrinsic motivations, interventions designed to strengthen accountability may have unintended consequences (e.g., Frey and Oberholzer-Gee, 1997; Gneezy and Rustichin, 2000), needs to be

[3] See Banerjee and Duflo (2006) for a discussion of absenteeism problems in developing countries more generally and for a review of evidence on policy interventions intended to mitigate this problem.

given careful consideration. The function of the lab-type experiment was to (a) provide a measure of teachers' intrinsic motivations that could be used to address Chaudhury et al.'s comment and test a number hypotheses concerning the interaction between intrinsic motivations and external incentives; and (b) act as a baseline in a series of lab-type experiments designed to investigate what would happen if SMCs were empowered to hold teachers to account.

3. Research Design

The DG is one of the simplest instruments in the experimental economist's tool kit; while the motivation for giving in the DG remains a topic for debate (List, 2007; Bardsley, 2008), the DG design precludes, to the extent possible, giving in the hope of reciprocation and giving to avoid punishment. Thus, it is well suited to the function of generating a measure of teachers' intrinsic or internal motivations.

Further, while it would be foolish to propose that the motivations for giving in a DG are a perfect match for the motivations affecting teachers' time allocations, the DG compares well to a reasonable albeit stylized characterization of the context in which Ugandan primary school teachers currently make their time allocation decisions. Thus, it may also be well suited to function as a baseline in a series of experiments exploring the effects of empowering SMCs. The stylized characterization takes the following form: Ugandan teachers sell a contracted amount of time to the government each month; the government gives them back this time and sends them off to schools in remote and isolated communities charged with the duty of allocating the time to teaching the children of those communities how to read, write, add, subtract, and so on; once in their posts, the teachers decide how much of their contracted time to allocate to the communities and how much to allocate to themselves and they make these decisions in the absence of any effective formal monitoring or enforcement. This is a DG in which the teachers are the dictators, the communities are the recipients, the currency is the teachers' time, and the size of the stake is specified in the contract.

The next step was to design an experiment involving the DG that: to maximize external validity, closely replicated the characterization above (Loewenstein, 1999); and, to maximize the usefulness of the measure of teachers' levels of motivation, preserved the DG in a recognizable form. To this end, while remaining mindful of the conceptual nature of the artificiality of alteration critique, we first set out to replicate, to the extent possible, the relational space observed in nature in the lab. The DG would be conducted with teachers in the role of dictator and parents of pupils in the role of passive recipient. Headmasters and SMC members would also be present, but with no role to play in the DG. Having field researchers present to run the DG was unavoidable, although even here steps were taken to minimize the likely impact on the relational space. The field researchers were predominantly young employees of the Uganda Bureau of Statistics, from the same regions as the schools in which they worked, and experienced in making respondents feel comfortable in their presence through prior work on household surveys investigating sensitive topics such as health outcomes. Further, the paperwork they carried and displayed and the identifying badges they wore were designed to indicate that, while they had the permission of the Minister for Education to conduct the research, they were not in the employ of and should not be perceived as monitors sent by the Ministry. In addition, we aimed to hold the teachers' physical decision-making contexts constant across nature and the lab by arranging for a make-shift lab to be set up in each and every school included in the study and ensuring that the sample design supported one experimental session in each school.

We were under no illusion that all the features of the teachers' time allocation decision-making environment would be replicated in the DG described above. At least five differences remained. First, in the natural decision-making context the teachers may coordinate their absences, taking advantage of the substitutability of their time in the production of education services and, thereby minimizing the impact on their pupils. Alternatively, teachers may be called upon to cover for the absences of their colleagues, so that absenteeism exerts a negative externality. In the DG individual decisions are made in isolation; no coordination is allowed and negative externalities are present only to the extent that selfish behaviour by one teacher may cause reputational harm to the other teachers in the same experimental session. Second, in nature, the teachers' time allocation decisions are observable to anyone in the community, including the SMC, who might choose to pay attention, whereas in the DG the teachers' allocation decisions are neither observed nor observable, although, given the size of the sessions, the resulting anonymity is probably best described as limited. Third, the DG is one-shot, whereas the time allocation decisions are repeated. Fourth, the teachers are contracted to allocate a certain amount of time to teaching and even though the contracts are not enforced they may have an effect that is not replicated in the lab. And fifth, DGs are played with Ugandan Shillings, whereas the time allocation decisions involved time not money. However, these differences notwithstanding, we considered the case for selecting the DG to be sufficiently strong to warrant putting it to the test.[4]

The subject sampling approach was as follows. Four out of Uganda's eighty administrative districts were purposefully selected, one from each region, and each broadly representative of the educational challenges faced in that region. In each of the selected districts 25 rural, state primary schools were randomly selected from the complete listings. Then, either five teachers (excluding the head teacher) were randomly selected from the complete list of teachers working at each school or, if five or fewer teachers were working at the school, all were selected. For reasons relating to the broader study, this sample of teachers was stratified to place greater weight on teachers of maths and literacy to primary 3 and primary 6 classes. At the same time, in each school, three pupils from the third year and two pupils from the sixth year were randomly selected from the full pupil lists and either their mother or their father (randomly determined) was selected for inclusion in the study. Finally, from each school's SMC we selected the head master, the chairperson, one foundation-body representative (randomly selected from all such members), and two parents (randomly selected from all parent members). The fifteen individuals selected in relation to each school and corresponding community were surveyed with different survey instruments being applied to the four types of individual and one further survey instrument was applied to the school as a whole. In approximately half the schools the surveys

[4] A number of other games were considered. These included the gift-exchange game (Fehr, Kirchsteiger and Riedl, 1993), which was rejected on the grounds that the salary-setting employer, i.e., the Ministry of Education, was both remote and irrelevant given the focus on local accountability. A novel variant on the DG designed to capture the potential substitutability of different teachers' time in the production of educational services was also rejected on the grounds that it would not generate a readily understandable measure of teachers' motivations and we could find no evidence of teachers exploiting the substitutability of their time in the literature.

were conducted first, and in the other half the experiment was conducted first. All the individuals invited to participate in the experiment did, indeed, participate.[5]

Each experimental session was conducted using a classroom large enough to seat all fifteen subjects and the four field researchers running the session, and three decision-making stations, located outside of this classroom and at a sufficient distance to ensure complete privacy for one-on-one interviews. Each session proceeded as follows. On arrival, each subject was registered and given a badge bearing a number and either an orange, green or blue figure. Parents received badges bearing orange figures, teachers, badges bearing green figures, and SMC members, including the head master, badges bearing blue figures. Following registration, each subject was invited to sit in the area in the classroom assigned to their badge colour.

Once all the subjects had arrived and were seated, one of the field researchers, standing at the front of the classroom, proceeded to introduce the research team and go through a series of formalities relating to ethics and control.[6] Then, before describing the DG, the presenter invited all those with green badges to raise their hands and then asked "Am I right in thinking that you are all teachers?" Having received confirmation, she or he said "During the workshop I will refer to you as green players." The presenter then went through the same procedure for orange badge wearers, i.e., parents, and blue badge wearers, i.e., SMC members. The aim of this exercise was to ensure that badge colours and the roles that the badge wearers assumed in their everyday lives and in the DG were linked in the minds of the participants. This was the only framing applied to the experiment.

Finally, the DG was described using wall mounted visual aids depicting green and orange figures and moving representations of real Ugandan Shillings (oversized plastic-coated pictures of coins with Velcro tabs on the back) about to show allocations being made (see Figure 1).[7]

Once the subjects had been taught the game, the teachers were called to one-on-one meetings with field researchers, were taken through the game again, tested, and were then invited to express their chosen allocation by dividing the real-money stake between the green figure representing themselves and the orange figure representing the parent they had been anonymously paired with on the table in front of them.[8] The stake in the DG was 5,000

[5] This arrangement meant that, in each district, two teams, one specializing in the surveys, the other in the experiments could complete the enumeration of two schools in a day. Each team would work in one school in the morning and the other in the afternoon, swapping at lunchtime. No order effects are apparent in the data.

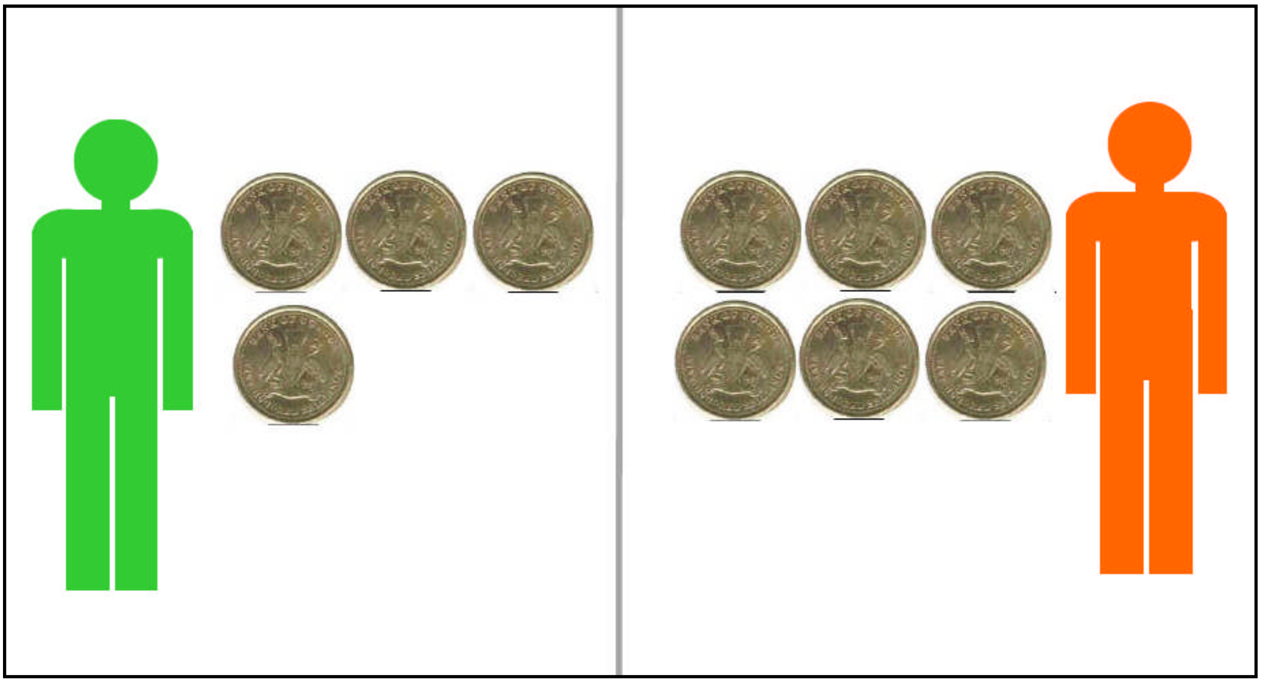
[6] On ethics: anticipating a greater than zero rate of illiteracy among the parents, in place of signed written informed consent, the session script included a statement early on indicating that anyone who felt uncomfortable about what they were being asked to do would be free to leave. On control: the subjects were asked not to talk to one another once the session had started and were informed that, if they talked, the session would have to be drawn to a close.

[7] Note that, while the SMC members (blue players) were not directly involved in the game, each was endowed with half of the stake that each teacher was given to allocate between him or herself and a parent. We did this to emphasize the passive presence of the SMC members during the DG and to ensure that the SMC members did not enter the second game, a third party punishment game in which they assumed the punishing role, feeling under-remunerated.

[8] We had no reason to doubt the literacy and numeracy skills of the teachers. So, we could have conducted the DG using written instructions, pencils and paper or possibly even computers. This would have economized on presentation time and, potentially, improved control. However, we knew that many of the parents and some of the SMC members would be illiterate and we wanted, first, the teachers to be in no doubt that the parents understood the game and, second, to apply the same protocol – group presentation with visual aids followed by one-on-one

Ugandan Shillings (just under $3) and was presented in the form of ten 500 Ugandan Shilling coins.[9]

Figure 1: Example of the visual aids used in the Dictator Game



Once all the teachers had made their decisions and returned to the classroom, a second game, a third party punishment game, was presented and played. This was followed by a modified third party punishment game. The subjects were informed about the outcomes of and remunerated for all three games at the end of the session.[10] No talking was allowed throughout. Sessions lasted between one- and-a-half and two hours.

Finally, in order to test the external validity of the DG by correlation, one needs an observational or survey measure that, under the assumption of external validity, will be correlated with the allocations made by the dictators in the DG. As a potential correlate to the DG allocations that the teachers made to the parents, we selected the proportion of contracted time that each teacher allocated to teaching during the month preceding the survey and experiment. During the surveys each teacher was asked how many days they were absent from the classroom in the month prior to the survey. In addition, the survey instrument applied to each school asked for each teacher's absences as they appeared in the school records. In the analysis, to minimize the effects of absenteeism being underreported (strategically on the part of teachers and resulting from poor record-keeping on the part of schools), the maximum of the two measures of absenteeism is used in conjunction with the assumption that teachers are contracted to work 20 days a month to

interviews – to all active subjects (teachers in the DG, teachers and SMC members in a subsequent third party punishment game).

[9] The 5,000 Shilling stake corresponded to half a day's pay for the average teacher in our sample.

[10] Ideally, we would have remunerated the subjects for only one game, randomly picked at the end of the session. However, we know from experience that this approach can be difficult to explain to illiterate and only partially numerate subjects and considered it very important that the teachers be in no doubt that the parents and SMC members completely understood what was going on.

calculate the proportion of contracted time that each teacher allocates to the community they have been sent to serve.

4. Theoretical Framework

This section presents a simple theoretical model that yields a testable prediction about how teachers' allocations in the DG and contracted-time allocations in nature are related under the assumption that a single preference parameter is driving behaviour in both contexts. The model also provides a framework within which variations in decision-determining factors between nature and the lab and across the sample in either nature or the lab or both can be characterized and investigated.

Consider a teacher *t* deciding how to divide a resource between him or herself and the community he or she has been sent to serve or a representative thereof (both referred to as "the community" below) and with the following utility function

$$U_{it}^{\kappa} = x_{it}^{\kappa} - w^{\kappa}\alpha_t \sum_{j=i,-i}(x_{jt}^{\kappa} - \gamma_j^{\kappa})^2 \qquad (1)$$

where x_{it}^{κ} is *t*'s allocation to him or herself in context κ, α_t is the parameter capturing teacher *t*'s preference to adhere to reference point allocations (RPAs), γ_j^{κ} is the RPA to *j* in context κ, w^{κ} is the weight assigned to this preference in context κ, *j* = *i* or *-i*, *i* indicating the teacher, *-i* indicating the community, and $\kappa = s$ or l, s indicating the allocation of the teacher's contracted time, and l indicating the allocation of money in the lab DG. Note that, for the time being, the weights and RPAs are assumed to be common across all teachers.

The utility function in (1) has elements in common with many in the behavioural economics literature (see for example Kőszegi and Rabin (2006) and Capellan et al (2007)). However, to our knowledge, the context-specific weights attached to the common-across-contexts preference parameter are novel. These weights correspond to variations in the salience of a given preference across contexts and, thus, could be interpreted as formally capturing one dimension of the artificiality of alteration. In the current case, the contract relating to teachers' time allocations may strengthen the salience of the preference in the natural as compared to the lab context. Alternatively, the fact that the lab-type DG is played with a windfall could strengthen the salience of a preference to share-and-share-alike in the lab as compared to the natural context.

Maximizing (1) subject to the normalizing constraints that $x_{it}^{\kappa} + x_{-it}^{\kappa} = 1$ and $\gamma_i^{\kappa} + \gamma_{-i}^{\kappa} = 1$ yields teacher *t*'s optimal allocation to the community in each of the contexts:

$$x_{-it}^{s*} = \gamma_{-i}^{s} - \frac{1}{2w^s\alpha_t} \qquad \text{and} \qquad x_{-it}^{l*} = \gamma_{-i}^{l} - \frac{1}{2w^l\alpha_t}$$

Rearranging each of these and combining on $\propto_t$ yields the following prediction.

Prediction 1: *A teacher's allocations to the community of contracted time in nature and money in the DG are linearly related; holding γ_{-i}^{s}, γ_{-i}^{l}, w^s, and w^l constant and varying α_t, this relationship is given by*

$$x_{-it}^{s*} = \beta_0 + \beta_1 x_{-it}^{l*} \quad \textit{where} \quad \beta_0 = \gamma_{-i}^{s} - \frac{w^l}{w^s}\gamma_{-i}^{l} \quad \textit{and} \quad \beta_1 = \frac{w^l}{w^s}. \qquad (2)$$

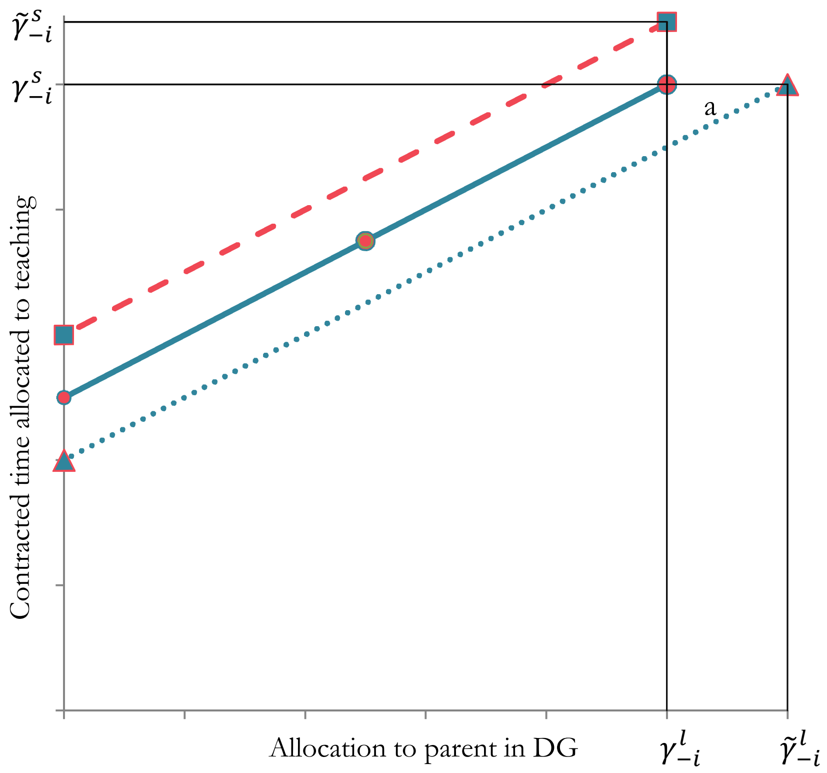


Figure 2: Predicted relationship between teachers' proportional allocations to parents in the Dictator Game and the proportion of contracted time allocated to teaching

The slope of the line deviates from 1 if the weights assigned to the preference to adhere to the RPA differ across contexts, being steeper the greater the weight in the lab-type DG and the smaller the weight in the contracted-time-allocation context. The intercept of the line deviates from zero if either the weights assigned to the preference to adhere to the RPA differ across contexts or the RPA differs across contexts or both (except in the case where $\gamma^s_{-i}/\gamma^l_{-i} = w^l/w^s$*).*

The solid line in Figure 2 plots the relationship. The higher a teacher's α_t *the further up the line towards point* ***a*** *the teacher will be. Point* ***a*** *lies at* $[\gamma^l_{-i}, \gamma^s_{-i}]$*. No teacher will be located above and to the right of point* ***a****.*

So, if RPAs and preference weights are (sufficiently) common across all teachers and the preference to adhere to a context-specific RPA is stable within teachers across contexts while varying across teachers, we will find a linear relationship between allocations in the two contexts.

This simple model can also be used to generate several predictions about how the relationship in (2) might be perturbed owing to variations in decision-determining factors between nature and the lab and across the sample in either nature or the lab or both. Here, we focus on two such predictions.

Prediction 2: *A variation in the cross-context difference in RPAs corresponds to a variation in the intercept of the relationship in (2).*

Consider, for example, two teachers who have the same RPA in the contracted-time-allocation context and different RPAs in the lab-type DG. They will be located on different, parallel versions of the relationship; one may have a DG RPA of γ_{-i}^{l} and, so, be on the solid line in Figure 2, while the other has the higher DG RPA of $\tilde{\gamma}_{-i}^{l}$ and is, therefore on the dotted line. Alternatively, suppose that the contracted-time-allocation RPA varies across schools. Then, one teacher may have a time-allocation RPA of γ_{-i}^{s} and, so, be on the solid line in Figure 2, while the other has a higher time-allocation RPA of $\tilde{\gamma}_{-i}^{s}$ and is, therefore on the dashed line. Thus, we see that if the cross-context difference in RPAs varies across teachers, the linear correlation between the allocations made in the two contexts is perturbed.

Now consider how the relationship would be perturbed if in some schools the SMC or the parents had found a way to hold the teachers to account informally. Suppose, for example, that an SMC credibly threatens to punish a teacher who allocates less than $\tilde{x}_{-i}^{s}$ to the community, $\tilde{x}_{-i}^{s} > \gamma_{-i}^{s} - w^{l}/w^{s}\, \gamma_{-i}^{l}$, the credibly threatened punishment is sufficiently severe that the teachers wish to avoid it, and there are some teachers for whom this added constraint binds because they have a sufficiently low α_t. Then, the relationship represented by the solid line in Figure 2 is transformed into the relationship represented by the solid line in Figure 3.

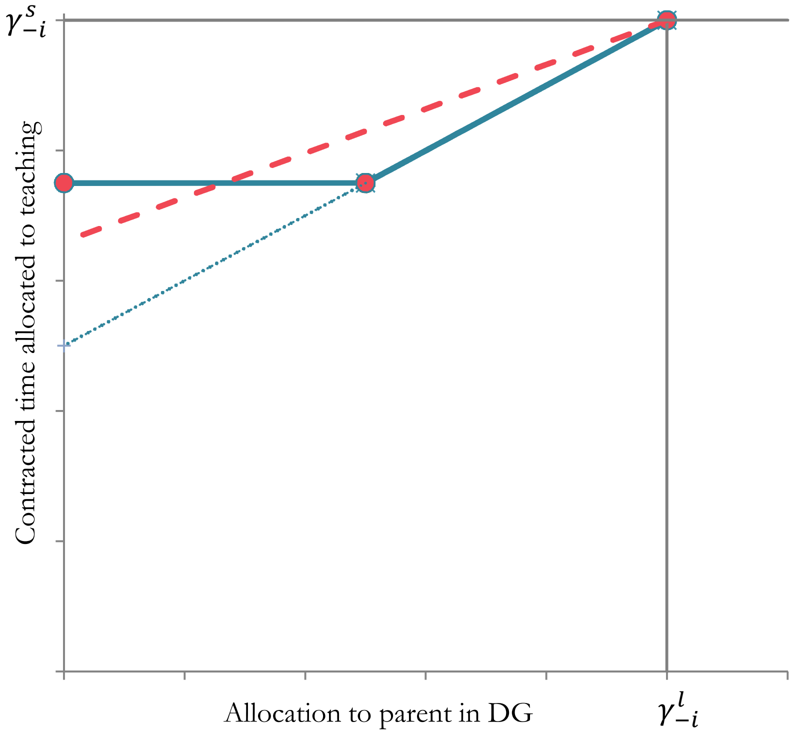


Figure 3: Predicted effects of informal enforcement on the relationship between teachers' allocations of contracted time and in the Dictator Game

Alternatively, the informal enforcement could be graduated with more pressure being brought to bear on a teacher the lower his or her α_t. Then, the relationship represented by the solid line in Figure 2 would be transformed into the relationship represented by the dashed line in Figure 3. This is graphically equivalent to a strengthening of the weight that a teacher applies to the preference to adhere to the reference point, w^{s}, although in this case the swivel is owing to an external force.

Either way, we have one final prediction.

Prediction 3: *In schools where the teachers are held to account by the parents or the SMC, a teacher who, in the absence of this holding to account, would choose a relatively low* x_{-it}^{l*} *and a correspondingly low* x_{-it}^{s*} *would now choose to allocate more of their contracted time to the community.*

So, if parents and SMCs in different schools have succeeded to varying degrees in finding informal ways to hold teachers to account, the linear correlation between the allocations made in the two contexts will, once again, be perturbed.

5. Data and Empirical Analysis

Figure 4 presents the distributions of the proportion of contracted-time that the teachers allocate to teaching (blue/dark) and the proportional allocations made by teachers to parents in the DG (pink/light). The difference in the distributions is immediately obvious. The DG allocations display a strong mode at 0.50 (0.45 was not an option in the DG), while time allocations display a strong mode at 0.85 to 0.90 (17 to 18 days in a 20 working day month), have a smaller variance than the DG allocations, and show signs of truncation at 1.00.[11] In the absence of a theory, the differences in the histograms may have led us to conclude that the DG is not a good match to the contracted-time allocation decisions being made by the teachers. However, the differences are consistent with the RPA and preference weight differing across the two contexts and do not preclude a correlation driven by within-teacher stability and cross-teacher variations in the preference parameter, α_t.

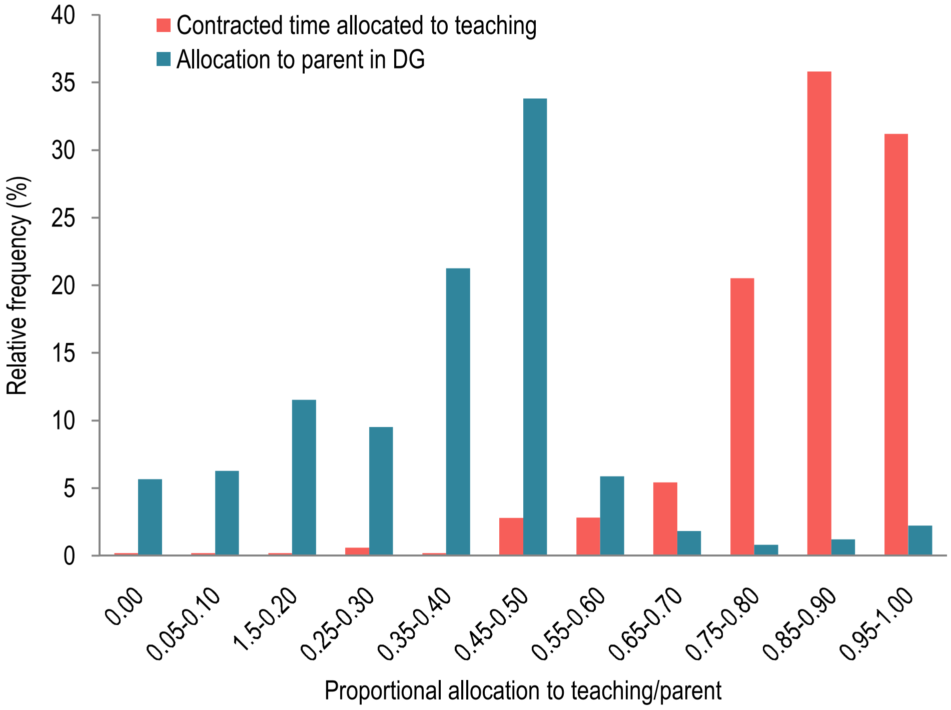


Figure 4: Distribution of contracted time allocations to teaching and allocations to parents in Dictator Game

[11] The average proportional allocations are 0.85 and 0.40 in the time and DG contexts respectively.

5.1 The external validity of the lab-type DG

Prediction 1 is explored in Table 1. The P-values corresponding to five tests are presented. Four of these test Prediction 1 directly as they relate to the existence of a monotonic or linear relationship between allocations of time to teaching and allocations to parents in the DG. Three of these, the Spearman's rho, the pairwise correlation, and the linear regression with no clustering of errors return significance levels better than 5 percent. One, the linear regression with the errors clustered to account for possible interdependence within schools (and, hence, experimental sessions) returns a significance level of 6.5 percent.[12]

Table 1: Statistical significance of relationship between proportion of contracted time allocated to teaching and proportional allocation to parent in Dictator Game (n=487)

P-value of Spearman's rho	0.013
P-value of pairwise correlation	0.042
P-value of linear bivariate regression#	0.042
P-value of linear bivariate regression, clustering by school#	0.063
P-value on F-test of joint sig. of teacher fixed effects##	0.039

Notes: # t-test relates to estimation of $x^{S*}_{-it} = b_0 + b_1 x^{l*}_{-it} + \eta_t$ where η_t is the error term; ## here, the discreteness in both variables was accounted, the proportion of time allocated to teaching was the limited dependent variable and ten dummies relating to each of the possible DG allocations were the only explanatory variables; ## F-test relates to estimation of $x^{\kappa*}_{-it} = a_0 S + f_t + \xi^{\kappa}_t$, where S is a dummy variable taking the value 1 in the contracted-time allocation context and zero in the DG, f_t is the vector of teacher fixed effects and ξ^{κ}_t is the error term.

The regression with the clustered errors is presented in full in the upper panel and first column of Table 2. As expected, the coefficient on the proportion allocated to the parent in the DG is positive. As well as being significantly different from zero, it is also significantly different from 1. This is consistent with the weight assigned to the preference to adhere to the reference point in the time allocation context being greater than the weight assigned to the same preference in the lab-type DG. Note also, the large, positive, and significant intercept.

Returning to Table 1, the fifth test relates to a fixed effects regression model of the form: $x^{\kappa*}_{-it} = a_0 S + \boldsymbol{f_t} + \xi^{\kappa}_t$, where S is a dummy variable taking the value 1 in the contracted-time allocation context and zero in the DG, $\boldsymbol{f_t}$ is the vector of teacher fixed effects and ξ^{κ}_t is the error term. The statistic of interest is the joint significance of the teacher fixed effects. Table 1 reports that the teacher fixed effects are jointly significant at the 4 percent level, which is consistent with some aspect of each teacher's decision-making process being common across the two contexts.[13]

[12] The worsened significance in the clustered regression is consistent with the allocations in both contexts being clustered at the school level. If the DG allocations are regressed on a full set of school dummy variables, an R^2 of 0.36 is returned. And if time allocations are regressed on a full set of school dummy variables, an R^2 of 0.37 is returned.

[13] Some readers my find this approach is more appealing than the regression of the time allocations on the DG allocations as it directly acknowledges the simultaneity of the two decisions or, put another way, the idea that both

The teacher fixed effects regression is presented in full in the lower panel and first column of Table 2. In this case, the coefficient on the contracted time context dummy gives us the mean difference between DG and contracted time allocations. The latter is significant and positive which is, once again, consistent with the reference point and the weight being applied in the time allocation context being higher than the reference point and weight being applied in the lab-type DG.[14]

These test results provide weak evidence of local external validity. However, they also reveal some significant differences between the two contexts. Further, even after accounting for these differences, the correlation between the allocations made in the lab and in nature is far from perfect and this generates an opportunity to investigate various hypotheses about differences in reference points across teachers and informal enforcement across schools.

Table 2: Regression analysis of relationship between Dictator Game allocations to parents and time allocations to teaching

Dependent variable = Proportion of contracted time allocated to teaching	**1a**		**2a**		**3a**		**4a**		**5a**	
Constant	0.823	***	0.836	***	1.001	***	1.206	***	1.212	***
	(0.017)		(0.014)		(0.042)		(0.042)		(0.045)	
Proportion allocated to parent in DG	0.066	*	0.067	**	0.067	**	0.072	**	0.166	***
	(0.035)		(0.030)		(0.030)		(0.031)		(0.053)	
ln(adj(wealth$_i$-wealth$_{-i}$))#					-0.008	***	-0.019	***	-0.019	***
					(0.002)		(0.002)		(0.002)	
SMC Meetings * Allocation to parent in DG									-0.046	**
									(0.022)	
School fixed effects included	no		no		no		yes		yes	
F-stat for school fixed effects							1.850	***	1.890	***
Observations	487		476		435		435		435	

Dependent variable = Proportion of contracted time allocated to teaching OR Proportion allocated to parent in DG	**1b**		**2b**		**3b**		**4b**		**5b**	
Constant	0.399	***	0.399	***	0.404	***			0.404	***
	(0.008)		(0.007)		(0.007)				(0.007)	
Contracted time context (*S*)	0.450	***	0.464	***	0.451	***			0.414	***
	(0.011)		(0.010)		(0.011)				(0.023)	
Contracted time context * ln(wealth$_i$)-ln(wealth$_{-i}$)					-0.010	**			-0.009	**
					(0.004)				(0.004)	
Contracted time context * SMC Meetings									0.018	*
									(0.010)	
F-stat for teacher fixed effects (***f_i***)	1.190	**	1.220	**	1.240	**			1.250	**
Observations	974		952		870				870	

Notes: # ln(adj(wealth$_i$-wealth$_{-i}$)) is the natural log of the teacher's wealth minus the wealth of the average parent in their experimental session minus the within sample minimum of this difference plus 1; *** significant at 1% level'; ** significant at 5% level; * significant at 10% level.

are driven by an underlying unobserved common factor. Others may be of the opinion that, as long as this simultaneity is born in mind and no causal inferences are drawn, the regressions of the time allocations on the DG allocations have the appeal of being more explicitly related to the theory, especially when its implications are presented graphically.

[14] Note that, in the fixed effects model we cannot disentangle the effect of a cross-context difference in the RPA from that of a cross-context difference in the weight applied to the preference to adhere to the RPA.

5.2 Variations in the cross-context difference in RPAs

In this sub-section we use Prediction 2 to guide our investigation. Note that we are not claiming to be testing the prediction; the findings below that are consistent with the prediction may also be consistent with other explanations. Rather, we are using the prediction to organize our thinking and guide our choice of variables to include in the analysis.

Prediction 2 states that, *ceteris paribus*, a variation in the cross-context difference in RPAs corresponds to a variation in the intercept of the relationship in (2). Here, we investigate three possible sources of variation in the cross-context difference in RPAs.

First, consider the possibility that some teachers were absent from school for genuine reasons. Suppose that a teacher contracted to work 20 days in a month is genuinely ill for five. Then, his or her time allocation reference point would shift down by 0.25 or a little less if he or she feels obliged to make up the for the time taken off. So, the relationship between the DG and time allocations would shift to the left for teachers who have been absent for genuine reasons and would shift more the longer their absence.

The baseline survey data includes some reasons for absence self-reported by the teachers. However, preliminary analyses suggested that these were subject to the sort of biases one might expect. For example, those who admitted being absent to pursue another income earning opportunity (an unacceptable reason for absence) were, in general, less absent and more generous in the DG. A less than ideal alternative approach is to explore the possibility that teachers who were absent a lot during the month preceding the survey and experiment, were absent for genuine reasons. Out of the 487 teachers for whom we have both time-allocation and DG data, only 11 (2.3 percent) were in school for less than 0.50 of their contracted time. On average, these teachers allocated 0.29 of their contracted time to teaching. Applying the estimated relationship presented in the first column of the upper panel of Table 2, this would correspond to a mean DG allocation to parents of zero (strictly less than zero if such an allocation was possible). However, on average, these 11 teachers allocated 0.41 of the stake to the parent in the DG. This is consistent with acceptable absences causing the relationship to shift to the left as described in the preceding paragraph.

Further, if these 11 teachers are dropped from the sample and the regression of contracted time allocations on allocations to parents in the DG re-estimated (see upper panel, column 2 of Table 2), the accuracy with which the slope coefficient is estimated and, hence, its significance are improved. The same improvement is also evident in the F-statistic on the teacher fixed effects when the fixed effects model is re-estimated on the restricted sample (see lower panel, column 2 of Table 2).

Second, recall that, while in standard DGs individuals are paired with others who are, best guess, like themselves, in Uganda, the teachers knew they were paired with one of the five parents present in the same session. This being the case, there is less reason *a priori* to expect a DG RPA of 0.50:0.50. Just as reasonable would be the conjecture that teachers positively condition their DG RPA on their own income or wealth relative to the expected or mean income or wealth of the five parents. All other things being equal, teachers who are richer relative to the parents in the session would, then, display a lower intercept when their time allocation is regressed on their

DG allocation and teachers who are poorer relative to the parents would display a higher intercept. And in the teacher fixed effects specification an interaction term between the contracted time context dummy and the wealth of the teacher relative to the mean parent would bear a negative sign.[15] Of course, if teachers also tend to allocate a greater amount of time to teaching when they are assigned to communities that are poorer relative to themselves, this effect may be cancelled out. However, survey-based analyses indicate that absenteeism tends to fall as community income rises even after controlling for school infrastructure (Chaudhury et al. 2006). In addition, it is important to bear in mind that the five parents selected from each school represent a very small sample of each school's parent population. This being the case, their wealth relative to each teacher's may not be a good measure of the parent population's wealth relative to each teacher's and so, while any effect of relative wealth on the teachers' DG offers should be clearly identified in the data, the identification of its effect on teachers' time allocations would be undermined by measurement error.

For 435 of the teachers, we have survey data on the teachers' and the parents' durable assets. Conjecturing that relative wealth differences should affect the reference point for DG allocations, but having no priors about functional form, we used this data to construct both the difference between the natural log of each teacher's wealth and the natural log of the wealth of the mean parent in the same session and the natural log of the difference between each teacher's wealth and the wealth of the mean parent in the same session.[16] Then, each of these variables was entered into both the regression of the time allocation on the DG allocation and the fixed effects regression. Somewhat frustratingly one of the variables is significant in one of the specifications and the other in the other specification (see column 3, Table 2). However, in all cases, the estimated coefficient takes the expected negative sign.

Third and finally with respect to Prediction 2, suppose that the contracted-time-allocation RPA varies across schools, possibly because different behavioural norms have emerged owing to frequent and ongoing interactions between teaching colleagues in isolation from other members of their profession. All other things being equal, this would lead to different school-specific intercepts when teachers' time allocations are regressed on their DG allocations, i.e., a vector of school-fixed effects would be jointly significant when added to the regression of time allocations on DG allocations. This is precisely what we see in column 3 of the upper panel of Table 2. Further, note the increase in the size of the coefficient on the DG allocation, indicating that the commonality between the time and DG allocations is better identified when school-level differences in RPAs are controlled for.[17]

Among the other possible explanations for the significance of the school fixed effects three are of particular interest. First, it may be the DG RPA rather than the contracted-time-allocation RPA that varies across schools. Second, it may be the weights, w^s and w^l, rather than the RPAs

[15] Note that the wealth of the teacher relative to the mean parent not interacted cannot be included in the fixed effects specification as it is perfectly correlated with the set of teacher fixed effects.

[16] The average teacher in our sample owned assets with a total combined value of 16,200,000 Ugandan Shillings. The average parent owned assets with a total combined value of 24,700,000 Ugandan Shillings. Prior to taking logs, the second difference was adjusted upwards by a fixed amount rendering the minimum value of the difference equal to 1. Thus, the teachers who were poor relative to the parents were not lost from the sample.

[17] School fixed-effects are implicitly accounted for in all of the teacher fixed effects specifications. For this reason no regression is reported in column 4 in the lower panel of Table 2.

that vary across schools. In principle, this could be investigated by interacting the school fixed effects with the DG allocations. However, this would be equivalent to assuming a common error structure, while estimating the relationship between the time and DG allocations for each school one at a time relying on, at most, five observations. This would be ill-advised. Third, it may be that the extent to which teachers are informally held to account varies across schools. This brings us to Prediction 3.

5.3 Variations in local accountability

Prediction 3 states that, in schools where teachers are held to account by parents or the SMC, a teacher who, in the absence of this holding to account, would choose a relatively low x_{-it}^{l*} and a correspondingly low x_{-it}^{s*} would now choose to allocate more of their contracted time to teaching.

While we have no data on parents informally holding teachers to account, the survey of SMC members provides a measure of SMC activity levels, that can be used to explore this prediction and, at the same time challenge the significance of the school fixed effects. Each of the five sampled SMC members in each school was asked how many SMC meetings had taken place during the six months preceding the survey.[18] Here, we take the median for each school as a proxy for the activity-level of the SMC.[19] In the regression of time allocations on DG allocations, we interact this proxy with the DG allocations with the aim of identifying a swivel effect. For Prediction 3 to be supported, the interaction term should bear a negative coefficient, i.e., the relationship between the two allocations should be flatter the more active the SMC.

The results of this exercise are presented in column 5 of the upper panel of Table 2. The interaction term bears a significant negative coefficient. Further, its inclusion in the model, causes the coefficient on the uninteracted DG allocation to more than double and to become a great deal more significant (P-value=0.002). A series of linear restriction tests indicate that where the SMC met three or more times during the six months preceding the survey (36 out of the 100 SMCs in the study did this), there is no discernable relationship between the teachers' time and DG allocations, whereas where they met one or two times or not at all there is a strong correlation.

In the fixed effects specification, we can interact the number of SMC meetings with the contracted time allocation dummy variable in order to identify the average effect of SMC activity on the differential performance of the teachers in the time relative to the DG allocation context. However, we cannot also interact this with the DG allocations as these make up part of the dependent variable sample. We expect the differential performance of the teachers in the time relative to the DG allocation context to be higher the more active the SMC and this is, indeed,

[18] The literature on teacher absenteeism in developing countries provides no evidence of a relationship between the level of activity of local entities that might be expected to hold teachers to account and teacher absenteeism (Chaudhury et al. 2006; Banerjee and Duflo 2006). However, in none of these studies are teachers' motivations accounted for.

[19] Across the sample of teachers, this measure of SMC activity varies between zero (4 percent) and five (2 percent), taking a mean value of 2.14.

what we find. The interaction between the median number of SMC meetings and the contracted time allocation dummy variable bears a significant positive coefficient.

6. Summary, Discussion and Conclusion

This paper set out to test by correlation the external validity of a simple one-shot DG as a laboratory analogue for a specific, naturally occurring and policy-relevant decision-making context and explore several explanations as to why the correlation was not perfect.

The naturally occurring and policy-relevant decision-making context was the one in which Ugandan primary school teachers decide how much of the time specified in their contract to actually allocate to teaching. A DG played in school classrooms using the local currency and in which teachers assumed the dictator role and pupils' parents assumed the passive recipient role was selected as the laboratory analogue for this context for two reasons: it would generate a readily interpretable measure of teachers' intrinsic motivations; and it was a good match to a stylized characterization of the teachers' current decision-making context.

A series of tests indicated that the teachers' allocations to parents in the DG were positively correlated with their time allocations to teaching in nature. However, the correlation was weak, especially after controlling for the likely interdependence of observations within schools. Further, consistent with a simple theoretical model, the data indicated that the teachers may be applying different reference points in the two contexts and that the preference to adhere to reference points may be more salient in nature as compared to the lab.

The investigation into why allocating behaviours in the two contexts were only weakly correlated was also guided by the simple theoretical model. This investigation revealed, first, that the difference in reference points between the two contexts may vary across teachers: teachers who were absent for much of the month preceding the study, probably for genuine reasons, appeared to have lower time allocation reference points relative to their DG reference points; teachers who were wealthier relative to the parents participating in the DG appeared to have higher DG allocation reference points relative to their time allocation reference points; and the reference points applied by the teachers in one or both of the contexts appeared also to vary across school, possibly owing to differences in local behavioural norms. Second, the investigation revealed that, in some schools, the teachers' time allocation decisions may have be informally enforced. Specifically, in schools where the School Management Committee had met at least three times in the preceding six months, the relationship between the teachers' DG and time allocations had, quite literally, been pulled apart; the mean difference between the two was significantly larger and this difference was greatest for the teachers making the lowest DG offers. This is consistent with active School Management Committees holding poorly motivated teachers to account in the time allocation decision.[20]

[20] That said, it is worth speculating briefly on whether the reported findings could be owing to a different scenario. Could SMC activity be causing teachers to reduce the DG allocations to parents while leaving their time allocations to teaching unchanged? Despite the fact that the DG is subsequent to the period of time over which SMC activity is measured, this seems unlikely. One can just about imagine a teacher wrongly targeted for non-pecuniary punishment by an SMC choosing to make a low offer in the DG. But why would an SMC target a teacher that is not apparently underperforming?

The correlation between the teachers' time and DG allocations indicates that the DG could be used as a baseline in as series of lab-type experiments designed to investigate what would happen if SMCs were empowered to hold teachers to account. However, the predicted and identified variations in the correlation across teachers suggest that caution would be required when interpreting the findings of such an endeavour.

Somewhat ironically, the finding that the DG may not be the best analogue for the teachers' natural decision-making environment in schools where the School Management Committees are more active provides strong support for the use the DG to generate a measure of teachers' intrinsic motivations. Until now, researchers interested in the impact of local accountability-related organizations on teacher absenteeism have been unable to find evidence of such an effect using observational data.[21] This could be because such organizations respond to underperformance and, without some sort of counterfactual data on teacher performance, a cross-section analysis cannot separate this negative response relationship from the positive impact relationship so the two cancel each other out. The DG provides the required counterfactual – it yields some information about what the teachers *would have* done had the SMCs *not* intervened.

The measured success of this endeavor suggests that future research efforts could usefully be directed towards: formally testing the simple theoretical model used to guide the investigation above; exploiting the DG measure of teacher motivations as either a control or a variable of interest in other policy-relevant analyses, both cross-sectional and relating to planned randomized policy intervention; and, possibly, to testing the conjecture that the external validity of the DG presented above was owing to the steps we took to replicate key aspects of the relational space found in nature in the lab.

References

Adair, J. G., (1984). The Hawthorne effect: a reconsideration of the methodological artifact, *Journal of Applied Psychology*, 69, 334-45.

Alatas, V., L. Cameron, A. Chaudhuri, N. Erkal, and L. Gangadharan (2009). Subject Pool Effects in a Corruption Experiment: A Comparison of Indonesian Public Servants and Indonesian Students, *Experimental Economics*, 12(1), 113-132.

Baran, N. M. P. Sapienza, and L. Zingales (2010). Can we infer social preferences from the lab? Evidence from the trust game, NBER Working Paper No. 15654.

Bardsley, N. (2008). Dictator game giving: altruism or artefact? *Experimental Economics*, 11, 1386-4157.

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer, and R. Sugden (2010). *Experimental Economics: Rethinking the Rules*, Princeton and Oxford: Princeton University Press.

[21] While the track record of field experiments aiming to improve education by strengthening local accountability is mixed, recent work by Duflo et al. (2009) suggests that communities are capable of effective monitoring. We view our approach as complementary to field experiments, since it yields a measure of teacher behaviour in the absence of SMC intervention.

Barr, A., M. Lindelow and P. Serneels (2009). Corruption in public service delivery: An experimental analysis, *Journal of Economic Behavior and Organization*, 72(1), 225-239.

Benz, M., and S. Meier (2008) Do people behave in experiments as in the field?—evidence from donations, *Experimental Economics*, 11, 268–281.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, Reciprocity, and Social History, *Games and Economic Behavior*, 10, 122-42.

Cameron, L., A. Chaudhuri, N. Erkal, and L. Gangadharan (2010). Propensities to Engage in and Punish Corrupt Behavior: Experimental Evidence from Australia, India, Indonesia and Singapore, forthcoming in *Journal of Public Economics*.

Cappelan, A. W, A. D. Hole, E. Ø. Sørenson, and B. Tungodden (2007) The Pluralism of Fairness Ideals: An Experimental Approach. *American Economic Review*, 97(3), 818-27.

Carpenter, J. and C. Myers (2007). Why Volunteer? Evidence on the Role of Altruism, Reputation, and Incentives. IZA Discussion Paper 3021.

Cox, J. C., K. Sadiraj, and V. Sadiraj (2008). Implications of Trust, Fear, and Reciprocity for Modeling Economic Behavior, *Experimental Economics*, 11, 1-24.

Croson, R., and S. Gächter (2010). The science of experimental economics, *Journal of Economic Behavior and Organization*, 73, 122–131.

de Oliveira, A.C.M., Croson, R.T.A., Eckel, C., 2008. Are preferences stable across domains? An experimental investigation of social preferences in the field. CBEES Working Paper #2008-3, University of Texas at Dallas.

Deininger, K. (2003) Does the cost of schooling affect enrolment by the poor? Universal primary education in Uganda, *Economics of Education Review*, 22, 291-305.

Duflo, E, P. Dupas, and M. Kremer (2009). Additional Resources versus Organizational Changes in Education: Experimental Evidence from Kenya. Unpublished, MIT.

Falk, A., and J. J. Heckman (2009) Lab Experiments Are a Major Source of Knowledge in the Social Sciences, *Scence*, 326(23 October, 2009), 535-8.

Fehr E, Kirchsteiger G, Riedl A. (1993). Does fairness prevent market clearing? An experimental investigation, *Quarterly Journal of Economics*, 108, 437–60.

Fehr, E. and A. Leibbrandt (2008). Cooperativeness and Impatience in the Tragedy of the Commons, Institute for Empirical Research in Economics, University of Zurich, Working Paper No. 378.

Frey, B. S. and F. Oberholzer-Gee (1997). The Cost of Price Incentives: An Empirical Analysis of Motivation Crowding-Out, *American Economic Review*, 87, 746-755

Gneezy, U. And A. Rustichini (2000). Pay enough or don't pay at all, *Quarterly Journal of Economics*, 115, 791-810.

Greenwood, J. D., (1982). On the relation between laboratory experiments and social behaviour: causal explanation and generalisation, *Journal of the Theory of Social Behaviour*, 12, 225-49.

Guala, F. (2001). Building economic machines: the FCC auctions, *Studies in the History and Philosophy of Science*, 32, 453-77.

Guala, F. (2002). On the scope of experiments in economics: comments on Siakantaris, *Cambridge Journal of Economics*, 26, 261-67.

Karlan, D. S. (2005). Using Experimental Economics to Measure Social Capital and Predict Real Financial Decisions, *American Economic Review*, 95(5), pp.1688-1699.

Kőszegi, B. & Rabin, M. (2006). 'A model of reference-dependent preferences', *Quarterly Journal of Economics*, 121(4), 1133-1165.

Levitt, S. D., and J. A. List (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, 21(2), 153–74.

List, John A. (2006). The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy*, 114(1), 1-37.

List, J. A. (2007). On the Interpretation of Giving in Dictator Games. *Journal of Political Economy*, 115(3), 482-493.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics, *The Economic Journal*, 109, F25-F34.

Miguel, E. and M. K. Gugerty (2005). Ethnic diversity, social sanctions, and pubic goods in Kenya, *Journal of Public Economics*, 89, 2325-3468.

Plott, C.R. (1982). Industrial organization theory and experimental economics, *Journal of Economic Literature*, 20, 1485–527.

Starmer, C. (1999). Experiments in economics: should we trust the dismal scientists in white coats? *Journal of Economic Methodology*, 6, 1-30.

Sugden, R., (2009). Experiments as exhibits and experiments as tests, *Journal of Economic Methodology*, 12(2), 291-302.