To pay or not to pay: Measuring risk preferences in lab and field

Pablo Brañas-Garza[*]   Lorenzo Estepa-Mohedano[†]   Diego Jorrat[‡]

Victor Orozco[§]   Ericka Rascón-Ramírez[¶]

Abstract

Measuring risk preferences using monetary incentives is costly. In the field, it might be also unfair and unsafe. The commonly used measure of Holt and Laury (2002) relies on a dozen lottery choices and payments, which make it time consuming and expensive. It also raises moral concerns as a result of the unequal payments generated by good and bad luck. Paying some but not all subjects may also create tensions between the researcher and subjects. In a pre-registered study in Honduras, Nigeria and Spain, we use a short version of Holt and Laury where we address all three concerns. We find in the field that not paying at all or paying with and without probabilistic rules makes no difference. Our hypothetical and short version makes our measurement of risk cheaper, fairer and safer.

Keywords: risk preferences, Holt Laury, field experiments, monetary payoffs, incentives

[*]LoyolaBehLab, Universidad Loyola Andalucía, Spain. Email: branasgarza@gmail.com. https://orcid.org/0000-0001-8456-6009.

[†]ETEA-Development Institute, LoyolaBehLab, Universidad Loyola Andalucía, Spain. Email: lorenzo.estepa@fundacionetea.org. https://orcid.org/0000-0002-2883-4280.

[‡]LoyolaBehLab, Universidad Loyola Andalucía, Spain. Email: dajorrat@gmail.com. https://orcid.org/0000-0002-5038-3877.

[§]Development Impact Evaluation (DIME), The World Bank, USA. Email: vorozco@worldbank.org. https://orcid.org/0000-0001-5251-8918.

[¶]Department of Economics, Middlesex University London, UK. ETEA-Development Institute, LoyolaBehLab, Universidad Loyola Andalucía, Spain. Email: E.Ramirez@mdx.ac.uk. https://orcid.org/0000-0002-2214-2224.

1   Introduction

Risk preferences influence a series of economic and social outcomes and behaviors. For example, human capital investment and career decisions (Weiss, 1972; Snow & Warren Jr, 1990), technology adoption and producer behaviour (Chavas & Holt, 1996; Liu, 2013) and the use of financial instruments (Jacobson & Petrie, 2009). Measuring risk preferences has been restricted to lab research, which is often reliant on financially incentivizing research subjects. With the growth of field experiments in development economics, it is unclear if this approach can be scaled up in the field. Risk preference modules can be time-consuming and costly. Paying incentives may fall outside researchers' budgets or be logistically challenging in the field. This study shows that a shorter version of a commonly used measure — Holt & Laury (2002) — and not paying monetary incentives, produce similar risk profiles to those of the more standard and expensive approach where real monetary incentives are used.

The measure of Holt & Laury (2002), hereafter HL, has been widely used to estimate risk parameters of utility functions. This allows researchers to profile individuals based on their risk preference, to study correlates of risk and choices, and to test theories on decision making. Other rigorous measures of risk such as the Balloon Analogue Risk Task (BART) and the Domain-Specific-Risk-Taking (DOSPERT) have been used to test correlation with risky behaviors such as addiction, safety and health-related behaviours (Lejuez et al., 2002; Blais & Weber, 2006) among other domains.[1] We focus here in the HL measure because we can compare our results to a wide range of studies using this measure in the lab.

A common practice among researchers when eliciting risk preferences in the lab is to use monetary incentives (non-hypothetical). However, in the field, the use of such incentives is less common. In particular, when data collection involves thousands of observations, as in national household surveys or impact evaluation data, researchers tend to use measures without monetary incentives (hypothetical).

The use of incentive compatible devices to elicit risk preferences entails two important implications: *i*) subjects make real choices with monetary consequences (hence, risk preferences are not self-reported as in hypothetical measures) and *ii*) final earnings depend on their choices and chance. Subjects are aware of what they can earn and in which circumstances using different devices, as suggested by Holt & Laury (2002); Gneezy & Potters (1997); Eckel & Grossman (2008); Charness & Viceisza (2016) and Crosetto & Filippin (2013).

In the field, researchers must consider various practical issues. First, the use of money or gifts may put at risk the researchers. If the experiment is run in highly deprived areas where carrying any type of reward can put at risk the safety of researchers.[2] Second, the use of lotteries creates inequalities among subjects, at least in the short-run, between winners

---

[1]See Charness et al. (2013) for a review of various elicitation methods of risk preferences.

[2]In some field settings, to reduce safety risks, researchers have adopted the use of non-monetary incentives such as mobile data cards, food baskets, or other in-kind payments. However, in-kind payments also involve considerable logistic effort when data collection takes place in remote areas.

and losers — and this may create tensions in the communities. These inequalities may also trigger frictions between subjects and researchers.[3]

The use of monetary incentives — the gold standard in lab experimental economics — instead of hypothetical choices as in observational studies (and many impact evaluations) is a matter of interest for empirical researchers. Two arguments for incentives are often made: *i*) when payoffs are hypothetical and subjects do not risk their own money, they are more likely to be more risk loving and *ii*) in the absence of monetary incentives, subjects do not put enough effort on the task and consequently, they make random choices.[4] While economists often presume that subjects will not work harder if they do not earn money, psychologists tend to assume that most subjects are intrinsically motivated and therefore will exercise steady effort even in the absence of rewards (Camerer & Hogarth, 1999).

However, paying in the lab and field may also adversely affect responses and behaviors. According to Harrison & List (2004), the nature of the stakes can affect field and lab responses differently. Subjects may adjust their behaviour based on the stake level. In the field, when using surveys as part of data collection for impact evaluation studies, stakes may affect not only the reliability of the measure itself, but they may also induce Hawthorne effects that can affect the identification of treatment effects. This might be the case if the subject believes that certain answers will increase the chances of getting higher stakes in subsequent data collections. Levitt & List (2007) also warns about inaccurate inference that may emerge if the analyst does not account for the differences in stakes across settings.[5]

While the question about the effect of incentives on risk elicitation is not new in the lab, the evidence is mixed and scarce about their effect in the field. *In the lab*, Wiseman & Levin (1996) found that student subjects made the same risky decisions under real and hypothetical consequences. In the same line, Kühberger et al. (2002) found that hypothetical choices match real choices for small as well as for large payoffs. Conversely, Holt & Laury (2002, 2005) found that increasing the size of real payoffs leads to more risk averse behavior than hypothetical payments. Etchart-Vincent & l'Haridon (2011) and Barreda-Tarrazona et al. (2011) found that real and hypothetical choices differ in the gain domain. Overall, when looking into different types of economic experiments, the evidence suggests that monetary rewards matter when effort or performance response to such incentives; Camerer & Hogarth (1999) provide a detailed review of 74 experiments. According to this review, some tasks that seem to affect performance involve recall, judgement on probabilities and

---

[3]While running a pilot of an intervention in another study, one of the co-authors observed a quarrel between a subject who did not win a well implemented public lottery and community members, including NGO facilitators, other subjects and even the community leader. Despite the NGO facilitators explained the chances of winning the lottery beforehand, the subject thought the lottery was manipulated because he did not win the prize.

[4]However, choices are not always random. For instance, in Holt and Laury (2002) subjects select the choice A and do not change their choice across the entire Multiple Price List (MPL) — even when the last choice is strictly dominated. See Amador-Hidalgo et al. (2021) for a detailed analysis of inconsistencies in Holt and Laury task.

[5]Kühberger et al. (2002) discusses further several features that differentiate hypothetical and real decisions.

clerical tasks. Conversely, tasks related to trading in markets, bargaining in games and choosing risky gambles were not found sensitive to such incentives.

In the field, using a sample of Rwandan adults who were subject to financial decisions, Jacobson & Petrie (2009) found that hypothetical and real choices resulted in the same risk preference profiles. When looking into low and high stakes, Bellemare & Shearer (2010) found no differences in risk preferences when comparing both types of payments among tree-planters in Canada. However, using a sample of farmers, the opposite results are found. In Senegal, Charness & Viceisza (2016) compared hypothetical questions with incentivized measures. They found that women were more likely to report willingness to take risks under the hypothetical measure; and when increasing the level of the incentives in India, Binswanger (1980) found that risk aversion tended to increase. The need for additional evidence on the elicitation of risk preferences in the field is highlighted by Levitt & List (2007) when they argue that the choices made by individuals depend not only on financial implications, but also on the nature and degree of others scrutiny, the context in which a decision is embedded, and the manner in which subjects are selected to participate.

The present study aims to answer the following question: Can we elicit risk preferences in the field without using monetary incentives? Our study offers a simple and cheap method to elicit such preferences in the field which we tested in three countries. To our knowledge, this is the first study using the same method of elicitation in both lab and field. Our lab exercise took place in Spain and this was replicated in the field in two developing countries: Honduras and Nigeria. Using our data from these three countries, we compare three different payment schemes widely used in economics and psychology.

We test to what extent the use of hypothetical payoffs has an impact on the measurement of risk preferences. We use the common measure of Holt & Laury (2002) as this is often used to estimate parameters relevant to risk attitude. Our study has two important features. *First*, we ran a lab experiment in Spain (April, 2019) where a reduced version of the Holt & Laury (2002) task with only 5 choices (hereafter HL5, which is half of the 10-choice list) was introduced to reduce the number of choices and time needed. Subjects were randomly assigned to three treatment arms with probability 1/3. Each treatment differs from the other in payment schemes: real payment, paying 1 out of 10 subjects (BRIS)[6] and no payment at all (hypothetical).

Second, we bring our reduced form of the Holt-Laury test to the field. We ran field experiments in two low-middle income countries, Nigeria and Honduras. The former took advantage of a large scale randomized control trial of an educational intervention that targeted parents of 6–9 year-old children. To make our studies as comparable as possible, the study in Honduras used the same target population (i.e., parents of 6–9 year olds) and the same survey instrument used in Nigeria. The main difference was that while the Nigeria study was conducted in a rural setting in Kano, in northern Nigeria; the Honduras study was conducted in Copán, in an urban and peri-urban setting. Eliciting risk preferences

---

[6]*Between Random Incentive Subjects — BRIS.*

in developing countries is relevant as they are the main target of development agencies to promote long-term investments (e.g., human capital, technology adoption, saving) and for which risk preferences are a key determinant of such outcomes. For these countries, we tested whether paying or not to the subjects produces different profiles of risk preferences. As in the lab experiment, all subjects were randomly assigned ($p = 1/3$) to a real payment, payment with probability 1 out of 10 and no payment at all. To compare risk preferences across treatment arms, we contrasted three types of dimensions of the measure: number of inconsistent subjects (and consistency *per* individual), number of safe options and response time.

We find that hypothetical and probabilistic payments have no impact on consistency, number of safe choices or response time. This suggests that when field researchers elicit risk preferences using hypothetical measures, they can trust that these are homogeneous across payment schemes (i.e., real, probabilistic and hypothetical) and that this can be done faster (and therefore cheaper) by using a short version of the Holt-Laury approach. However, paying only a fraction of the sample (probabilistic payment) seemed to introduce noise in the elicitation of risk preference in the lab: subjects became more risk loving when only a fraction of them is paid than when all of them are paid.

Our study was partially run to inform data collection of a large-scale impact evaluation that took place in Northern Nigeria. If our findings would have shown differences between hypothetical and non-hypothetical measures, this would have implied the support for the use of incentives in the field. For this experiment, the cost per subject when eliciting our measures was equivalent to 2.75 US dollars. For the impact evaluation data collection, this would have implied a cost of approximately 25.5 thousand US dollars. Our findings on the comparability of risk preference profiles across payment schemes support the use of hypothetical measures for subsequent analysis.

We now proceed to describe our experimental design, as well as the details of the three experiments encompassing this study. Our last section concludes and discusses some recommendations for future data collection. All studies were pre-registered at AsPredicted: lab experiment, April 10, 2019; field experiments, April 25.

2   Experimental Design

When eliciting risk preferences in the field, observational and experimental studies commonly face the pressing issue of the time spent by enumerators when visiting households. This issue becomes more salient when subjects' attention and time are limited as a result of domestic or job-related chores. To address this issue, we propose a short version of the measure suggested by Holt & Laury (2002) using 5 choices (HL5) instead of 10 (HL), which allows the elicitation of risk preferences in less time (Table 1).[7] We decided to

---

[7]This example is expressed in Naira, the local currency in Nigeria. The exchange rate is 411.5 Naira per US$.

elicit 5 choices, over other possibilities, to keep the extreme choices as in the standard HL (expected value: $q = 0.1 \times \text{High} + 1 - q = 0.9 \times \text{Low}$) and the complementary case ($q = 0.9 \times \text{High} + 1 - q = 0.1 \times \text{Low}$). Our selection of middle points ($q = 0.4, 0.5, 0.6$; $1 - q = 0.6, 0.5, 0.4$, respectively) was based on previous evidence suggesting that the majority of subjects switch between options at these points.

TABLE 1: Trimmed Holt & Laury measure, expected values. The actual choice was listed as a choice between (for example) "1/10 chance of 100N, 9/10 chance of 80N" and "1/10 chance of 200N, 9/10 chance of 2N".

Decision	q	Lottery A	Lottery B
1st	0.1	0.1×100N + 0.9×80N	0.1×200N + 0.9×2N
2nd	0.4	0.4×100N + 0.6×80N	0.4×200N + 0.6×2N
3rd	0.5	0.5×100N + 0.5×80N	0.5×200N + 0.5×2N
4th	0.6	0.6×100N + 0.4×80N	0.6×200N + 0.4×2N
5th	0.9	0.9×100N + 0.1×80N	0.9×200N + 0.1×2N

The use of a short (trimmed) version of the HL might have a direct impact on the elicitation of risk preferences; for instance, under the short version subjects are more likely to be consistent (never choosing A after B has been chosen once, reading down Table 1) than under the long version. However, our main focus in this study is to compare the differences in risk preference profiles across payment schemes.[8] Hence, we use in Study 1, 2 and 3 the HL5 measure. As in HL, every subject is asked to choose between two lotteries A and B. Both A and B offer a low and high payment with probability q and $(1 - q)$.

Note (Figure 1) that the first choice in Table 1 is "trivial" since the expected value of A is much higher than B (82N vs. 21.8N) while the opposite happens in decision 5 (98N vs. 180.2N). In decisions 2, 3 and 4 the expected values of A and B are much closer (88N vs. 81.2N, 90N vs. 101N, 92N vs. 120.8N, respectively) therefore subjects preferences for risk — rather than expected value — might play a critical role in the choice between A and B. Under risk neutrality, subjects should select: A, A, B, B, B. Under consistency, A, B, B, B, B implies risk loving and B, B, B, B, B strong risk loving behaviour. On the contrary, A, A, A, B, B is indicative of slight risk aversion, A, A, A, A, B risk aversion and A, A, A, A, A reflects strong risk aversion. This means that only very risk loving individuals would choose option B in the first decision and very risk-averse would choose A in the last decision.

---

[8]In a previous version of this paper (https://psyarxiv.com/8hknv/), we also included a large version of the HL for the lab experiment. Since we did not have the same long version for the field, we decided to remove this comparison from the paper.
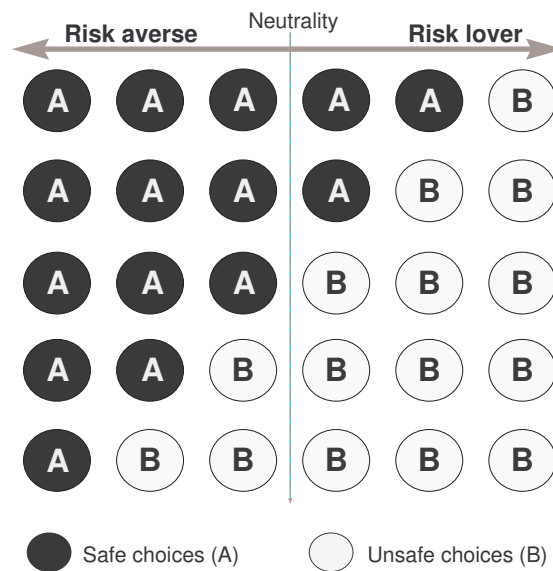
FIGURE 1: Possible choices under consistency. Each column represents possible choices of consistent subjects depending on their risk preferences. Risk averse individuals are represented by the first three columns (left to right) and risk loving by the last three. The first row (top) represents the 1st decision, second row the 2nd decision, and so on. The last decision is the last row of the figure.

2.1 Dimensions of Risk Preferences

In Studies 1, 2 and 3, we compare five dimensions of the elicitation of risk preference under the three payment schemes. The first one is *consistency*, C_i, that is, whether subjects make consistent choices. A typical example of inconsistency is multiple switching from A to B and then from B to A. Any switch from B to A is indeed another example of inconsistency (see Amador-Hidalgo et al. (2021) for a discussion). The second dimension is the *number of safe choices* or risk aversion, RA_i (A instead of B) under consistency. We also explore the number of safes choices for all subjects regardless of consistency, $RA_i(all)$. In the fourth dimension we compute *response time* (RT_i), that is, the duration of the task. This dimension was measured only in Study 2 and 3. Finally, we compute a Goodman and Kruskal's γ for each subject as a measure of the consistency observed within the sequence of choices.[9] Here we summarize our dimensions:

C_i Consistency: whether the subject makes *consistent* choices (= 1) or not (= 0).

RA_i Risk aversion: number of safe choices (A) for *consistent* subjects only, taking values from 0 to 5, where 5 refers to extremely risk averse.

---

[9]Γ is a measure of association (or rank correlation) for ordered classifications. In our case, this measure accounts for the number of pairs that rank the same in the HL5 tasks, minus the number of pairs that do not rank the same, divided by the sum of these two measures.

RA_i (all) Risk aversion: number of safe choices (A) for *all* subjects, taking values from 0 to 5, where 5 refers to extremely risk averse.

RT_i Time: number of seconds spent by the subject to complete the task.

γ_i Consistency *per* subject: number of choice pairs which are consistent over the total number of possible pairs, taking values from 0 to 1, where 1 refers to perfect consistency.

2.2 Experimental Arms and Implementation

Each of Studies 1, 2 and 3 has three experimental arms with probability 1/3. The main difference across arms was the probability of receiving a payment with probability p:

$$T_R: p = 1; T_B: p = 1/10 ; T_H: p = 0$$

Under T_R, all subjects are certain about receiving a payment, whereas under T_H subjects are certain of receiving no payment. For each study, the entire sample was randomly assigned to one out of three treatment arms: T_R refers to real payments, T_B to 1 out of 10 subjects gets paid and T_H to hypothetical payments. Subjects were fully aware of their payment scheme before the elicitation of their preferences.

Study 1 was based on a self-administered questionnaire where subjects were invited to a computer lab. Whereas Study 2 and 3 were conducted in the field by enumerators who were trained by the authors. Using enumerators implied that subjects did not self-manage the instructions. These were read and explained by the enumerator. Enumerators used computer-assisted personal interviewing (CAPI) questionnaires in Nigeria and paper-based interviewing in Honduras. In both cases, enumerators received a list of households they had to visit, including the type of questionnaire they had to apply (i.e., T_R, T_B or T_H questionnaire). The authors conducted the random allocation of treatments prior to the visit to the communities and the enumerators did not have any influence on such selection. For an additional check, a field coordinator monitored the correct use of the lists created by the researchers in the ground.

Prior to running both experiments, the authors piloted the risk preference questionnaire with around 20 subjects to ensure the translations in Hausa (Nigeria) and Spanish (Honduras) were appropriate to the context. All questionnaires and instructions were originally written in English. Data collection for both studies involved a one-day training of enumerators. Training involved in all cases a theoretical explanation of risk preferences and enumerator-subject role play to help enumerators to understand the questions. For both field experiments, enumerators conducted all face-to-face interviews in the households of the subjects. Only one experimental subject was interviewed per household.[10]

---

[10]All survey instruments can be found in the supplement.

3   Study 1 — Lab in Spain

This lab experiment was run in the School of Economics and Business at the University of Sevilla, the largest public university in Andalusia (Southern Spain). The experiment used paper-based questionnaires and was conducted the last week of April 2019.

In most lab experiments, subjects are university students who are self-selected into the experiment and hold high levels of education (see Exadaktylos et al. (2013)). The main advantage of lab experiments is the absolute control of the conditions faced by the subject: a) subjects cannot interact among them, unless interaction is required by the experiment; and b) there are no external distractions. For these reasons, a lab experiment is the cleanest approach to test our main research question: *do monetary incentives make a difference in the elicitation of risk preferences?*

The HL5 used in the lab is identical to the MPL shown in section 2 in the example. We just adjusted the monetary values to be meaningful to Spanish subjects. Thus, for Study 1, lottery A offers 5 euros with q probability and 4 euros with $(1 - q)$; lottery B offers 10 and 0.1 euros with probabilities q and $(1 - q)$, respectively.

3.1   Sample and balance

The entire sample consisted of n=178 Spanish subjects[11]. Each subject was randomly assigned to one out of three payment schemes (T_R: 60, T_B: 57, T_H: 62). Table 2 shows the balance between sub-samples. Table 2 shows no differences across treatments regarding age, gender, cognitive reflection test (CRT) and courses (freshmen, sophomore, etc).

TABLE 2: Balance across treatments in Study 1.

	obs.	mean$_{T_H}$	$T_R - T_H$	p	$T_B - T_H$	p
Age	179	21.721	–0.155	0.790	–0.318	0.589
Male	179	0.492	0.008	0.929	–0.018	0.846
Courses	179	2.984	–0.017	0.936	–0.370	0.086
CRT	179	0.710	0.090	0.531	–0.043	0.768

Note: H refers to hypothetical, R real payment and B probabilistic payment with 1 out of 10 chance of winning (BRIS).

[11]We invited n=298 subjects: Approximately 60% subjects entered in the study of the three different payment schemes (n=179) while around 40% (n=119) entered in the experiment where we compare the reduced HL5 vs. HL10 (see footnote 8).

3.2 Results

Table 3 shows the impact of payment schemes on the risk preference measure. The top of the table focuses on consistency (C_i) where n_c corresponds to the number of consistent subjects. The second row shows the number of safe choices (RA_i) under consistency and the third one $RA_i(all)$ the number of safe choices regardless of the consistency of the subjects' choices. The last row focuses on the gamma consistency measure (γ_i).

TABLE 3: Summary results in the Lab — Study 1, Spain

Dimensions		(1) T_R	(2) T_B	(3) T_H	(4) $\hat{\beta}_{T_B}$	(5) $\hat{\beta}_{T_R}$
C_i	n_c	49	49	53	–0.009	–0.052
	%	81.6	85.9	86.8	(0.890)	(0.428)
RA_i	mean	2.938	2.796	2.285	–0.540	0.133
	st. dev.	1.068	1.219	1.258	(0.107)	(0.670)
$RA_i(all)$	mean	2.850	2.333	2.774	–0.460	0.071
	st. dev.	1.070	1.285	1.151	(0.135)	(0.803)
γ_i	mean	0.929	0.950	0.964	–0.014	–0.035
	st. dev.	0.184	0.133	0.110	(0.605)	(0.185)
Obs.		60	57	62		

Note: Column 4 and 5 display LPM (C_i), negative binomial coefficients (RA_i & $RA_i(all)$) and OLS coefficients (γ_i) for each outcome variable (row) and for the Real and BRIS treatments (column). Our reference group is the hypothetical treatment (T_H). In $RA_i(all)$ we use all subjects, while in RA_i only the consistent ones.

Columns 1–3 show different statistics for each treatment. The percentage of subjects making consistent choices is shown in row 1. These percentages are not significantly different between treatments. Row 2 and 3 show the mean and the standard deviation of the number of safe choices. Again, we do not observe significant differences either.

Finally, columns 4 and 5 provide the regression coefficients of each treatment for the two types of outcome variables: consistency and risk aversion responses. For C_i, we estimate a linear probability model (LPM), for both RA_i measures a negative binomial regression model and for γ_i an OLS.[12] $\hat{\beta}_{T_B}$ refers to the comparison between BRIS and Hypothetical payoffs, whereas $\hat{\beta}_{T_R}$ to the comparison between Real and Hypothetical payoffs.

---

[12]The number of safe choices is a discrete variable, which reflects the number of occurrences of choosing A. The distribution of this variable is discrete, not continuous, limited to non-negative values and present a high number of 0. So, we applied a negative binomial regression model which account for all these restrictions.

1299

Column 5 in Table 3 shows that risk preferences measured under real payments do not differ from those under hypothetical payment. This is true for the number of consistent individuals (C_i), the level of risk aversion (RA_i and $RA_i(all)$) and the gamma consistency measure (γ_i).[13]

The cumulative distributions of the number of safe choices by treatments shown in Figure 2 confirms the results found on the level of risk aversion (RA_i) for treatment T_R, but not for T_R. While there is no difference between T_H and T_R (black and blue lines, respectively), the cumulative distribution of T_B is above these two. The latter suggests that paying 1 out of 10 makes people choose less safe choices than the other two payment schemes. We performed a Kolmogorov-Smirnov test for equality of distributions between the hypothetical and each of the incentivized schemes. The results confirm that there are no statistical differences between the distribution of safe choices under the Hypothetical and Real treatments; however, there are for the BRIS treatment when compared with the hypothetical and real payment scheme (T_H vs. T_R, $p = 0.80$; T_H vs. T_B, $p = 0.03$; T_R vs. T_B, $p = 0.02$). The same conclusion is found for $RA_i(all)$. Since T_B involves two lotteries (to be selected for payment and the HL), a possible explanation might be that subjects consider that both lotteries are correlated.[14]

One particular feature highlighted by Camerer & Hogarth (1999) is higher variance in choices or in task performance when using hypothetical measures. To study this, we performed a series of variance ratio tests. In all of them, we do not reject the null hypothesis of equal variance of our dimensions across treatments: RA_i (T_H vs. T_R, $p = 0.18$; T_H vs. T_B, $p = 0.41$; T_R vs. T_B, $p = 0.13$). The same conclusion is found for $RA_i(all)$ (T_H vs. T_R, $p = 0.29$; T_H vs. T_B, $p = 0.41$; T_R vs. T_B, $p = 0.22$).[15]

To summarize our main findings for Study 1:

Result 1 — Lab (H vs. R): *Compared to hypothetical payments in the lab: paying all the subjects has no impact on consistency or number of safe choices.*

Result 2 — Lab (H vs. B): *Compared to hypothetical payments in the lab: paying 1/10 of the subjects has no impact on consistency, but decreases the level of risk aversion.*

Study 1 analyses risky decision-making in the lab. We address whether the payment scheme impacts the measurement of risk preferences.

---

[13]These results hold if we use a non-parametric test, as the Kendall rank correlation (K), where the null hypothesis is that both variables are independent (in this case, the set of pairs under treatments). To carry out the test, we firstly define a categorical variable (0 for T_H, 1 for T_B and 2 for T_R) and then perform the test. Our results show no significant differences in γ_i coefficients across treatments ($p = 0.37$). Our findings are also true for paying 1 out 10 (T_B).

[14]Under the idea of a "hot hand", subjects may assume that if they were lucky in the first lottery, they will also be lucky in the second one (Miller & Sanjurjo, 2018).

[15]Our results hold after adjusting for age, gender, cognitive reflection test score and session fixed effects, see supplement.

FIGURE 2: Cumulative distributions for the number of safe choices.
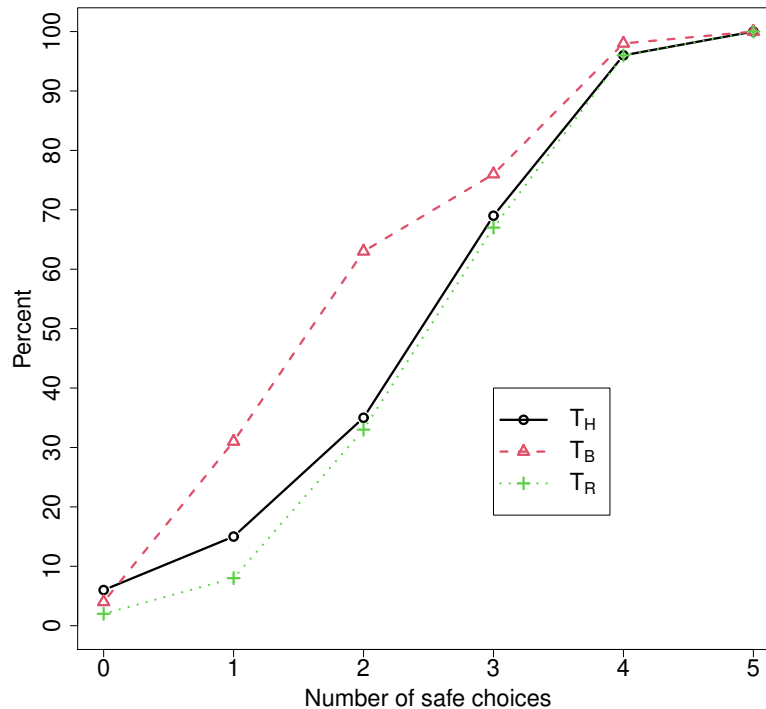
Result 1 shows that paying all the subjects or none have no significant effect on consistency or number of safe choices. This is a critical issue for field data collection since samples (and therefore costs) are substantial when household visits are required.

Result 2, on the other hand, shows that BRIS payments may affect risk taking (although do not alter inconsistency).

Overall, Study 1 served to test whether incentives are important (and apparently are not). In the following two sections, we will test Result 1 and 2 in the field — Nigeria and Honduras — using the short version of Holt-Laury.

4   Study 2 — Field in Nigeria

This study was conducted in Kano (Nigeria) in April 2019. We ran the field experiment in three rural villages in Kano: Dorayi, Ja'en and Gidan Maharba, where 360 households were randomly selected. Our sample in each village was selected according to the eligibility criterion of having at least one child between 6 and 9 years old. We selected households with these characteristics as this experiment was used to inform the elicitation of risk preferences for a large-scale impact evaluation that took place in rural areas in Northern Nigeria.

The experiment was composed of four modules: social norms (coordination games), subjective expectations, time and risk preferences. The HL5 was the last module we elicited. Random assignment to T_R, T_B or T_H treatments remained the same throughout the entire experiment. Response time (RT_i) was collected using CAPI for all four modules.

To elicit risk preferences, we used the same HL5 as in the lab experiment. We adapted the currency to Naira. On average, subjects earned 126 Naira (equivalent to 0.33 US dollars), see example in subsection 2.[16]

4.1 Sample and balance

Sample size is $n = 360$ (by treatments, T_R: 120, T_B: 124, T_H: 116). In this sample, 25% has no education, 8% has completed primary education and 32% has completed the secondary education. Around 40% of the sample is female and the average age is 39 years old. This sample is quite different from the sample used in the lab, but reflects common characteristics of the populations studied in Development Economics, which are usually the target of interventions aiming at increasing savings and human capital investment.

Table 4 shows the balance across treatments. We observe significant differences in age, where subjects in the real and BRIS treatment (T_R and T_B) have 2.8 years more than those in the hypothetical treatment (T_H). However, these few differences are no longer significant after adjusting p-values using Bonferroni corrections.

TABLE 4: Balance across treatments in Study 2.

	obs.	mean$_{T_H}$	$T_R - T_H$	p	$T_B - T_H$	p
Age	360	39.605	2.695**	0.049	2.807**	0.038
Female	360	0.403	−0.003	0.957	0.033	0.606
Education	360	9.245	−0.154	0.856	0.548	0.514
Sufficient[+]	360	0.816	−0.007	0.885	−0.006	0.902

Note: H refers to hypothetical, R real payment and B probabilistic payment with 1 out of 10 chance of winning (BRIS). Significance levels: ***$p < 0.01$ and **$p < 0.05$. Sufficient[+] refers to sufficient money to eat.

4.2 Results

Table 5 reports a summary of results. The top of the table shows consistency (C_i), where n_c and % correspond to the number and percentage of consistent subjects, respectively. The second and third lines focus on the number of safe choices or risk aversion (RA_i and $RA_i(all)$), the fourth one reports the response time (RT_i) and the last row focuses on the gamma consistency measure (γ_i). The number of observations per treatment appears at the bottom of the table. Columns 1, 2 and 3 report consistency, the number of safe choices, response time and gamma measure by experimental arms T_R, T_B and T_H. The last columns

---

[16]We planned payments in order to cover one-day average wage for the entire experiment. This daily average wage was equal to 1080 Naira (3 US$). We paid (on average) 350N in coordination, 405N in time discount and 126N in risk.

show the regression coefficients of outcome variables on the treatment variables T_B and T_R, having as reference group the hypothetical measure.

TABLE 5: Summary results in the field – Study 2, Nigeria

| Dimensions		(1) T_R	(2) T_B	(3) T_H	(4) $\hat{\beta}_{T_B}$	(5) $\hat{\beta}_{T_R}$
C_i	n_c	117	119	111	–0.029	0.001
	%	97.5	94.4	97.4	(0.227)	(0.957)
RA_i	mean	3.085	2.899	3.135	–0.238	–0.049
	st. dev.	1.534	1.607	1.615	(0.303)	(0.832)
$RA_i(all)$	mean	3.083	2.857	3.088	0.230	0.004
	st. dev.	1.515	1.578	1.621	(0.501)	(0.990)
RT_i	mean	3.345	3.385	3.469	–0.084	–0.124
	st. dev.	1.107	1.132	1.101	(0.561)	(0.397)
γ_i	mean	0.982	0.955	0.975	–0.020	–0.007
	st. dev.	0.121	0.199	0.152	(0.337)	(0.747)
Obs (n)		120	126	114		

Note: Column 4 and 5 display LPM (C_i), negative binomial (RA_i & $RA_i(all)$) and OLS coefficients (RT_i in logs and γ_i) for each outcome variable (row) and for the Real and BRIS treatments (column). Our reference group is the hypothetical treatment (T_H). The time variable is expressed in logs.

As shown in Table 5 the fraction of consistent individuals, the mean of safe choices and the response time are nearly the same across treatments, which implies that there are no significant differences. To test differences in coefficients across treatments, as in Study 1, we ran LPM for C_i and a negative binomial regression for RA_i and $RA_i(all)$. For RT_i and γ_i, we ran an OLS. The last two columns show the results of these regressions: column (4) focuses on the BRIS treatment and column (5) on the real case. The reference group is the hypothetical payment in both cases.

For C_i (first row), we do not find significant differences in the chances of making consistent choices between treatments. The estimated impact of paying 1/10 (vs. hypothetical payment) is not significant ($p = 0.227$), the same finding we observe when comparing the real payment with the hypothetical ($p = 0.957$). Our results are robust to different model specifications, see the supplement.

For RA_i and $RA_i(all)$ (second and third rows), we do not find significant differences in the number of safe choices either. The estimated coefficients reported in column (4) and (5) are not significantly different from zero (RA_i: $p = 0.303$, $p = 0.832$; $RA_i(all)$: $p = 0.501$,

FIGURE 3: Cumulative distributions for the number of safe choices in Nigeria.
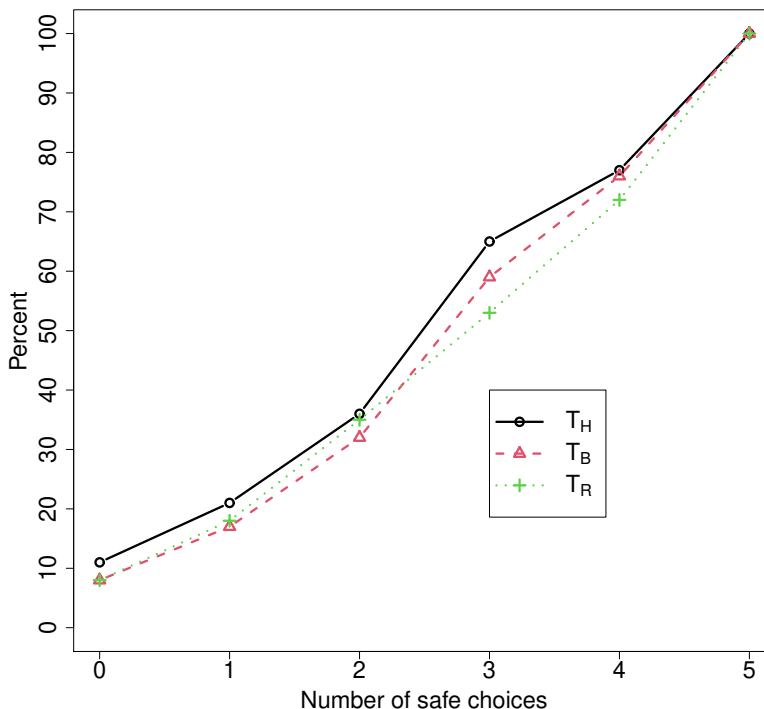
$p = 0.990$).[17]. As in Study 1, we analyzed the variance of RA_i across treatments. We do not find significant differences between them (T_H vs. T_R, $p = 0.29$; T_H vs. T_B, $p = 0.48$; T_R vs. T_B, $p = 0.31$). The same conclusion is found for when using all choices $RA_i(all)$ (T_H vs. T_R, $p = 0.23$; T_H vs. T_B, $p = 0.38$; T_R vs. T_B, $p = 0.33$).

For RT_i, subjects making incentivized choices are 4 seconds faster than those making hypothetical choices, but the difference is not significant. (The coefficient is expressed after transforming the logs). When comparing the hypothetical payment with $1/10$, the mode of payment does not make any difference — subjects take the same time. And for the γ_i consistency measure (last row), we do not find any effect of T_B or T_R when compared with the hypothetical measure ($p = 0.337$ and $p = 0.747$)[18].

Finally, Figure 3 shows the cumulative distributions of the number of safe choices RA_i for the three treatments. The three lines are similar to each other and they cross for some values of safe choices. A Kolmogorov-Smirnov test found no statistical differences between the distributions (T_H vs. T_R, $p = 0.99$; T_H vs. T_B, $p = 0.43$; T_R vs. T_B, $p = 0.99$). The same conclusion is found for $RA_i(all)$.

---

[17]Our results remain the same after adjusting for age, gender, education and enumerator fixed effect — see the supplement. We adjusted for enumerator fixed effect to account for the potential differences in the level of understanding of the instructions by the enumerators. To reduce measurement error derived from the misunderstanding of instructions, in both Honduras and Nigeria, we used a show card where all A and B choices where shown to the subjects.

[18]As before, these results hold when we test the Kendall correlation of the payment discrete variable and gamma ($p = 0.92$).

In this study, we find that the payment scheme has no impact on consistency, the number of safe choices an the response time. It also shows no difference in distributions between payment schemes.

To summarize our main findings for Study 2:

Result 3 — Nigeria (H vs. R): *Compared to hypothetical payments in the field: paying all subjects has no impact on consistency, number of safe choices or response time.*

Result 4 — Nigeria (H vs. B): *Compared to hypothetical payments in the field: Paying 1/10 of the subjects has no impact on consistency, number of safe choices or response time.*

Our findings in Study 2 show that the elicitation of risk preferences with monetary incentives in the field provides the same results as the elicitation without incentives.

5   Study 3 — Field in Honduras

We ran the same experiment in Honduras. Our aim with this sample was to test whether Results 3 and 4 replicate in a different location, having subjects with similar socioeconomic characteristics to the Nigerian sample, but living in peri-urban areas. Our goal responds to the increasing demand for replication and validation of experimental studies (Banerjee et al., 2017; Peters et al., 2018; Al-Ubaydli et al., 2017). Study 3 took place in Santa Rosa de Copán (Honduras) between May 1 and 14, 2019 in the districts of Osorio, El Carmen, Prado Alto and Santa Teresa. We selected 360 households based on the same eligibility criterion as the Nigerian sample: having a child between 6 and 9 years old.

As in Nigeria, the experiment consisted of four tasks (coordination, expectations, risk and time preferences). The main difference with the Nigerian experiment is that the risk preference task was the third one, instead of last one as in Study 2. The original random assignment of subjects to T_R, T_B or T_H remained the same for the entire experiment. RT_i was recorded by the enumerator for the entire block of risk preferences.

To elicit risk preferences, we used the same reduced HL task (HL5) as in Study 2 and Study 3. We adapted our experiment to the local currency (Lempiras). On average, subjects earned 18 Lempiras in the HL5; for the first set of choices, in lottery A we paid 50L or 40L (instead of 100N and 80N) and in lottery B 100L or 1L (instead of 200N and 2N). We followed the same structure of payments as described in see section 2.[19]

5.1   Sample and balance

The total sample consisted of 360 subjects. Each subject was randomly assigned to 1 out of 3 arms resulting the following distribution. T_R: 109, T_B: 126, T_H: 125. Table 6 shows the balance across treatments. We found significant differences in age, where the subjects

---

[19]The daily wage of unskilled workers in this peri-urban communities is about 150L (6 US$).

allocated to the real payment (T_R) had 3.3 years less than those allocated to the hypothetical payment (T_H). However, this difference is not economically important (9% of the average age).

TABLE 6: Balance across treatments in Study 3.

	obs.	$mean_{T_H}$	$T_B - T_H$	p	$T_R - T_H$	p
Age	359	35.750	−1.059	0.406	−3.310**	0.006
Male	360	0.144	0.015	0.746	0.034	0.437
Education	358	9.048	0.170	0.763	−0.002	0.997
Sufficient[+]	360	0.768	0.009	0.854	0.003	0.962

Note: H refers to hypothetical, R real payment and B probabilistic payment with 1 out of 10 chance of winning (BRIS). Inference was made using robust standard errors. Significance levels: **$p < 0.05$. Sufficient[+] refers to sufficient money to eat.

5.2 Results

Table 7 summarizes the results. As in Study 2, the top of the table refers to consistency (C_i), the second and third lines to risk aversion (RA_i and $RA_i(all)$), the fourth to response time (RT_i) and the last row focuses on the gamma consistency measure (γ_i). The number of observations per treatment appears in the bottom of the table. Columns 1, 2 and 3 respectively show the relevant values for T_R, T_B and T_H. We use LPM, negative binomial and OLS regressions, as in Study 1 and Study 2. Column 4 focuses on the BRIS payment and column 5 on the real case, in both cases the reference group is the hypothetical payment. For consistency C_i (first row), both treatments T_B and T_R are not significant ($p = 0.605$ and $p = 0.378$, respectively). These results are robust to different specifications (see the supplement). Also, our payment schemes have no impact on risk aversion. However, we found a marginally significant effect of both treatments on response time.

As before, Figure 4 shows the cumulative distributions for the three treatments and confirms the results found on the level of risk aversion (RA_i). While there is no difference between T_R and T_B lines, T_H line is below the other two lines. However, a Kolmogorov-Smirnov test found no significant differences between treatments in their cumulative distributions (T_H vs. T_R, $p = 0.80$; T_H vs. T_B, $p = 0.99$; T_R vs. T_B, $p = 0.98$), and the same conclusion holds for $RA_i(all)$. As in Study 1 and Study 2, our results suggests no difference in variance for RA_i (T_H vs. T_R, $p = 0.42$; T_H vs. T_B, $p = 0.37$; T_R vs. T_B, $p = 0.45$) or $RA_i(all)$ (T_H vs. T_R, $p = 0.48$; T_H vs. T_B, $p = 0.47$; T_R vs. T_B, $p = 0.48$).

For RT_i, we find a positive and marginally significant effect of T_B and T_R on response time ($p = 0.08$ and $p = 0.05$). These results suggests that using hypothetical measures might be quicker to implement (or subjects pay less attention to it) than when using monetary

1306

TABLE 7: Summary results in the field — Study 3, Honduras

Dimensions		(1) T_R	(2) T_B	(3) T_H	(4) $\hat{\beta}_{T_B}$	(5) $\hat{\beta}_{T_R}$
C_i	n_c	80	102	98	0.027	–0.048
	%	73.4	80.9	78.2	(0.605)	(0.378)
RA_i	mean	1.787	1.833	2.030	–0.193	–0.241
	st. dev.	1.572	1.554	1.608	(0.392)	(0.320)
$RA_i(all)$	mean	2.000	1.936	2.144	–0.206	–0.141
	st. dev.	1.472	1.479	1,468	(0.267)	(0.462)
RT_i	mean	5.120	5.104	4.994	0.110	0.126
	st. dev.	0.424	0.499	0.543	(0.082)	(0.054)
γ_i	mean	0.884	0.911	0.901	0.010	–0.016
	st. dev.	0.237	0.213	0.222	(0.724)	(0.585)
Obs (n)		109	126	125		

Note: Column 4 and 5 display LPM (C_i), negative binomial (RA_i & $RA_i(all)$) and OLS coefficients (RT_i in logs and γ_i) for each outcome variable (row) and for the Real and BRIS treatments (column). Our reference group is the hypothetical treatment (T_H). The time variable is expressed in logs.

incentives. Finally, for the γ_i consistency measure, we do not find significant differences between the incentivized measure (T_B and T_R) and the hypothetical one ($P = 0.724$ and $p = 0.585$, respectively). As before, these results are robust using the Kendall correlation between gamma and the categorical variable that distinguishes between treatments ($p = 0.46$).

Our main findings from Study 3 show that the payment scheme has no impact on consistency or the number of safe options. We also find that there are no differences in the distributions of neither of these variables. However, our results show that paying all or 1 out of 10 subjects increase the time response by 20 and 15 seconds, respectively.

To summarize our main findings for Study 3:

Result 5 — Honduras (H vs. R): *Result 3 is partially replicated in Honduras: Compared to hypothetical payments in the field, paying all subjects has no impact on consistency or number of safe choices. R (weakly) increases response time.*

Result 6 — Honduras (H vs. B): *Result 4 is partially replicated in Honduras: Compared to hypothetical payments in the field, paying 1 out of 10 subjects has no impact on consistency*

FIGURE 4: Cumulative distribution for the number of safe choices in Honduras.
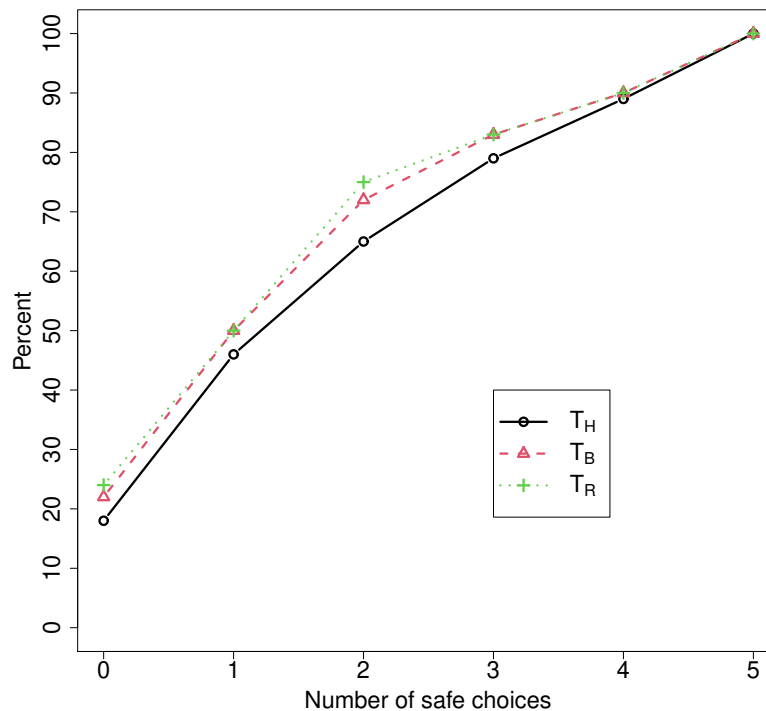
or number of safe choices. B (weakly) increases response time.

Our findings from Study 3 reinforce our conclusion from Study 2. The elicitation of risk preferences with monetary incentives in the field yields the same results as in the hypothetical scheme. When comparing our field results with previous studies, the percentage of risk averse subjects in Nigeria and Honduras, 65.7% and 29.8% respectively, falls within the range observed in other countries: 40% of risk averse subjects in Senegal (Charness & Viceisza, 2016) and 85% in India (Chakravarty et al., 2011). In the lab, our study finds that 74.3% of our subjects are risk averse. This percentage is similar to other lab experiments ranging from 64% to 84% (Charness et al., 2018; Taylor, 2013).

To test whether other alternative specifications may provide different results, we estimated a Random Utility Model[20] to analyse both Honduras and Nigeria data. Using the whole sample (or even restricting our analysis to only consistent subjects) our findings remain the same. We also applied this method to the lab data with identical results.

Our findings in Nigeria and Honduras are important for two reasons. *First*, existing hypothetical measures collected in the field are indeed informative. *Second*, future measurements of risk preferences in the field do not need to use monetary incentives to get accurate proxies of risk preferences.

---

[20]See Andersen et al. (2008); Apesteguia & Ballester (2018); Apesteguia et al. (2019).

6   Equivalence Tests

In this section we test whether our estimates are different across treatments comparing estimates within a range instead of with respect to a point estimate.

A possible alternative is to explore whether or not the observed effect is large enough to be deemed worthwhile. This procedure is called *equivalence* (Lakens, 2017; Wellek, 2010). It testes whether the observed effect falls within or outside of an equivalence interval, defined by two exogenous bounds: lower ($-\gamma_L$) and upper (γ_U). To test for equivalence, a two one-sided test (TOST) approach is applied. Two composite null hypotheses are tested: $H_{01} \rightarrow \gamma \leq -\gamma_L$ and $H_{02} \rightarrow \gamma \geq \gamma_U$. When both null hypotheses are rejected, we can conclude that $-\gamma_L < \gamma < \gamma_U$ or that the observed effect falls within the equivalence bounds and it is *close enough to zero to be practically equivalent* (Lakens, 2017).

Our objective and exogenous bounds are defined based on Holt & Laury (2005) where they found a difference of one in the average number of safe choices between real and hypothetical incentives. Hence, our equivalence level is equal to one where we define the equivalence interval for each TOST as $H_{01} \rightarrow \gamma \leq -1$ and $H_{02} \rightarrow \gamma \geq 1$. To check the robustness of our results, we also use two additional equivalence levels: 0.50 and 0.75.

To consider a thorough analysis of our main findings, we use both the null hypothesis significance test (NHST) and the equivalence test (ET).[21] Table 8 summarizes the results from the equivalence test. Panel A shows that in the lab, paying 1 out 10 subjects yield to *relevant* differences with respect to hypothetical decisions, while using real payments to *equivalent* results. In both field studies (panel B and C), equivalence tests suggest that both treatments (real and BRIS) yield equivalent results to the hypothetical payment scheme.

To summarize the equivalence tests results:

Result 7 — Equivalence (H vs. R):   *Hypothetical and real measures yield to equivalent results in both the lab and field.*

Result 8 — Equivalence (H vs. B):   *BRIS and Hypothetical payment mechanisms yield to equivalent results **only** in the field.*

7   Conclusion

This paper studies the impact of monetary incentives in the elicitation of risk preferences using Holt & Laury (2002). We report on a lab and two field experiments with different subject pools. Our lab experiment (Study 1 — Spain) took place in a developed country using

---

[21]According to these tests, there are four possible outcomes in the analysis: *i*) the observed effect can be statically equivalent ($-\gamma_L < \gamma < \gamma_U$) and not statistically different from zero (Equivalence or *E*); *ii*) statistically different from zero but not statistically equivalent (Relevant Difference or *RD*); *iii*) statistically different from zero and statistically equivalent (Trivial Difference or *TD*); and *iv*) neither statistically different from zero nor statistically equivalent (Undetermined or *U*).

TABLE 8: Equivalence testing results: Lab and Field.

Equivalence level:	0.5		0.75		1	
	$P-value$	Conclus.	$P-value$	Conclus.	$P-value$	Conclus.
Panel A: Lab						
T_B	0.518	RD	0.154	RD	0.019	TD
T_R	0.004	E	0.000	E	0.000	E
Panel B: Nigeria						
T_B	0.104	I	0.007	E	0.000	E
T_R	0.016	E	0.000	E	0.000	E
Panel C: Honduras						
T_B	0.088	E	0.007	E	0.000	E
T_R	0.141	I	0.017	E	0.000	E

Note: P-value refers to the maximum p-value obtained from the two one-side test. Conclus. shows the conclusion considering both test NHST and ET. The baseline is the group making decisions with Hypothetical money (T_H).

a pool of undergraduate students, the common population used in lab experiments. The second and third experiment (Study 2 and 3 — Nigeria and Honduras) were implemented in rural villages of a middle-income country and in peri-urban towns of a low-income country. In both cases using a subject pool with similar characteristics: parents with children between 6–9 years old living in deprived areas.

Our study answers a simple question: *Can we elicit risk preferences in the field without using monetary incentives to pay subjects?* We found in this study that this is possible. Lab experimentalists, who have widely used monetary incentives in the lab when eliciting such measures can trust the measures that their field counterparts collect, in most of the cases, using hypothetical payments; and field experimentalists can now trust that these measures are consistent across payment schemes (i.e., real, probabilistic and hypothetical) and that this can be done faster (and therefore cheaper) by using a short version of the Holt-Laury approach.

Our concrete findings are summarized in here. Using a short version of Holt & Laury (2002) in the lab and field, produces the same risk preference profiles (in terms of consistency and safe choices) regardless of the payment scheme (hypothetical or real payment). In particular:

- When comparing the HL5 measures, hypothetical (no payment) and real payment schemes generate the same consistency and risk levels, that is, they yield equivalent measures.

- When comparing the HL5 measures, hypothetical (no payment) and BRIS scheme

generate the same consistency and risk levels in the field, while in the lab the measurement of risk preferences display higher levels of risk loving attitudes under the probabilistic scheme.

Empirical researchers interested in the elicitation of risk preferences at a low cost, but also concerned about the potential negative consequences of the use of lotteries in the field, may want to consider our reduced version of the hypothetical HL to measure risk preferences without incentives. This approach is faster (and cheaper) and does not create any asymmetries among subjects. Field experimentalists who not only aim at reducing data collection costs, but also at minimising feelings of unfairness among experimental subjects, should consider hypothetical payment schemes to minimise potential frictions between subjects and researchers, or among winners and losers.

References

Al-Ubaydli, O., List, J. A., & Suskind, D. L. (2017). What can we learn from experiments? Understanding the threats to the scalability of experimental results. *The American Economic Review*, *107*(5), 282–86.

Amador-Hidalgo, L., Brañas-Garza, P., Espín, A. M., García-Muñoz, T., & Hernández-Román, A. (2021). Cognitive abilities and risk-taking: Errors, not preferences. *European Economic Review*, *134*, 103694.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2008). Eliciting risk and time preferences. *Econometrica*, *76*(3), 583–618.

Apesteguia, J. & Ballester, M. A. (2018). Monotone stochastic choice models: The case of risk and time preferences. *Journal of Political Economy*, *126*(1), 74–106.

Apesteguia, J., Ballester, M. Á., & Gutierrez, A. (2019). Random models for the joint treatment of risk and time preferences. *Mimeo*.

Banerjee, A. V., Chassang, S., & Snowberg, E. (2017). Decision theoretic approaches to experiment design and external validity. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of Economic Field Experiments*, volume 1 (pp. 141–174). North-Holland.

Barreda-Tarrazona, I., Jaramillo-Gutiérrez, A., Navarro-Martínez, D., & Sabater-Grande, G. (2011). Risk attitude elicitation using a multi-lottery choice task: Real vs. hypothetical incentives. *Spanish Journal of Finance and Accounting*, *40*(152), 613–628.

Bellemare, C. & Shearer, B. (2010). Sorting, incentives and risk preferences: Evidence from a field experiment. *Economics Letters*, *108*(3), 345–348.

Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, *62*(3), 395–407.

Blais, A.-R. & Weber, E. U. (2006). A domain-specific risk-taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*(1), 33–47.

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), 7–42.

Chakravarty, S., Harrison, G. W., Haruvy, E. E., & Rutström, E. E. (2011). Are you risk averse over other people's money? *Southern Economic Journal*, *77*(4), 901–913.

Charness, G., Eckel, C., Gneezy, U., & Kajackaite, A. (2018). Complexity in risk elicitation may affect the conclusions: A demonstration using gender differences. *Journal of Risk and Uncertainty*, *56*(1), 1–17.

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization*, *87*, 43–51.

Charness, G. & Viceisza, A. (2016). Three Risk-elicitation Methods in the Field: Evidence from Rural Senegal. *Review of Behavioral Economics*, *3*(2), 145–171.

Chavas, J.-P. & Holt, M. T. (1996). Economic behavior under uncertainty: A joint analysis of risk preferences and technology. *The Review of Economics and Statistics*, *78*(2), 329–335.

Crosetto, P. & Filippin, A. (2013). The 'bomb' risk elicitation task. *Journal of Risk and Uncertainty*, *47*(1), 31–65.

Eckel, C. C. & Grossman, P. J. (2008). Forecasting risk attitudes: An experimental study using actual and forecast gamble choices. *Journal of Economic Behavior & Organization*, *68*(1), 1–17.

Etchart-Vincent, N. & l'Haridon, O. (2011). Monetary incentives in the loss domain and behavior toward risk: An experimental comparison of three reward schemes including real losses. *Journal of Risk and Uncertainty*, *42*(1), 61–83.

Exadaktylos, F., Espín, A. M., & Brañas Garza, P. (2013). Experimental subjects are not different. *Nature Scientific Reports*, *3*, 1213.

Gneezy, U. & Potters, J. (1997). An experiment on risk taking and evaluation periods. *The Quarterly Journal of Economics*, *112*(2), 631–645.

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

Holt, C. A. & Laury, S. K. (2002). Risk aversion and incentive effects. *The American Economic Review*, *92*(5), 1644–1655.

Holt, C. A. & Laury, S. K. (2005). Risk aversion and incentive effects: New data without order effects. *The American Economic Review*, *95*(3), 902–912.

Jacobson, S. & Petrie, R. (2009). Learning from mistakes: What do inconsistent choices over risk tell us? *Journal of Risk and Uncertainty*, *38*, 143–158.

Kühberger, A., Schulte-Mecklenbeck, M., & Perner, J. (2002). Framing decisions: Hypothetical and real. *Organizational Behavior and Human Decision Processes*, *89*(2), 1162–1175.

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362.

Lejuez, C. W., et al. (2002). Evaluation of a behavioral measure of risk taking: the Balloon Analogue Risk Task (BART). *Journal of Experimental Psychology: Applied*, *8*(2), 75–84.

Levitt, S. D. & List, J. A. (2007). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives*, *21*(2), 153–174.

Liu, E. M. (2013). Time to change what to sow: Risk preferences and technology adoption decisions of cotton farmers in China. *Review of Economics and Statistics*, *95*(4), 1386–1403.

Miller, J. B. & Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, *86*(6), 2019–2047.

Peters, J., Langbein, J., & Roberts, G. (2018). Generalization in the tropics–development policy, randomized controlled trials, and external validity. *The World Bank Research Observer*, *33*(1), 34–64.

Snow, A. & Warren Jr, R. S. (1990). Human capital investment and labor supply under uncertainty. *International Economic Review*, *31*(1), 195–206.

Taylor, M. P. (2013). Bias and brains: Risk aversion and cognitive ability across real and hypothetical settings. *Journal of Risk and Uncertainty*, *46*(3), 299–320.

Weiss, Y. (1972). The risk element in occupational and educational choices. *Journal of Political Economy*, *80*(6), 1203–1213.

Wellek, S. (2010). Equivalence tests for multivariate data. In *Testing Statistical Hypotheses of Equivalence and Noninferiority* (pp. 251–280). Chapman and Hall/CRC.

Wiseman, D. B. & Levin, I. P. (1996). Comparing risky decision making under conditions of real and hypothetical consequences. *Organizational Behavior and Human Decision Processes*, *66*(3), 241–250.