UNSW Business School Research Paper No. 2015 ECON 20
First version September 2015
UNSW Business School Research Paper No. 2015 ECON 20A
This version October 2015

MTurk 'Unscrubbed': Exploring the good, the 'Super', and the unreliable on Amazon's Mechanical Turk

A.M. Jeanette Deetlefs
Mathew Chylinski
Andreas Ortmann

MTurk 'Unscrubbed': Exploring the good, the 'Super', and the unreliable on Amazon's Mechanical Turk[♦]

A.M. Jeanette Deetlefs[*], Mathew Chylinski[♯], Andreas Ortmann[§]

Abstract

Widely accepted as a low-cost, fast-turnaround solution with acceptable validity, Amazon's Mechanical Turk (MTurk) is increasingly being used to source participants for academic studies. Yet two commonly raised concerns remain: the presence of quasi-professional respondents, or "Super-Turkers", and the presence of "Spammers", those that compromise quality while optimising their pay rate. We isolate the influence on research results of experienced subjects (Super-Turkers), and of unreliable subjects (Spammers), jointly and separately. Jointly including these subjects produces very similar results to jointly excluding them, yet effect sizes decrease disproportionately to their sample representation. Furthermore, separately including experienced subjects in research results is shown to be as problematic as inclusion of unreliable subjects, although the noise introduced by these subjects is divergent and measure dependent. Hence removing only one of these types of respondents can be even more damaging to the reliability of results, than including both.

Keywords: data collection, experimentation, field experiment, internet, Mechanical Turk

JEL Classification: C90, C91, D80

[♦] Research support from Australian Research Council LP110100489 and the UNSW Business School (Deetlefs) is gratefully acknowledged.

[*] School of Marketing, University of New South Wales, NSW 2052, Australia, j.deetlefs@unsw.edu.au

[♯] School of Marketing, University of New South Wales, NSW 2052, Australia, m.chylinski@unsw.edu.au

[§] School of Economics, University of New South Wales, NSW 2052, Australia, a.ortmann@unsw.edu.au

Introduction

Amazon's Mechanical Turk (MTurk) is an increasingly used platform for academic research (Horton et al. 2011). In commerce too, ready availability of on-line respondents has solved the traditional market research problem of sourcing subjects (Leet 2015). However, as MTurk has grown in popularity, the reliability and trustworthiness of results has increasingly been questioned (e.g., Marder and Fritz 2015 and the work of David Rand cited therein[1]). In particular the presence of quasi-professional respondents, or "Super-Turkers", and "Spammers", those who maximise their pay rate with little regard for response quality, is a major concern (Bohannon 2011; Mason and Suri 2012).

Research to date has shown that laboratory and MTurk experimental results (Horton et al. 2011; Paolacci et al. 2010) can reach similar conclusions. These conclusions are often based on results after the removal of unreliable subjects, since unreliable subjects have the potential to generate outliers resulting in misleading statistics and poor interpretation (Ratcliff 1993; Stevens 1984; Tukey 1977). Indeed, Chandler et al. (2014) found that approximately a third of MTurk based research had between 3% and 37% of 'questionable' responses removed. Yet no agreed protocol exists in terms of identifying, removing and reporting on unreliable subjects. Less easily identified for removal are the excessively experienced subjects, like Super-Turkers, who can also detrimentally alter standard measures like the cognitive reflection test (Chandler et al. 2014) and reduce effect sizes on commonly used experimental tasks (Chandler et al. 2015). Practice effects can also lead to unnatural strategizing and sharpened response times, potentially marring experiments designed to emulate one-off type decisions, like retirement investment (Camerer and Loewenstein 2004). Yet here too, there is

[1] About their running example, Marder and Fritz say: "'I am never going to be absolutely undistracted, ever,' Marshall says, and smiles. Her employers don't know that Marshall works while negotiating her toddler's milk bottles and giving him hugs. They don't know that she has seen studies similar to theirs maybe hundreds, possibly thousands, of times."

no consistent approach to identify the excessively experienced post-hoc. Participation of subjects in specific types of research is not tracked in MTurk. Researchers either have to ask subjects to self- report their level of experience (Berinsky et al. 2012; Chandler et al. 2015; Paolacci and Chandler 2014), or maintain a database to identify experienced subjects (Berinsky et al. 2012; Paolacci et al. 2010). It is unclear to what extent the post-hoc exclusion or inclusion of experienced and unreliable subjects alters results and may affect their validity. In this paper we examine the interaction of these two types of problematic subjects on results.

Among the 2736 completed responses, one per subject, across twelve studies that we conducted on MTurk in 2014, we identify approximately 9% of subjects as having previously participated in our experiments (the Experienced) and 11% of subjects[2] as being unreliable (the Unreliable). The Experienced, we suggest, represent Super-Turkers given the practice effects gained through the highly consistent and repetitive nature of our experimental tasks. The Unreliable, we suggest, represent the Spammers combining faster overall completion times with poor question completion. Focusing on the most recent three of our MTurk studies, we compare results when responses of the Experienced and the Unreliable are included and jointly or separately excluded (the joint condition producing so-called clean responses). In addition, we directly compare one of these three recent MTurk studies with a laboratory study. In support of previous research (as discussed in Mason & Suri 2012), we find little difference when both types of responses are excluded. Yet their exclusion doubles our effect sizes, indicating that Super-Turkers and Spammers add opposing noise and impact results differently. Indeed, measures with an objective interpretation like response times, demographic data or financial-literacy assessments, highlight significant differences between the Experienced and the Unreliable. For example, when compared with the clean responses, the Experienced are almost 38% faster on timed tasks; and the Unreliable score 10% lower on

[2] 13% if we control for over-quota and software problems.

objective financial-literacy assessments. Moreover we note when including these responses separately, the Experienced markedly dampen differences in response times on critical tasks while unreliable subjects confound research outcomes more randomly. Ultimately both types of responses are shown to be untrustworthy and undesirable.

Our research isolates the joint and separate influence of experience and unreliability on MTurk results. We demonstrate, first, empirically the scope of the problem by reviewing our research studies between May and December 2014. We show, second, their detrimental influence on the absolute outcome of timed tasks and objective assessments. Third, we highlight their impact on effect sizes when jointly included. Fourth, we demonstrate that removing responses of only one of either the Experienced or the Unreliable (the latter being common practice) could possibly be even more detrimental to conclusions, than including both. Fifth, we propose a rigorous approach to identify these responses post-hoc, with stringent design requirements for experiments conducted online.

We proceed with a discussion of MTurk and the relevant studies. Following this we discuss our studies, profiling the experienced and unreliable subjects and examining how they influence results. Finally we discuss the impact of our findings.

Amazon's Mechanical Turk

MTurk offers an on-demand scalable online labour market with access to more than 500 000 workers from 190 countries (Amazon Mechanical Turk 2015). Almost 50% of the MTurk workforce (Turkers) are US-based (Paolacci and Chandler 2014). This pool of subjects is willing to work for compensation well below the minimum wage, e.g. between $1.40 and $6 per hour (Berinsky et al. 2012; Bohannon 2011; Jacquet 2011; Mason and Suri 2012). Researchers, termed requesters, post their requests for a HIT (Human Intelligence Task) online with details of expected task time, task details, and remuneration offered. Turkers select the HIT that appeals to them. The HIT can be linked to the unique URL of an experiment. On completion, Turkers are given a code which they paste on the HIT page. Once this is done, the HIT is no longer accessible to them and Turkers can be automatically remunerated within a certain time period e.g. 8 hours, unless the requester chooses to block the Turker. Since requesters can limit the accessibility of their HIT to Turkers who have never been blocked or with a certain acceptance rate e.g. 75%/95%/99%, blocking Turkers limits their work prospects. Requesters can also limit the accessibility of their HIT to Turkers in a certain country, for example the US.

MTurk is recognised as a low-cost, painless, diverse, large, and stable source of participants for fast-turnaround studies (Bohannon 2011; Jacquet 2011; Mason and Suri 2012). Turkers compare favourably with a general internet sample (Buhrmester et al. 2011) but are described as over-educated and under-employed, less religious and more liberal than the average US resident (Berinsky et al. 2012; Paolacci and Chandler 2014; Paolacci et al. 2010).

MTurk results have been found to be just as valid as laboratory and field experiments. After restricting MTurk results to only those who correctly answered comprehension questions, Horton et al. (2011) successfully replicated laboratory results for a prisoner’s dilemma game. After removing unreliable responses, Paolacci et al. (2010) successfully replicated three of the psychological ‘bias’ tests, the “Linda problem”, the “Asian disease problem” and the “physician problem” (Tversky and Kahneman 1981, 1983) among 131 Turkers and 141 mid-western laboratory students and found no discernible difference. (Note however, that the research of Charness et al. (2010) suggests that replication of the “Linda problem” may be a reflection of weak incentivisation and limited communication and learning opportunities for subjects. This implies that the validity of MTurk still faces some substantive questions that have long been the subject of contention in the methodological debates among economists and psychologists see for example Hertwig and Ortmann (2001).)

MTurk use requires caution. Turkers, unlike laboratory subjects, can be distracted (Chandler et al. 2014; Marder and Fritz 2015). In Chandler et al. (2014), Turkers surveyed admit to multi-tasking with 18% watching television, 14% listening to music and 6% instant messaging while completing HITs. Spammers may target the highest paying HITs (Bohannon 2011; Horton et al. 2011; Marder and Fritz 2015; Mason and Suri 2012). Programs or 'bots' may complete tasks (Crump et al. 2013; Mason and Suri 2012). Demand effects have also been noted when using objective assessments as checking questions, for example asking the name of a politician (Paolacci and Chandler 2014).

Concerns around excessively experienced Turkers remain, particularly since this problem is likely to grow as academic use of MTurk increases. Surveying 291 Turkers with a minimum 75% approval rate for their work, Rand et al. (2014) found their median number of completed academic surveys to be 300, as opposed to 15 for the 118 physical laboratory subjects surveyed. This high level of experience has obvious implications for the speed at which they complete surveys, but also increases the likelihood of demand effects. Participants are likely to recognise commonly used attention checks (Marder and Fritz 2015) and successfully respond to them, making the inattentive ever more difficult to eliminate from studies. Chandler et al. (2015) examine this problem by persuading 638 Turkers to complete the same twelve experimental tasks across two waves, spaced at intervals of days, a week or a month apart. Between the two waves, they note a reduction in effect sizes of about 25%. With excessive experience, standard assessment tools such as the cognitive reflection test (CRT) may fail in their diagnosticity. Indeed Chandler et al. (2014) find that CRT scores are significantly moderated by the number of previously completed HITs. Furthermore, practice effects can mean that participants lose the freshness required to respond naturally to the tasks that experiments require (Marder and Fritz 2015, quoting Rand). Faced with a task that may be interpreted as strategic and is designed to be a one-off task, like a lottery, their previous exposure to similar tasks may also change their response (Camerer and Loewenstein 2004; Hertwig and Ortmann 2001). For example if subjects had chosen the highest risk option before and lost, this could affect their likelihood of choosing this option again.

Assessing the scope of the problem

We conducted twelve studies on MTurk between May 2014 and December 2014. Each of these studies shared common tasks, attitudinal scales and demographic details designed to explain the difference in behaviour among participants randomly allocated to either the treatment or no-treatment conditions.

Table 1 shows a summary of responses to the studies conducted including the month in which the study was undertaken, the expected pay and expected time as shown to participants, the number of responses started and the number completed. (Note that although Study S4 had a within-subjects design, the numbers in **Table 1** relate to individual subjects). The completed responses are broken down into the usable and the unusable. Unusable responses include over-quota responses and responses lost due to software problems.

One important MTurk drawback is the limited subject selection criteria in the standard interface. This can create a need for over-sampling when applying quotas. In our last two studies (S4 and S5), subjects were quota-controlled. We experienced software problems since our experiment software, Qualtrics, was unexpectedly offline during some of the experiments. In the 'Experienced' column, we report on subjects previously exposed to our experiments. Within each study we identified common self-reported Turker identification numbers and IP addresses. These were classified as duplicates and were typically around

0.4% of completed responses.[3] Following the first study, (S1) we created a master database to identify whether subjects had previously participated in one of our experiments. In addition to identifying these based on Turker identity (id) number, we also excluded common IP addresses to err on the side of caution since we did not want to include subjects who may have discussed our experiment with others.[4] Subjects who had seen our experiments before were classified as Experienced and together with the duplicates made up 9% of all completed responses. The ‘Unreliable’ column refers to haphazard responses. We used a multiple criteria approach to identify the Unreliable (as described under Studies S4 and S5). This was used for studies S1 and the last three studies; for studies S1a to S3 we made use of a single checking question. (One of the financial-literacy questions was repeated near the end of the experiment and subjects who did not recognise it, were marked as unreliable.) Despite this slight difference in approach, the Unreliable responses comprised 13% of all completed responses, controlling for software problems and over-quota. In the last two columns, we compare first the number of unusable and usable responses, and second calculate the proportion of usable responses of all who completed the experiment but were not excluded due to software problems.

[3] While it is not possible for the same Turker to participate more than once in a common HIT with batches, where the Turker id was blank, but the IP address was common, we classified these as duplicates.

[4] As part of their terms and conditions, MTurk only allows an individual Turker to have a single Turker id number. It is possible for Turkers to have a common IP address if they all work at a large firm or are working from a common coffee shop or home (Berinsky et al. 2012).

Table 1: Overview of studies conducted

									Percent of completed						
Name	**Date created**	**Expected pay**	**Expected time**	**Started**	**Attrition rate**	**Completed**	**Over-quota**	**Software problems**	**Experienced**	**Unreliable**	**Unreliable/ Completed (net)**	**Unusable responses**	**Usable responses**	**Unusable/ Usable**	**Usable/ Completed**
Pay rate per hour: $7.70															
S1	May-14	$3.85	30m	378	0.08	348	0	0	0.01	0.16	0.16	54	291	0.19	0.84
Pay rate per hour: $6.42															
S1a	Jun-14	$2.14	20m	154	0.07	143	0	0	0.09	0.10	0.10	27	116	0.23	0.81
Pay rate per hour: $6.00															
S1b	Jun-14	$2.00	20m	133	0.16	112	0	0	0.13	0.10	0.10	25	87	0.29	0.78
S1c	Jun-14	$2.00	20m	81	0.26	60	0	0	0.05	0.18	0.18	14	46	0.30	0.77
S1d	Jun-14	$2.00	20m	180	0.20	144	0	1	0.06	0.20	0.20	37	106	0.35	0.74
S2a	Jun-14	$2.00	20m	89	0.18	73	0	1	0.10	0.15	0.15	18	54	0.33	0.75
S2b	Jul-14	$2.00	20m	121	0.20	97	0	0	0.09	0.14	0.14	23	74	0.31	0.76
S2c	Jul-14	$2.00	20m	180	0.21	143	0	1	0.06	0.10	0.11	24	118	0.20	0.83
S3	Nov-14	$0.80	8m	336	0.28	243	0	0	0.12	0.06	0.06	43	200	0.22	0.82
Pay rate per hour: $3.00															
MTurk vs Lab	Nov-14	$1.00	20m	222	0.31	154	0	0	0.09	0.15	0.15	37	117	0.32	0.76
S4 within	Dec-14	$1.00	20m	689	0.24	522	224	27	0.10	0.08	0.15	94	177	0.53	0.36
S5 between	Dec-14	$1.00	20m	922	0.24	697	179	8	0.12	0.09	0.13	151	354	0.43	0.51
Total				3485	0.21	2736	408	38	0.09	0.11	0.13	547	1740	0.32	0.65

Table shows details of 12 studies conducted between May 2014 and December 2014. Subjects were informed what they could expect to be paid, 'Expected pay', i.e. basic plus average bonus, and how long the experiment would take, 'Expected time'. The pay rate per minute was adjusted over time and the four panels organise the studies accordingly. 'Started' shows the number of responses started. 'Attrition rate' shows the difference between those that 'Started' and those 'Completed' as a percentage of 'Started'. 'Completed' shows all collected responses. To achieve minimum sample sizes for quotas, over-sampled responses are shown in 'Over-quota'. 'Software problems' are incomplete responses due to off-line software. Experienced are subjects who have previously been exposed to our experiments. ‘Experienced’ shows these as a percentage of 'Completed'. ‘Unreliable’ shows haphazard responses as described in the text as a percentage of 'Completed' and as a percentage of Completed (net) when controlling for software and over-quota responses. 'Unusable responses' is the sum of Experienced and Unreliable responses. 'Usable responses' is the balance after removing 'Over-quota', 'Software problems' and 'Unusable responses' from 'Completed'. 'Unusable/Usable' shows the ratio of unusable responses for each study. 'Usable/Completed' shows 'Usable' responses as a proportion of all that 'Completed' excluding 'Software problems'.

Across the studies the basic rate of pay was held constant at $0.50. In addition to this base pay, subjects could earn a bonus. Together they are presented as the ‘Expected pay’. The ‘Expected pay’ remained consistent when a common objective was being tested across studies. Including the bonus, the first study (S1) offered subjects $7.70 per hour, while studies S1b through to S3 offered subjects $6 per hour. The last three studies halved that rate of pay to $3 per hour. These rates of pay highlight the appeal of MTurk as an inexpensive source of subjects. However, the last column of **Table 1** also shows the added sample cost caused by the Experienced and the Unreliable, aggravated by inefficient quota-control capabilities in the last two studies. Whereas 79% of the completed responses proved usable generally, this dropped to only 44% on these two studies, thus increasing the cost per completed response.

Table 1 also highlights the role that incentivisation plays. Attrition, the difference between those that started the experiment and those that completed it, is shown as a percentage of the starters as an attrition rate. The average attrition rate of 22% varied from 8% for S1, to 20% across studies S1a to S3 and to 25% for the three most recent studies. This suggests that the attrition rate can be managed by offering a larger bonus in addition to the basic pay.

Studies S4 and S5

In this section we focus on the results of the last two of our twelve studies. This allows us to identify a sizeable proportion of subjects as either having been exposed to our experiment and measures more than once (Experienced), or as completing the experiment and other measures in a haphazard fashion (Unreliable).

In December 2014 we conducted two experiments (S4 and S5) to test the time spent on a risky choice in the presence or absence of a treatment. As is the typical approach when emulating uncertainty in economics, or judgment and decision research, our experiments made use of incentivised risky gambles or so-called “lotteries” (Charness et al. 2013; Eckel and Grossman 2008; Gneezy and Potters 1997; Payne and Bettman 2004; Payne et al. 1993). To test the interaction effect of level of expertise, type of choice and the applied treatment on the time taken, each experiment had a 3 (financial expertise: low, medium, high) x 2 (choice 1; choice 2) x 2 (no treatment; treatment) full factorial design. The experiments were identical although the first (S4) used a within-subjects approach to the choice factor and the second (S5) used a between-subjects approach. Turkers were redirected to the experiment in Qualtrics via a URL posted in the HIT. The experiment consisted of five stages. In stage 1, subjects completed a financial-literacy assessment to allow quota-control based on level of financial expertise before being randomly allocated to a condition. In stage 2, subjects completed a filler task for which they earned 6000ECU.[5] In stages 3 and 4 subjects were tasked with investing these earnings in risky choices (lotteries) that we timed. In stage 5, subjects completed a questionnaire with demographics and attitudinal scales. Our treatment was applied to stage 4.

Measures

To objectively assess and compare subjects’ financial expertise and cover three levels of financial expertise in our studies, the financial-literacy questions were as previously used in Europe, the US and Australia (Bateman et al. 2012; Lusardi and Mitchell 2007, 2009; van Rooij et al. 2011). These consist of basic numeracy, financial-literacy and sophisticated financial-literacy with each question having a single correct answer. (Appendix 1 contains a copy of the questions used.)

The filler task in stage 2 served to create a sense of ownership of the earned, rather than endowed, earnings to be used for the investments (Cherry et al. 2002). In stages 3 and 4, each

[5] ECU = experimental currency units. We converted these to ensure the ‘Expected pay’ would be the average bonus plus the $0.50 basic pay.

subject made risky choices in the form of lotteries. Stage 3 used a modified version of the lottery of Eckel and Grossman (2008). This allowed us to establish subjects' risk preference while timing the choice. Stage 4 used lotteries based on that of Gneezy and Potters (1997) to test if subjects would consistently match their risk preference with stage 3 and to time their choices. We created two versions of the stage 4 lottery task, choice 1 and choice 2. (Choice 1 had very similar expected value outcomes, while choice 2 had divergent expected value outcomes.) In the within-subjects study (S4), subjects were therefore asked to complete three risky choices, the Eckel-Grossman (2008) lottery together with choice 1 and choice 2. In the between-subjects study (S5) subjects saw either choice 1 or choice 2 after completing the Eckel-Grossman (2008) lottery. All the risky choices were timed as in Wilcox (1993).

In stage 5, subjects completed basic demographics such as age, gender, level of education, income and whether they were employed or not. Subjects also completed fourteen attitudinal scales taken mainly from psychology and marketing literature. These scales varied on the number of items each measured, the number of points on the Likert agreement scale and whether they were consistently positively or negatively phrased. Contrary to our priors, which expected scale reliability to be compromised by the separate or joint exclusion of the Experienced and the Unreliable, there were few consistent differences. (Full details on these attitudinal scales, their reliability and ratings are included in Appendix 5.)

Identifying the Experienced and the Unreliable

'Super-Turkers' are effectively professional respondents and consequently are likely to have completed the experiment before. To minimise their number in our experiments, our HITs informed Turkers of how many times we had run this experiment and asked them not to participate if they had done so before. (Appendix 2 contains screenshots of the HIT). Turkers were also told that they would not receive the bonus if we identified them. [6] Despite this, using our master database of Turker id numbers and IP addresses, we were able to identify 11%[7] of participants for these two studies as having seen the experiment, and most of the other measures contained in the questionnaire, more than once.[8] (Across all twelve studies conducted in 2014, basic and sophisticated financial-literacy, the lottery tasks and most of the attitudinal scales were common.)

In contrast with laboratory subjects, Turkers are widely criticised for being easily distracted (Buhrmester et al. 2011; Marder and Fritz 2015; Paolacci et al. 2010) and are known to multi-task (Chandler et al. 2014). Distracted subjects may take longer on a decision or may even leave the computer and resume the experiment at a later stage. This behaviour would render our timed choice results unreliable. At the other extreme, subjects can distort results by skimming through tasks and giving random responses to complete the experiment as quickly as possible. Hence we tried to minimise the participation of unreliable subjects by limiting access to the HIT to Turkers with a 99% acceptance rate. Post hoc, we adopted a stringent approach to identifying those who had completed the experiment in a haphazard fashion. Our approach is multi-pronged to avoid measurement error as can occur when using a single question (Mason and Suri 2012; Paolacci and Chandler 2014). We flagged subjects who answered a repeated question inconsistently, or if they failed to recognise it. Similar attitudinal scale items were compared, flagging subjects with large discrepancies. On scales with more than eight items, subjects who consistently gave the same response to all items were flagged. If a subject was flagged three times for haphazardness, the maximum possible number, the response was deemed unreliable. If subjects completed scale questions or even the entire questionnaire faster than readable, they were deemed unreliable. Extremely long

[6] The scale of the problem is likely to be even larger when subjects are not asked in advance to refrain from repeat participation.

[7] To repeat, across all twelve studies conducted in 2014, 9% of responses were Experienced.

[8] It is possible with the use of Turkprime.com to specify Turker id numbers for those that may not participate.

response times (more than three standard deviations away from the mean) were also deemed unreliable. (Appendix 3 discusses this process in more detail.)

The Datasets

By tracking the IP addresses of subjects we were able to verify that 97.5% of our subjects were in fact in the US when taking the experiment as required in our HIT design. For both studies, our distribution of subjects across the US in terms of time zones, matched that of the population.[9]

The final datasets for the two studies were weighted to adjust the three levels of financial-literacy from quota-controlled proportions to their naturally occurring proportions.[10]

Who are the Experienced and the Unreliable?

In **Table 2** we compare the results when the Experienced and the Unreliable are separately excluded as well as when they are jointly excluded (Excluding All) or jointly included (Including All) for each of studies S4 and S5.

The Unreliable are significantly more likely to be male and younger, typically aged around 28 years. The Experienced are significantly more likely to earn less than $75000 p.a. than subjects with clean responses, while the Unreliable are significantly more likely to claim to earn more than $75000 p.a. Focusing on the larger sample of study S5, we note that the Experienced are significantly less likely to be in full-time employment and earn less despite being more educated than the Unreliable. This confirms the running example of Marder and Fritz (2015), Marshall, as prototypical (see footnote 1).

[9] We compared our time zone distribution with that sourced from Fuller Dynamic (2012), http://fullerdynamic.com/news/2012/8/8/distribution-of-us-population-by-time-zone, accessed on 20 May 2015.

[10] Across studies S1 to S3 we ascertained the mean and identified the level of financial-literacy one standard deviation above and below the mean to use for classifying subjects as high, medium or low.

Table 2: Comparing the Experienced and the Unreliable with Clean Responses (studies S4 and S5)

	Study S4						Study S5					
	Excluding All	Excl. Experienced	Experienced	Excl. Unreliable	Unreliable	Including All	Excluding All	Excl. Experienced	Experienced	Excl. Unreliable	Unreliable	Including All
Unweighted obs	177	218	53	230	41	271	354	420	85	439	66	505
Weighted obs	179	207	55	234	28	262	353	411	78	431	58	489
	mean	mean	mean	mean	mean	mean	mean	mean	mean	mean	mean	mean
	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)
Demographics												
Gender (F=1)	0.56	0.54	0.37	0.52	0.393	0.50	0.53	0.51	0.48	0.52	0.38[1]	0.51
	(0.49)	(0.50)	(0.47)	(0.49)	(0.58)	(0.50)	(0.49)	(0.50)	(0.51)	(0.50)	(0.51)	(0.50)
Age	37.30	35.87	36.48	37.12	26.75**[2]	35.25	34.15	33.39	34.24	34.17	28.78**[3]	33.53
	(12.76)	(13.02)	(11.80)	(12.54)	(8.12)	(12.78)	(11.64)	(11.54)	(11.88)	(11.69)	(9.36)	(11.60)
Full-time employed	0.09	0.09	0.13	0.10	0.09	0.10	0.18	0.19	0.10[4]	0.16	0.24	0.17
	(0.29)	(0.30)	(0.33)	(0.30)	(0.35)	(0.30)	(0.38)	(0.39)	(0.31)	(0.37)	(0.45)	(0.38)
Highest level of school is High school	0.37	0.37	0.39	0.37	0.35	0.37	0.31	0.32	0.27	0.30	0.41	0.31
	(0.47)	(0.05)	(0.47)	(0.47)	(0.57)	(0.48)	(0.45)	(0.46)	(0.46)	(0.46)	(0.52)	(0.46)
Earn less than $75000p.a.	0.88	0.88	0.87	0.88	0.86	0.88	0.89	0.88	0.96*[5]	0.91	0.79*[6]	0.89
	(0.32)	(0.33)	(0.32)	(0.32)	(0.41)	(0.33)	(0.30)	(0.32)	(0.20)	(0.29)	(0.42)	(0.31)

Table shows mean scores and standard deviations (sd) for each of the Experienced, Unreliable as well as when these responses are excluded, 'Excluding' or included, 'Including' for studies S4 and S5. Female are allocated a 1 for gender: Unreliable are significantly more likely to be male (S5). Subjects in full-time employment are coded a 1, versus others coded a 0: Experienced subjects are less likely to be in full-time employment (S5). Subjects whose highest level of school is high school are coded a 1, versus 0 if more highly educated. Subjects that earn less than $75000p.a. are coded a 1 versus those who earn more, coded a 0: Experienced subjects are more likely to earn less than $75000p.a.(S5)

	Study S4						Study S5					
	Excluding All	Excl. Experienced	Experienced	Excl. Unreliable	Unreliable	Including All	Excluding All	Excl. Experienced	Experienced	Excl. Unreliable	Unreliable	Including All
Unweighted obs	177	218	53	230	41	271	354	420	85	439	66	505
Weighted obs	179	207	55	234	28	262	353	411	78	431	58	489
	mean	mean	mean	mean	mean	mean	mean	mean	mean	mean	mean	mean
	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)	(sd)
Numeracy (5 items)	4.58	4.51	4.34	4.52	4.07^{*7}	4.47	4.68	4.64	4.34^{*8}	4.62	4.40^{9}	4.59
	(0.72)	(0.86)	(0.81)	(0.75)	(1.51)	(0.85)	(0.62)	(0.71)	(1.13)	(0.74)	(1.11)	(0.80)
Basic financial-literacy (5 items)	3.85	3.74	3.62	3.80	3.07^{*10}	3.72	3.71	3.63	3.60	3.69	3.19^{**11}	3.63
	(1.04)	(1.22)	(1.24)	(0.77)	(2.04)	(1.22)	(1.07)	(1.13)	(1.33)	(1.12)	(1.41)	(1.16)
Sophisticated financial-literacy (4 items)	2.83	2.75	3.08^{12}	2.89	2.28^{**13}	2.82	2.79	2.76	2.70	2.77	2.60	2.75
	(1.00)	(1.08)	(1.11)	(1.03)	(1.45)	(1.10)	(1.11)	(1.13)	(1.28)	(1.14)	(1.22)	(1.15)
Time on lottery (s)	53.54	51.59	33.69^{**14}	48.88	39.27	47.84	63.92	65.50	39.48^{**15}	59.49	75.10	61.35
	(28.42)	(46.73)	(23.49)	(28.54)	(119.91)	(43.29)	(35.47)	(92.88)	(36.51)	(36.87)	(242.18)	(87.28)
min.	13.51	2.40	3.76	3.76	2.40	2.40	11.64	2.28	2.87	2.87	2.28	2.28
max.	250.31	673.80	102.57	250.31	673.80	673.80	332.23	1354.25	284.83	332.23	1354.25	1354.25

Significance testing between sample excluding either previous or unreliable:

^{1} z=-2.12, p=0.03
^{**2} t=-5.30, p=0.00
^{**3} t=-4.24, p=0.00
^{4} z=-1.79, p=0.09
^{*5} z=1.96, p=0.04
^{*6} z=-2.25, p=0.02
^{*7} t=-2.02, p=0.05
^{*8} t=-2.34, p=0.02
^{9} t=-1.77, p=0.08
^{*10} t=-2.14, p=0.03
^{**11} t=-2.73, p=0.01
^{12} t=1.80, p=0.07
^{**13} t=-2.53, p=0.01
^{**14} t=-3.30, p=0.00
^{**15} t=-3.68, p=0.00

Table shows mean scores and standard deviations (sd) for each of the Experienced, Unreliable as well as when these responses are excluded, 'Excluding' or included, 'Including' for studies S4 and S5. Mean scores are shown for sum of 5 numeracy items, sum of 5 basic financial-literacy items and sum of 4 sophisticated financial-literacy items (Bateman et al. 2012, Lusardi & Mitchell 2009; van Rooij et al., 2007, Lusardi & Mitchell 2007). Unreliable score significantly lower on each of these assessments (S4). Experienced subjects score higher on sophisticated financial-literacy (S4). Mean time on the first lottery (Eckel & Grossman 2008) is shown together with minimum and maximum times. Experienced subjects are significantly faster at this, the first choice (S4 & S5).

How do the Experienced and the Unreliable influence results?

Beyond the demographics of subjects from studies S4 and S5 already discussed **Table 2** also shows financial-literacy measures and response times. The financial-literacy measures contained numeracy, basic financial-literacy and sophisticated financial-literacy questions, each progressively more difficult. (See Appendix 1 for details together with a comparison, where possible, with weighted US data from the American Life Panel (Lusardi and Mitchell 2009).) The Unreliable and the Experienced responses have significantly fewer correct numeracy answers than the clean responses. This problem is also evident among the Unreliable for the basic and sophisticated financial-literacy questions. Overall, across all fourteen financial-literacy questions, the unreliable responses contain 10% fewer correct answers ($t(409) = 3.22$, $p < 0.001$, $M_{\text{Unreliable}} = 10.19$, SD = 2.98; $M_{\text{Excluding}} = 11.17$, SD = 1.98).

In terms of the absolute results of response time (Table 2) those with previous exposure to the task are more than 38% faster than those seeing our experiment for the first time. When comparing the time spent on the first lottery task, subjects in S5 (S4) previously exposed to our experiment spent 39.48s (33.69s) as opposed to the 63.92s (53.54s) of subjects with clean responses ($t_{S5}(429) = -3.68$, $p < 0.001$, $M_{\text{Experienced}} = 39.48$ seconds, SD = 36.51; $M_{\text{Excluding}} = 63.92$ seconds, SD = 35.47).

Nevertheless, **Table 2** reveals that despite significant differences between the results of the Experienced and the Unreliable and the clean sample (‘Excluding’), results at an overall level are only slightly influenced by their inclusion (‘Including’). This suggests that the differences in the results of the Experienced and the Unreliable wash out. Indeed, Figure 1 illustrates this. We compare the demographic, financial-literacy and response time mean results of the

Unreliable and the Experienced subjects by indexing them with the clean (‘Excluding’) responses.

Figure 1 shows that the influence of the significantly lower proportion of the Experienced in full-time employment will be neutralised by the higher proportions of the Unreliable when both are included in the results. A similar net effect is noted for both the level of schooling and the proportions earning less than $75000 p.a. Hence results for ‘Including’ or ‘Excluding’ are virtually identical as the divergent influences cancel one another out.

The lower response times of the Unreliable are also evident in Figure 1.[11] The resulting net effect of including both sets of poor quality data in the ‘Including’ sample is that the influence of the significantly faster response times of the Experienced, is suppressed.

Figure 1: Contrasting the Experienced and the Unreliable

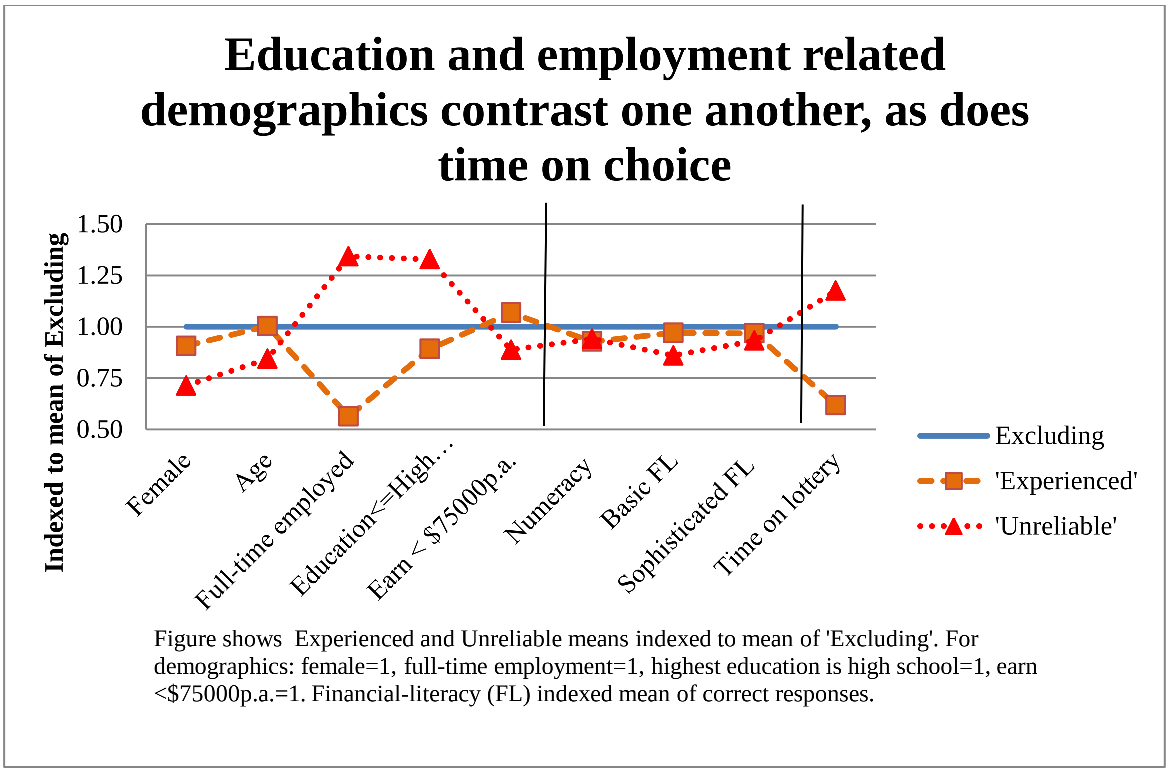


Figure shows Experienced and Unreliable means indexed to mean of 'Excluding'. For demographics: female=1, full-time employment=1, highest education is high school=1, earn <$75000p.a.=1. Financial-literacy (FL) indexed mean of correct responses.

[11] While our expectation is that Spammers will rush through the questionnaire to maximise their rate of pay, the bonus payment was determined by this choice. Consequently, Spammers may find it worthwhile to spend time on this choice while rushing on others.

In the studies reviewed in detail so far, comparing the ‘Including’ with those ‘Excluding’ the unusable responses, showed small differences caused in part by the two types of data suppressing each other’s influence. Another explanation could be that the sample sizes at an overall level are large enough to minimise the influence. Hence we now review results at sub-sample level to examine the influence of the Experienced and the Unreliable on our results when samples are very small.

Results at sub-sample level

Study S5 tested two different choices in the second lottery task between subjects. The sample was also split by level of financial expertise with half of the subjects exposed to the treatment. We further restricted the sample to subjects making an active choice.[12] Thus the sample of 505 subjects resulted in sub-samples ranging from 17 to 42, when all responses were included, or from 14 to 30 when unusable responses were excluded.

Significant moderators of the time spent on the choice varied depending on which responses were included. We used three-way factorial ANCOVA models with age as the covariate. Including unreliable and/or experienced subjects caused heteroscedasticity and distribution problems in the data. Consequently we ran the models on the unweighted dataset and transformed the dependent variable with the Box and Cox (1964) method.[13] Across our four models (‘Excluding all’, ‘Excluding Experienced’, ‘Excluding Unreliable’, ‘Including all’, see **Table 3**), expertise level proved significant. This influence was significantly interacted with the type of choice in all the models which included the Experienced and/or the Unreliable. In contrast, the influence of the treatment became significant only when the Experienced were excluded (the ‘Excluding Experienced’ and ‘Excluding all’ models). While

[12] A default option allowed subjects to passively accept the default rather than make an active choice.

[13] We also ran multiple regression models for the untransformed dependent variable with robust standard errors. These models produced similar results although the extreme skewness of the data still distorted outcomes. The interaction effect between the treatment and expertise was found in the “Excluding All” model once the dependent variable was transformed (even using a log10 rather than a Box-Cox transformation).

the significant influence of the treatment had been shown in our previous research (studies S1, S1a and S1b), the significant interaction of expertise and the treatment shown in the 'Excluding all' model, supported our hypothesis.

Table 3: Estimation results when jointly or separately excluding the Experienced and the Unreliable

Dep. variable: (time^L-1)/L	Excluding All			Excluding Experienced			Excluding Unreliable			Including All		
Source	df	F	Sig.	df	F	Sig.	df	F	Sig.	df	F	Sig.
Corrected Model	(12.0)	3.6	.000	(12.0)	4.7	.000	(12.0)	4.4	.000	(12.0)	5.2	.000
Intercept	(1.0)	1854.4	.000	(1.0)	452.0	.000	(1.0)	462.6	.000	(1.0)	436.8	.000
Age	(1.0)	11.0	.001	(1.0)	18.3	.000	(1.0)	7.7	.006	(1.0)	15.5	.000
Treatment	(1.0)	2.0	.161	(1.0)	4.9	.028	(1.0)	2.1	.144	(1.0)	3.6	.060
Expertise	(2.0)	5.5	.004	(2.0)	7.0	.001	(2.0)	10.8	.000	(2.0)	11.3	.000
Choice	(1.0)	3.6	.057	(1.0)	2.2	.137	(1.0)	.0	.988	(1.0)	.0	.921
Treatment * expertise	(2.0)	4.0	.020	(2.0)	1.2	.318	(2.0)	2.4	.089	(2.0)	1.0	.372
Treatment * choice	(1.0)	.1	.817	(1.0)	1.3	.261	(1.0)	.0	.852	(1.0)	.2	.629
Expertise * choice	(2.0)	2.3	.102	(2.0)	3.6	.029	(2.0)	4.2	.016	(2.0)	4.5	.012
Treatment * expertise * choice	(2.0)	.6	.554	(2.0)	.4	.680	(2.0)	2.3	.102	(2.0)	1.2	.301
Error	(270.0)			(320.0)			(339.0)			(389.0)		
Total	(283.0)			(333.0)			(352.0)			(402.0)		
Corrected Total	(282.0)			(332.0)			(351.0)			(401.0)		
Adj. R^2		0.101			0.104			0.117			0.113	
L (Box & Cox 1964)		-0.147			0.088			0.189			0.120	
Levene's test		0.373			0.057			0.255			0.092	

Table shows results from ANCOVA with age as a covariate. 'Excluding all' excludes the Experienced and the Unreliable; 'Excluding Experienced' includes Unreliable, 'Excluding Unreliable' includes Experienced, and 'Including all' includes both Unreliable and the Experienced. Results run on unweighted data. Time taken by subjects was transformed with (time^L-1)/L to ensure that there was zero skew (Box & Cox 1964). Dark shading highlights significant results where p<0.05, medium shading p<0.10.

(Appendix 4 shows the analysis in more detail.)

The dampening effect of the Experienced on response time is revealed. Excluding the Experienced, even if including the Unreliable, highlights the significance of the treatment for those medium in expertise.

Our results have thus far shown that jointly including the Experienced and the Unreliable sees their influence on results cancelled out, but the separate inclusion of these can alter results. Next we compare our results on MTurk with a laboratory study and observe the effect sizes.

A laboratory and MTurk comparison study

Utilising the same five stages as in S4 and S5, we conducted a study in the UNSW Australia Business School laboratory with 149 student subjects and repeated it on MTurk with 154 US subjects. This study aimed to test the influence of the first lottery task (Eckel and Grossman 2008), termed the prime, and the treatment applied on the time taken to make the risky choice (a modified Gneezy & Potters (1997) choice similar to choice 2 in S5). Both studies were 2 (no treatment; treatment) x 2 ('lottery' prime: no prime) between-subjects' experiments. On average, MTurk subjects earned $1 for their 20 minutes spent ($0.50 basic fee and $0.50 bonus), whereas laboratory subjects earned an average of $11 ($5 show-up fee and $6 bonus) for their 30 minutes spent. Unlike studies S4 and S5, in this study, the financial-literacy questions formed part of stage 5.

Who are the Experienced and the Unreliable?

As shown in **Table 1**, of the 154 MTurk subjects, only 117 proved usable with 14 classified as Experienced and 23 classified as Unreliable. As in studies S4 and S5, the Unreliable and the Experienced were more likely to be male (Gender (F): $M_{Unreliable\&Experienced}$= 0.35, SD = 0.48; $M_{Excluding}$=0.52, SD = 0.502; z = 1.80; p = 0.07) and the Unreliable were significantly younger at 27 years of age (Age $M_{Unreliable}$=26.96, SD = 5.65; $M_{Excluding}$=34.44, SD = 12.04; $t(138) = 2.91$ $p < 0.001$).

How do the Experienced and the Unreliable influence results?

Consistent significant results as seen with study S5 sub-samples are also evident when comparing the MTurk sub-sample results with the laboratory results. **Table 4** compares regression results between the laboratory, MTurk excluding, and MTurk including samples, focusing on the active choosers. The significant influence of the prime in reducing the time taken for the second task is shown in both the laboratory and either of the MTurk results. However, the effect sizes are noticeably different. The effect size for the prime goes from small for the 'MTurk incl.' results, to medium when only the clean MTurk responses ('MTurk excl.') are used.

These changes in the effect sizes are disproportionate to the change in sample size and replicate effect size reductions of Chandler et al. (2015).

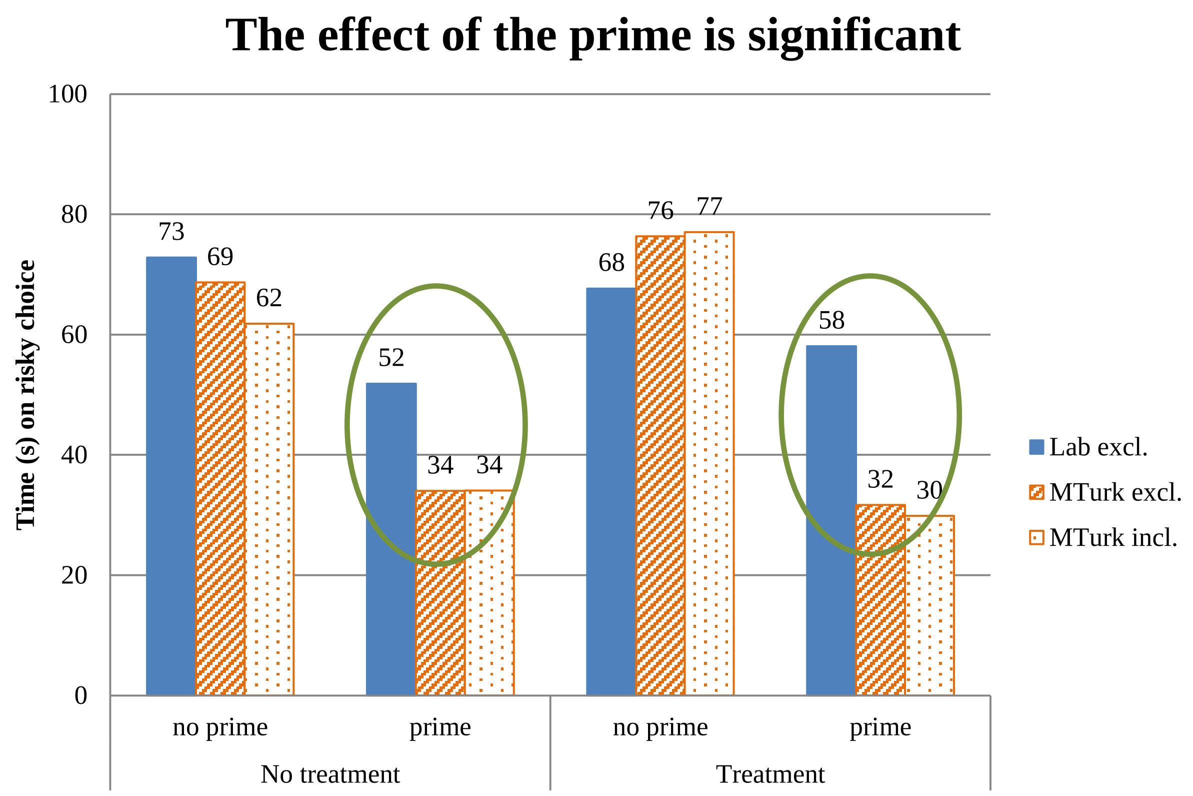
Table 4: Regression results for time on choice

	Lab excl.	**MTurk excl.**	**MTurk incl.**
F	4.32	23.90	14.80
Obs	123	104	135
Adj R-squared	0.076	0.395	0.236
	Coefficient	Coefficient	Coefficient
(time on choice^L-1)/L	(std. err)	(std. err)	(std. err)
	eta-squared	*eta-squared*	*eta-squared*
treatment	-0.837	0.342	0.349
	(1.356)	(0.271)	(0.254)
	0.00	*0.01*	*0.01*
prime	-4.268**	-1.459***	-0.956***
	(1.386)	(0.257)	(0.243)
	0.07	*0.19*	*0.09*
treatment x prime	2.051	-0.335	-0.522
	(1.837)	(0.390)	(0.367)
	0.01	*0.00*	*0.01*
L (Box & Cox 1964)	*0.593*	*0.165*	*0.124*

Table shows regression results when (time on choice^L-1/L) is dependent variable for three models (Box&Cox 1964). Lab. excluding model excludes Unreliable, MTurk excluding model excludes Experienced and Unreliable and MTurk including model includes both. Coefficients, standard error and effect sizes are shown for each of treatment, prime and their interaction.

These results do more than reveal the impact of poor quality data on effect size. As shown graphically in Figure 2, they highlight the difference in the time spent on the tasks whether in the laboratory or when outside of laboratory constraints. Whereas there are only minor differences in time spent on the risky choice between laboratory and MTurk subjects when it is the first task, i.e. without the prime, when it is the second task and there has been an element of learning, there is a significant difference in time spent between laboratory and MTurk subjects ($M_{LabPrimeNotreat}$=51.86, SD = 29.63; $M_{MTurkExclPrimeNotreat}$=34.00, SD = 28.35; *t*(55) = -0.2.32; *p* = 0 .024). Subjects invited to the laboratory may have allowed themselves additional time on the choice given that they had set aside 30 minutes for the experiment. Subjects on MTurk seem to behave more realistically.

Figure 2: Comparing the Laboratory and MTurk on time spent on the risky choice



Graph shows time spent on risky choice with or without prime and/or treatment. 'Excl.' excludes both Experienced and Unreliable for lab and MTurk. 'MTurk incl.' includes both. Note significant time difference between lab and MTurk shown in oval.

Discussion

In this article we focus on the impact of previous exposure to specific tasks as well as unreliable responses on experiment results. We begin by noting that among the 2736 completed responses of our twelve studies, approximately 9% of subjects have previously conducted similar experiments of ours despite our asking that they refrain from this and our subsequent withholding of their bonus[14]. Our studies addressed a particular research question regarding the impact of a treatment on the time taken to make a risky choice. These studies made use of lottery tasks to represent risky choices. While lottery tasks are fairly typical of experiments that try to elicit risk preference (Harrison and Rutström 2008), these types of experiments are perhaps less commonplace than those asking subjects the "Linda problem", or the "Asian disease problem" (Paolacci et al. 2010). On that premise our findings may even under-report the influence of experienced subjects. Similarly since we requested only Turkers with a 99% HIT acceptance rate, we may also be under-reporting the influence of the Unreliable, compared with surveys where the HIT acceptance rate is 95% or even 75%. Notwithstanding our rather high participation hurdle, across all twelve studies, controlling for quotas and software problems, approximately 13% of all subjects offer unreliable responses.

These less reliable subjects are significantly more likely to be young and male. The Experienced are less likely to be in full-time employment, and are significantly more likely to earn less despite being more educated, than those giving unreliable responses. Hence the influence of the Experienced and the Unreliable when included in the overall results is minimised through these divergent differences. The same neutralisation of influence is noted on response times to the choice when both types of subjects are excluded. (In relation to subjective self-report scales, including their responses does not unduly influence the overall

[14] 11% if we focus only on the last three studies discussed in more detail.

results, although we note the importance of including both positively and negatively phrased items in a scale in order to detect unreliable responses.)

At sub-sample level, the influence of the Experienced and the Unreliable seems mitigated by random allocation to conditions and the likelihood of identifying significant differences between sub-samples is inhibited by the very small sample sizes and their large confidence intervals. Nevertheless, excluding the Experienced and the Unreliable doubles the effect sizes from small to medium; an increase disproportionate to their sample size contribution. Furthermore among objective measures we also note disturbing differences.

Subjects who have previously been exposed to the experiment spend 38% less time on a critical task. Unreliable responses also lower the absolute assessment scores on numeracy and basic financial-literacy. While this could be an indication that these subjects are less educated, it is more likely that they have rushed through the experiment by not paying attention, since they typically conduct the entire questionnaire in significantly less time ($t_{S5}(409) = -2.98$; $p = 0.003$; $M_{\text{Excluding}} = 1219$s, SD = 480; $M_{\text{Unreliable}} = 993$s, SD = 794). These influences flow through to the point where including the Experienced mars our hypotheses.

Whereas comparisons between MTurk and laboratory studies have suggested validity (Berinsky et al. 2012), despite the question-marks around some of the experiments compared (see Charness et al. 2010), our results suggests that such validity can only be achieved with proper response management.

Guidelines suggested

We suggest the following guidelines when conducting research on MTurk to minimise the impact of the Experienced and Unreliable responses on results:

- Structure Turker pay to include a bonus:

  - including a financial incentive can reduce the attrition rate.
  - including a financial incentive can also improve performance and reduce risk seeking behaviour (Camerer and Hogarth 1999).
- Add time limited instructions at the start of the experiment to eliminate Spammers or 'bots'. (For example if the instructions take 60 seconds to read, adding a 40 second time limit will frustrate those Spammers rushing through the experiment and they may abandon the experiment).
- Record the Turker id number and IP address to allow elimination of duplicates and to pay bonuses.
- Maintain a master database of Turker identity numbers and IP addresses to identify the Experienced.
- Stringently clean the data using a multi-pronged approach based on total experiment time, time spent on critical tasks, multiple checking questions, scale validation between positively and negatively phrased items, and time spent on scale questions.
- Over-sample to collect the desired quota and to account for losses caused by duplicates and validation.

Conclusions

Across 12 studies from May 2014 to December 2014 our unusable responses grew to represent almost one third of all completed responses increasing our risk of incorrect conclusions[15]. While the proportion of unreliable subjects remained fairly constant, Turkers were obviously becoming experienced with our research. As academics continue to use MTurk and continue to use common tasks like lotteries or the CRT, the extent of this problem will grow. Already 55% of Turkers report that they follow certain requesters aggravating the

[15] TrueSample estimate "If a sample has 10% bad respondents, the increased risk is relatively small; but at 30% the risk is doubled, and at 40% the risk is nearly tripled"(TrueSample 2013)

likelihood of this problem (Chandler et al. 2014). While we are all able to monitor our own experiments, we cannot identify subjects who have done our type of experiment with other academics[16]. The implications of using MTurk and not scrubbing responses of the Experienced and the Unreliable, are concerning. Retaining these responses will materially affect objective measures, lessen the reliability of results and undermine effect sizes.

These findings demand an acceptance that MTurk experiment data needs scrubbing. But, scrubbing of data brings with it the risk of retaining responses more supportive of experiment hypotheses. Consequently acceptance of MTurk data scrubbing to remove the experienced and the unreliable subjects should occur concurrently with agreed minimum standards of data quality control.

References

Amazon Mechanical Turk. (2015). Service Summary. https://requester.mturk.com/tour. Accessed 27 May 2015

Bateman, H., Eckert, C., Geweke, J., Louviere, J. J. J., Thorp, S., & Satchell, S. E. (2012). Financial Competence and Expectations Formation: Evidence from Australia*. *Economic Record*, *88*(280), 39–63. doi:10.1111/j.1475-4932.2011.00766.x

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating Online Labor Markets for Experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*(3), 351–368. doi:10.1093/pan/mpr057

Bohannon, J. (2011). Social Science for Pennies. *Science*, *334*(6054), 307. doi:10.1126/science.334.6054.307

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society, Series B*(26), 211–252.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822–848. doi:10.1037/0022-3514.84.4.822

[16] While this problem may also exist in a laboratory setting, it is manageable through a laboratory user group and a steady turnover of students.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science, 6*(1), 3–5. doi:10.1177/1745691610393980

Camerer, C. F., & Hogarth, R. M. (1999). The Effects of Financial Incentives in Experiments : A Review and Capital-Labor-Production Framework. *Journal of Risk and Uncertainty, 19*(1-3), 7–42.

Camerer, C. F., & Loewenstein, G. (2004). *Behavioral Economics: Past, Present, Future*. Princeton: Princeton University Press.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior research methods, 46*(1), 112–130. doi:10.3758/s13428-013-0365-7

Chandler, J., Paolacci, G., Peer, E., Mueller, P., & Ratliff, K. (2015). Using non-naive participants can reduce effect sizes. *Psychological Science, 26*(7), 1131–1139.

Charness, G., Gneezy, U., & Imas, A. (2013). Experimental methods: Eliciting risk preferences. *Journal of Economic Behavior & Organization, 87*, 43–51. doi:10.1016/j.jebo.2012.12.023

Charness, G., Karni, E., & Levin, D. (2010). On the conjunction fallacy in probability judgment: New experimental evidence regarding Linda. *Games and Economic Behavior, 68*(2), 551–556. doi:10.1016/j.geb.2009.09.003

Chaudhuri, A., & Holbrook, M. B. (2001). The chain of effects from brand trust and brand affect to brand performance: the role of brand loyalty. *Journal of Marketing, 65*(2), 81–93.

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the dictator. *The American Economic Review, 92*(4), 1218–1221.

Cox, D., & Cox, A. D. (2001). Communicating the consequences of early detection: The role of evidence and framing. *Journal of Marketing, 65*(3), 91–103.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE, 8*(3), 1–18.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340.

DeVellis, R. F. (2003). *Scale development : theory and applications* (Second edi.). London: Sage, Thousand Oaks, Calif.

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. doi:10.1016/j.jebo.2008.04.006

Fuller Dynamic. (2012). Distribution of the US population by time zone. http://fullerdynamic.com/news/2012/8/8/distribution-of-us-population-by-time-zone. Accessed 20 May 2015

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

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis: A global perspective* (Seventh.). New Jersey: Pearson Education, Inc.

Harrison, G. W., & Rutström, E. E. (2008). Risk aversion in the laboratory. *Research in Experimental Economics*, *12*, 41–196. doi:10.1016/S0193-2306(08)00003-3

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: a methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*(3), 383–403; discussion 403–51.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425. doi:10.1007/s10683-011-9273-9

Jacquet, J. (2011). The pros & cons of Amazon Mechanical Turk for scientific surveys. *Scientific American*. http://blogs.scientificamerican.com/guilty-planet/2011/07/07/the-pros-cons-of-amazon-mechanical-turk-for-scientific-surveys/. Accessed 5 May 2015

Kaufman, C. F., Lane, P. M., & Lindquist, J. D. (1991). Exploring more than 24 hours a day: A preliminary investigation of polychronic time use. *Journal of Consumer Research*, *18*, 392–401.

Langer, E. J. (1989). *Mindfulness*. Addison-Wesley/Addison Wesley Longman.

Leet, B. (2015). Throw the panel in the river. *research*. http://www.research-live.com/4013335.article?utm_source=google&utm_medium=email&utm_campaign=research-live.com. Accessed 21 May 2015

Lown, J. M. (2011). Development and Validation of a Financial Self-Efficacy Scale. *Journal of Financial Counseling and Planning*, *22*(2), 54–76.

Lusardi, A., & Mitchell, O. S. (2007). Baby Boomer retirement security: The roles of planning, financial literacy, and housing wealth. *Journal of Monetary Economics*, *54*(1), 205–224. doi:10.1016/j.jmoneco.2006.12.001

Lusardi, A., & Mitchell, O. S. (2009). *How ordinary consumers make complex economic decisions: Financial literacy and retirement readiness* (No. w15350). NBER working paper w15350 Cambridge MA.

Marder, J., & Fritz, M. (2015). The Internet's hidden science factory. *PBS Newshour*. http://www.pbs.org/newshour/updates/inside-amazons-hidden-science-factory/#.VN1mmfGgeyY.facebook. Accessed 4 May 2015

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. doi:10.3758/s13428-011-0124-6

Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, *23*(3), 184–188. doi:10.1177/0963721414531598

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419.

Payne, J. W., & Bettman, J. R. (2004). Walking with the Scarecrow: The Information-processing Approach to Decision Research. In *Blackwell Handbook of Judgment and Decision Making* (pp. 110–132).

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The Adaptive Decision Maker*. Cambridge University Press.

Plocher, T., Goonetilleke, R. S., Yan, Z., & Liang, S.-F. M. (2002). Time orientation across cultures. *Proceedings of the 4th …*, 23–31.

Rand, D. G., Peysakhovich, A., Kraft-Todd, G. T., Newman, G. E., Wurzbacher, O., Nowak, M. A., & Greene, J. D. (2014). Social heuristics shape intuitive cooperation. *Nature Communications, 5*(3677), 1–12. doi:10.1038/ncomms4677

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114*(3), 510–532. doi:10.1037/0033-2909.114.3.510

Reilly, M. D. (1982). Working Wives and Convenience Consumption. *Journal of Consumer Research, 8*(4), 407–418. doi:10.1086/208881

Ritchie, T. D., & Bryant, F. B. (2012). Positive state mindfulness: A multidimensional model of mindfulness in relation to positive experience. *International Journal of Wellbeing, 2*(3), 150–181. doi:10.5502/ijw.v2.i3.1

Shim, S., & Drake, M. F. (1988). Apparel selection by employed women: A typology of information search patterns. *Clothing and Textiles Research Journal, 6*(2), 1–9.

Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, *95*(2), 334–344. doi:10.1037/0033-2909.95.2.334

TrueSample. (2013). What Impact Do "Bad" Respondents Have on Business Decisions? http://truesample.com/white-paper/what-impact-do-bad-respondents-have-on-business-decisions/. Accessed 1 June 2015

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*(4481), 453–458. doi:10.1126/science.7455683

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315.

Van Rooij, M. C. J., Lusardi, A., & Alessie, R. J. M. (2011). Financial literacy and stock market participation. *Journal of Financial Economics*, *101*(2), 449–472. doi:10.1016/j.jfineco.2011.03.006

Weinstein, N., Przybylski, A. K., & Ryan, R. M. (2012). The index of autonomous functioning: Development of a scale of human autonomy. *Journal of Research in Personality*, *46*(4), 397–413. doi:10.1016/j.jrp.2012.03.007

Wilcox, N. T. (1993). Lottery choice: Incentives, complexity and decision time. *The Economic Journal*, *103*(421), 1397–1417.

Appendix 1: Financial-literacy Assessment

	Excluding	Experienced	Unreliable	Including	ALP#
Weighted obs	353	78	58	489	989
Numeracy - Sale					
Q17 In a sale, a shop is selling all items at half price. Before the sale a sofa costs $300. How much will it cost in the sale?					
$150	98.2%	90.7%	86.5%	95.6%	
$300	0.0%	0.0%	4.7%	0.6%	
$600	1.3%	6.5%	5.9%	2.7%	
Do not know	0.5%	2.1%	2.8%	1.0%	
Prefer not to answer	0.0%	0.7%	0.0%	0.1%	
Numeracy - 10% Chance					
Q18 If the chance of getting a disease is 10%, how many people out of 1000 would be expected to get the disease?					
10	5.9%	7.4%	2.8%	5.8%	
100	93.6%	88.2%	93.4%	92.7%	
1000	0.0%	2.3%	0.9%	0.5%	
Do not know	0.5%	2.1%	2.8%	1.0%	
Prefer not to answer	0.0%	0.0%	0.0%	0.0%	
Numeracy - Discount					
Q19 A second hand car dealer is selling a car for $6000. This is two-thirds of what it cost new. How much did the car cost new?					
$4000	3.6%	5.3%	5.0%	4.0%	
$6600	0.2%	1.4%	2.8%	0.7%	
$9000	95.1%	88.3%	86.5%	93.0%	
Do not know	1.1%	4.9%	4.7%	2.2%	
Prefer not to answer	0.0%	0.0%	0.9%	0.1%	
Numeracy - Lottery Share					
Q20 If five independent unrelated people all have the winning numbers in the lottery and the prize is $2 million, how much will each of them get?					
$40 000	0.2%	3.0%	1.9%	0.8%	
$400 000	94.5%	83.9%	86.8%	91.9%	
$500 000	3.7%	8.1%	3.8%	4.4%	
Do not know	0.9%	4.9%	7.6%	2.4%	
Prefer not to answer	0.7%	0.0%	0.0%	0.5%	
Numeracy - 1 in 10 Chance					
Q21 If there is a 1 in 10 chance of getting a disease, how many people out of 1000 would be expected to get the disease?					
10	11.5%	14.2%	9.0%	11.6%	
100	86.3%	83.0%	86.3%	85.7%	
1000	2.0%	0.7%	0.9%	1.7%	
Do not know	0.3%	2.1%	3.8%	1.0%	
Prefer not to answer	0.0%	0.0%	0.0%	0.0%	

Table shows column percentages of responses to financial literacy questions taken from S5 weighted to correct for financial expertise quotas. 'Excluding' excludes Experienced and Unreliable subjects. 'Including', includes these. # - ALP results are taken from the on-line USA Rand Life Panel and, unlike our MTurk S5 sample, are weighted to represent the population (Lusardi & Mitchell 2009).

	Excluding	Experienced	Unreliable	Including	ALP[#]
Weighted obs	353	78	58	489	989
Basic financial-literacy - Numeracy					
Q12 Suppose you had $100 in a savings account and the interest rate was 2 percent per year. After 5 years, how much do you think you would have in the account if you left the money to grow?					
More than $102	90.6%	83.2%	75.6%	87.6%	91.8%
Exactly $102	3.7%	7.4%	14.0%	5.5%	
Less than $102	3.4%	4.4%	2.8%	3.5%	
Do not know	2.0%	4.9%	6.6%	3.0%	1.0%
Prefer not to answer	0.3%	0.0%	0.9%	0.3%	0.4%
Basic financial-literacy - Compound interest					
Q8 Suppose you had $100 in a savings account and the interest rate is 20 per cent per year and you never withdraw money or interest payments. After 5 years, how much would you have on this account in total?					
More than $200	59.4%	66.2%	56.9%	60.2%	69.0%
Exactly $200	27.2%	21.4%	25.6%	26.1%	
Less than $200	9.6%	7.4%	10.9%	9.4%	
Do not know	3.8%	5.1%	5.7%	4.2%	1.9%
Prefer not to answer	0.0%	0.0%	0.9%	0.1%	0.0%
Basic financial-literacy -Inflation					
Q9 Imagine that the interest rate on your savings account was 1 per cent per year and inflation was 2 per cent per year. After 1 year, how much would you be able to buy with the money in this account?					
More than today	3.8%	5.8%	14.0%	5.3%	
Exactly the same	5.0%	10.6%	7.6%	6.2%	
Less than today	84.5%	75.8%	66.9%	81.0%	87.1%
Do not know	6.6%	7.1%	11.6%	7.2%	4.1%
Prefer not to answer	0.2%	0.7%	0.0%	0.2%	0.1%
Basic financial-literacy -Time Value of Money					
Q15 Assume a friend inherits $10,000 today and his sibling inherits $10,000 three years from now. In 3 years, who is richer because of the inheritance?					
My friend	42.9%	44.7%	36.4%	42.4%	73.8%
His sibling	11.5%	10.3%	14.6%	11.7%	
They are equally rich	31.6%	35.5%	39.3%	33.1%	
Do not know	13.8%	9.5%	9.6%	12.6%	6.6%
Prefer not to answer	0.1%	0.0%	0.0%	0.1%	0.0%
Basic financial-literacy - Money Illusion					
Q11 Suppose that in the year 2020, your income has doubled and prices of all goods have doubled too. In 2020, how much will you be able to buy with your income?					
More than today	1%	1%	3%	1%	
Exactly the same	93.3%	90.5%	82.9%	91.6%	78.4%
Less than today	4.6%	6.0%	7.7%	5.2%	
Do not know	0.7%	2.8%	5.7%	1.6%	1.2%
Prefer not to answer	0.0%	0.0%	0.9%	0.1%	0.1%

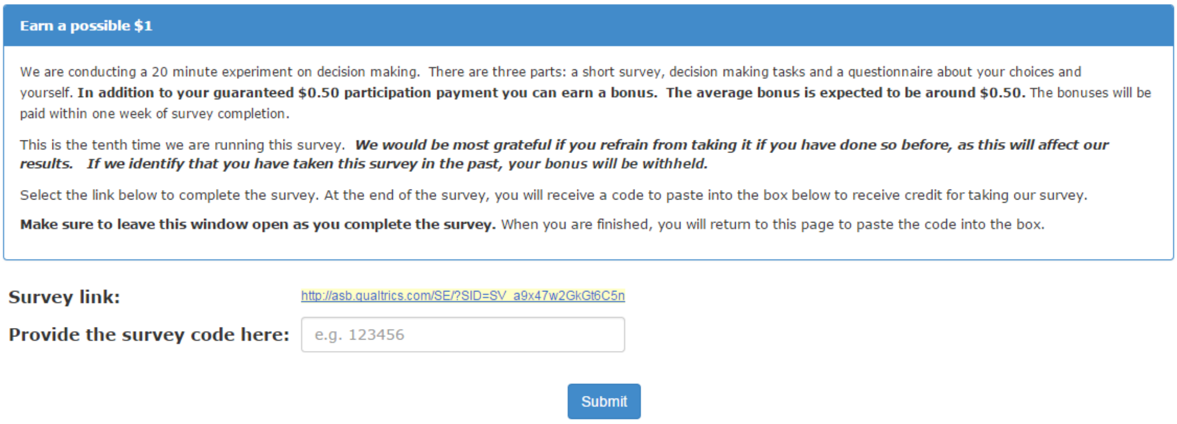
Table shows column percentages of responses to financial literacy questions taken from S5 weighted to correct for financial expertise quotas. 'Excluding' excludes Experienced and Unreliable subjects. 'Including', includes these. # - ALP results are taken from the on-line USA Rand Life Panel and, unlike our MTurk S5 sample, are weighted to represent the population (Lusardi & Mitchell 2009).

	Excluding	Experienced	Un-reliable	Including	ALP#
Weighted obs	353	78	58	489	989
Sophisticated financial-literacy - Risky assets					
Q14 Is the following statement true or false? Shares are normally riskier than bonds					
True	80.3%	76.5%	75.0%	79.1%	80.2%
False	3.0%	5.8%	10.6%	4.4%	
Do not know	16.4%	17.7%	13.5%	16.2%	14.4%
Prefer not to answer	0.3%	0.0%	0.9%	0.3%	0.1%
Sophisticated financial-literacy - Volatility					
Q16 Normally, which asset displays the highest fluctuations over time?					
Bonds	3.2%	5.3%	4.7%	3.7%	
Savings accounts	1.1%	1.4%	1.9%	1.3%	
Shares	82.6%	78.1%	80.2%	81.6%	88.3%
Do not know	13.0%	15.2%	13.2%	13.3%	7.1%
Prefer not to answer	0.2%	0.0%	0.0%	0.1%	0.0%
Sophisticated financial-literacy - Risk diversification					
Q13 When an investor spreads his money among different assets, does the risk of losing money:					
Increase	8%	14%	15%	10%	
Decrease	77.5%	70.5%	65.7%	75.0%	74.9%
Stay the same	6.8%	10.4%	9.7%	7.7%	
Do not know	7.5%	4.9%	9.7%	7.4%	6.7%
Prefer not to answer	0.3%	0.0%	0.0%	0.2%	0.1%
Sophisticated financial-literacy - Long Period Returns					
Q10 Considering a long time period (e.g. 10 or 20 years), which asset normally gives the highest return?					
Bonds	37.4%	23.4%	29.2%	34.2%	
Savings accounts	7.1%	2.8%	4.7%	6.2%	
Shares	38.4%	44.8%	39.6%	39.5%	
Do not know	17.1%	29.0%	26.5%	20.1%	
Prefer not to answer	0.0%	0.0%	0.0%	0.0%	

Table shows column percentages of responses to financial literacy questions taken from S5 weighted to correct for financial expertise quotas. 'Excluding' excludes Experienced and Unreliable subjects. 'Including', includes these. # - ALP results are taken from the on-line USA Rand Life Panel and, unlike our MTurk S5 sample, are weighted to represent the population (Lusardi & Mitchell 2009).

Appendix 2: HIT Screen

Below is a screenshot of the recruitment page displayed on MTurk to Turkers. Note that we inform Turkers of the number of times we have run the survey as well as our intention to withhold their bonus if they participate against our wishes.



Earn a possible $1

We are conducting a 20 minute experiment on decision making. There are three parts: a short survey, decision making tasks and a questionnaire about your choices and yourself. **In addition to your guaranteed $0.50 participation payment you can earn a bonus. The average bonus is expected to be around $0.50.** The bonuses will be paid within one week of survey completion.

This is the tenth time we are running this survey. ***We would be most grateful if you refrain from taking it if you have done so before, as this will affect our results. If we identify that you have taken this survey in the past, your bonus will be withheld.***

Select the link below to complete the survey. At the end of the survey, you will receive a code to paste into the box below to receive credit for taking our survey.

Make sure to leave this window open as you complete the survey. When you are finished, you will return to this page to paste the code into the box.

Survey link: http://asb.qualtrics.com/SE/?SID=SV_a9x47w2GkGt6C5n

Provide the survey code here: e.g. 123456

Appendix 3: Identifying Unreliable Responses

An important part of our research was our multi-pronged approach to identifying the Unreliable. This involved flagging subjects for each violation of our pre-determined rules. Here we describe this approach in more detail.

Flagging based on question answers

Subjects were shown one of the financial-literacy questions at the start of the experiment and again near the end, and were asked whether they recognise it. If not, the record was flagged. If they did recognise the question but answered inconsistently, the record was also flagged. Subjects' treatment of a negatively phrased question was compared with their answer to a similar positively phrased question on the same scale. Where subjects had not selected the midpoint yet showed little difference, they were flagged. On another scale, the answers to two extremely similar questions were compared. If these were widely disparate, i.e. three or more scale points apart, the record was flagged.

On scales where between eight and fifteen items were measured, responses were flagged if the same answer was selected throughout.

If a subject was flagged three times, the maximum possible number, for haphazardness the response was deemed unreliable.

Flagging based on timing

Subjects who completed scale questions with eight to fifteen items in less than 20 seconds (faster than the questions could be read) were also flagged and deemed unreliable.

If subjects completed the entire questionnaire in less than 7 minutes, the record was flagged as unreliable (the average time for questionnaire completion is 20 minutes and arguably the questionnaire cannot be read thoroughly in 7 minutes). Similarly, if subjects completed the first ‘lottery’ task in less than ten seconds, the record was flagged as unreliable.

If subjects completed the second lottery task in less than 10 seconds the record was only deemed unreliable if the subject had also been flagged as haphazardly answering questions.

Subjects taking inordinately long on either choice e.g. more than 150 seconds were also flagged but only deemed unreliable if more than three standard deviations away from the mean (around 1% of collected responses).

An example of one of the spreadsheets used to identify the Unreliable, taken from study S5, is shown in **Table A3.1**.

Table A3.1: Example of a spreadsheet used to identify the Unreliable (study S5)

Quest id	q49==2	q487_7> q487_8 (diff 3 plus)	q487_9== q487_11 (diff==0)	q496_7> q496_8 (diff 3 plus)	q496_9==q496_11 (diff==0)	q48<>q8	Poor completion	Inattentive Score	Lottery time	Choice 1 time	Choice 2 time	Total Duration	Unreliable
a	**b**	**c**	**d**	**e**	**f**	**g**	**h**	**i**	**j**	**k**	**l**	**m**	**n**
92	92					92	1	2				458	1
119	119							1		3.515			1
129	129							1			9.619		1
185	185							1			5.205		1
213	213			213				2			8.779		1
301	301							1		9.026			1
361	361						1	1		9.176		434	1
370					370			1			9.762		1
379	379							1		9.128			1
380					380	380	1	2	3.771		2.458	320	1
449	449							1		9.798			1
509						509		1			5.143		1
578	578					578		2			6.386		1
621	621							1				467	1
636	636						1	1			8.24	457	1

Table shows an example spreadsheet used to identify Unreliable subjects. Columns b to g identify subjects who have been flagged on validation questions. ‘Poor completion’ flags subjects for poor scale completion identified in the database of responses. ‘Inattentive score’ sums flags in columns b to g. Extreme response times to risky choices are recorded in columns j to l. Extremes for total duration of survey are recorded in column m. Subjects tagged as Unreliable are recorded in column n.

Appendix 4: ANCOVA Results of Study S5

Table A4.1 shows the ANCOVA results for study S5. We tested the interaction effect of the treatment, choice and level of expertise on the time taken to make the risky choice by active choosers. As would be expected, including the unreliable and the experienced created problems with the homoscedasticity and the likelihood of a normal distribution in all datasets excepting ‘Excluding all’. Hence all results were run in Stata version 13.1 on unweighted data. The time taken by subjects was transformed with the (timechoice^L-1)/L, where L is shown for each of the models in the following table. to ensure that there was zero skew (Box and Cox 1964). We repeated the analyses on the untransformed dependent variable, time, using multiple regressions with robust standard deviations. These models produced similar results although the extreme skewness of the data still distorted outcomes. The three way interaction between the treatment, choice and expertise on effort is not significant even if excluding the unusable responses. However, the interaction effect between the treatment and expertise was found in the “Excluding All” model once the dependent variable was transformed. We also ran the ANCOVA models with a log 10 transformation and found this interaction effect to be present.

All results were run in four ways:

- ‘Excluding all’ i.e. excluding the Experienced and the Unreliable;
- ‘Excluding Unreliable’ and thus including Experienced,
- ‘Excluding Experienced’ and thus including Unreliable, and
- ‘Including all’ including the Unreliable and the Experienced.

Table A4.1: ANCOVA results of study S5

	Excluding all					Excluding Experienced					Excluding Unreliable					Including all				
Dep. variable: (time^L-1)/L	SS	df	MS	F	Sig	SS	df	MS	F	Sig	SS	df	MS	F	Sig	SS	df	MS	F	Sig
Corrected Model	4.3	(12)	.4	3.6	.00	44.9	(12)	3.7	4.7	.00	73.4	(12)	6.1	4.4	.00	71.2	(12)	5.9	5.2	.00
Intercept	181.4	(1)	181.4	1854.4	.00	361.9	(1)	361.9	452.0	.00	644.3	(1)	644.3	462.6	.00	493.6	(1)	493.6	436.8	.00
Age	1.1	(1)	1.1	11.0	.00	14.7	(1)	14.7	18.3	.00	10.7	(1)	10.7	7.7	.01	17.5	(1)	17.5	15.5	.00
Treatment	.2	(1)	.2	2.0	.16	3.9	(1)	3.9	4.9	.03	3.0	(1)	3.0	2.1	.14	4.0	(1)	4.0	3.6	.06
Expertise	1.1	(2)	.5	5.5	.00	11.3	(2)	5.6	7.0	.00	30.1	(2)	15.1	10.8	.00	25.5	(2)	12.8	11.3	.00
Choice	.4	(1)	.4	3.6	.06	1.8	(1)	1.8	2.2	.14	.0	(1)	.0	.0	.99	.0	(1)	.0	.0	.92
Treatment* expertise	.8	(2)	.4	4.0	.02	1.8	(2)	.9	1.2	.32	6.8	(2)	3.4	2.4	.09	2.2	(2)	1.1	1.0	.37
Treatment * choice	.0	(1)	.0	.1	.82	1.0	(1)	1.0	1.3	.26	.0	(1)	.0	.0	.85	.3	(1)	.3	.2	.63
Expertise * choice	.5	(2)	.2	2.3	.10	5.7	(2)	2.9	3.6	.03	11.7	(2)	5.8	4.2	.02	10.1	(2)	5.1	4.5	.01
Treatment * expertise * choice	.1	(2)	.1	.6	.55	.6	(2)	.3	.4	.68	6.4	(2)	3.2	2.3	.10	2.7	(2)	1.4	1.2	.30
Error	26.4	(270)	.1			256.2	(320)	.8			472.2	(339)	1.4			439.6	(389)	1.1		
Total	2078.6	(283)				5370.7	(333)				8237.5	(352)				6958.7	(402)			
Corrected Total	30.7	(282)				301.1	(332)				545.5	(351)				510.7	(401)			
	Mean Diff.	Std. Error		F	Sig	Mean Diff.	Std. Error		F	Sig	Mean Diff.	Std. Error		F	Sig	Mean Diff.	Std. Error		F	Sig
Expertise low *treatment	.104	.070		2.2	.14	.000	.178		.0	.96	.218	.233		.9	.35	.012	.192		.0	.95
Expertise med *treatment	-.135	.066		4.1	.04	-.371	.180		4.2	.04	-.362	.230		2.5	.12	-.315	.198		2.5	.11
Expertise high *treatment	-.132	.063		4.4	.04	-.296	.171		3.0	.09	-.426	.210		4.1	.04	-.320	.182		3.1	.08
Levene's test for homoscedasticity				1.085	.373				1.774	.057				1.246	.255				1.614	.092
Adjusted r squared					.101					.117					.104					.113

Table shows results from ANCOVA run on each choice with age as a covariate. Excluding all' excludes the Experienced and the Unreliable; 'Excluding Experienced' includes Unreliable, 'Excluding Unreliable' includes Experienced, and 'Including all' includes both Unreliable and the Experienced. Results run on unweighted data. Time taken by subjects was transformed with (time^L-1)/L to ensure that there was zero skew (Box & Cox 1964). Dark shading highlights significant results where p<0.05, medium shading p<0.10.

Appendix 5: Attitudinal Scale Analysis

In stage 5 of our studies, subjects completed fourteen attitudinal scales taken mainly from psychology and marketing literature. We expected scale reliability to be compromised by the separate or joint exclusion of the Experienced and the Unreliable, but found few consistent differences as discussed here.

Measures

All but two of the attitudinal scales required seven point Likert agreement ratings (see **Table A5.1**). These attitudinal scales varied in terms of the number of items used to measure the attitude and in terms of whether the scale used only positively phrased items or also included negatively phrased items. (When developing attitudinal scales, it is common practice to include negatively phrased items in addition to positively phrased items. This can make it more difficult for subjects to complete the scale, but is intended to minimise subjects' agreeing to all statements (DeVellis 2003)).

We wanted to measure subject's perception of the choices in terms of how risky they thought each one was, how easy each was to make, and how confident subjects felt making the choice. To assess the perceived risk of each of choice 1 and choice 2, we used the scale developed by Cox and Cox (2001) with five negatively phrased items. We measured the perceived level of cognitive effort required for each of the two choices using a modified version of Davis (1989) with two of the three items negatively phrased. Confidence in each of the choices was measured using the two positively phrased and one negatively phrased item scale of Shim and Drake (1988). We also wanted to assess the perceived risk of the treatment and whether subjects were likely to trust the treatment. We again used the Cox and Cox (2001) scale to measure the perceived risk of the treatment, while trust in the treatment used the scale developed by Chaudhuri and Holbrook (2001) with four positively phrased items.

While subjects' attitudes towards the choices and treatment are of interest in terms of their reliability among the Experienced and the Unreliable, the following scales were used to gain more insight into the subjects themselves. These also shed light on differences in mindset between the Experienced and Unreliable. Perceived risk in choosing investments generally was assessed with the Cox and Cox (2001) scale. We measured the extent to which subjects are Mindful, or show a predisposition to pay attention to the novel (Brown and Ryan 2003; Langer 1989) using the Mindfulness Trait scale of fourteen negatively phrased items of Brown and Ryan (2003). We also measured the extent to which the investment choices made subjects more mindful using the eight positively phrased items of the Mindfulness State scale (Ritchie and Bryant 2012). We measured whether subjects considered themselves time-poor with the modified five item positively phrased Time Poverty scale of Reilly (1982). Polychronic attitude to time use was measured with three positively phrased items (Kaufman et al. 1991; Plocher et al. 2002). While the financial-literacy assessment of stage 1 objectively assessed subjects' financial-literacy, we were also interested in their financial self-efficacy perceptions and used six positively phrased items (Lown 2011) for this.

Scale accuracy

Here we examine the reliability of the scaled instruments included in the experiments by comparing the Cronbach alpha score across the subsets. (Cronbach alpha is a measure of scale reliability based on the inter-relatedness of scale items (DeVellis 2003), i.e. to what extent are all items measuring the same thing.) We look first at study S5, before comparing the scale results of MTurk and the laboratory.

Table A5.1 compares Cronbach alpha scores taken from study S5.[17] When there are positively phrased and negatively phrased items in the scale, the alpha scores for the scales are noticeably lower. This occurs not only among the Unreliable but also among the

[17] We examine the scales on S5 rather than S4 due to S5's larger sample sizes.

Experienced subjects. In terms of the recommendation that the acceptable Cronbach alpha of the scale be above 0.700 (Hair et al. 2010), only one of the scales is adversely affected.

Table A5.1: Comparison of Cronbach alpha for each scale

			Study S5			
			Excluding	**Experienced**	**Unreliable**	**Including**
Unweighted obs			354	85	66	505
Weighted obs			353	78	58	489
Scales (positively +negatively phrased items)						
Perceived risk choice 1	(0+5)	7pt	0.719	0.735	0.757	0.743
Perceived risk choice 2	(0+5)	7pt	0.744	0.697	0.653	0.727
Cognitive effort choice 1	(1+2)	7pt	0.730	0.677	0.752	0.726
Cognitive effort choice 2	(1+2)	7pt	0.728	0.351	0.247	0.658
Confidence choice 1	(2+1)	7pt	0.751	0.479	0.721	0.723
Confidence choice 2	(2+1)	7pt	0.748	0.472	0.737	0.716
Perceived risk treatment	(0+5)	7pt	0.771	0.693	0.826	0.773
Trust treatment	(4+0)	7pt	0.789	0.840	0.840	0.808
Perceived risk investments	(0+5)	7pt	0.758	0.659	0.706	0.745
Mindfulness state	(8+0)	7pt	0.821	0.797	0.763	0.813
Mindfulness trait	(0+14)	6pt	0.914	0.922	0.915	0.916
Time poverty	(5+0)	7pt	0.881	0.862	0.820	0.872
Polychronic time use	(3+0)	7pt	0.792	0.755	0.860	0.795
Financial efficacy	(6+0)	4pt	0.812	0.772	0.808	0.804

Table shows attitudinal scales used and the Cronbach alpha reliability measure of each scale for each of the Experienced and Unreliable as well as when these responses are excluded, 'Excluding' or included, 'Including'. The number of positively and negatively phrased items is shown in brackets after each scale's name. The number of points on the Likert agreement scale is also shown for each scale. Where the Cronbach alpha is below the recommended 0.700 (Hair et al. 2010) it is highlighted.

Figure A5.1 indexes the mean self-ratings on these scales of the Experienced and the Unreliable with the mean of the clean responses. As with some of the demographics and the response time measures, we note that differences in perceptions of the choices, and to a far lesser extent, the treatment are often cancelled out between the Experienced and the Unreliable. In terms of subjects’ self-perceptions, despite their unreliable responses generally, the Unreliable can be differentiated from the Experienced by their propensity to see themselves as time-poor.

Figure A5.1: Comparison of Scale Ratings

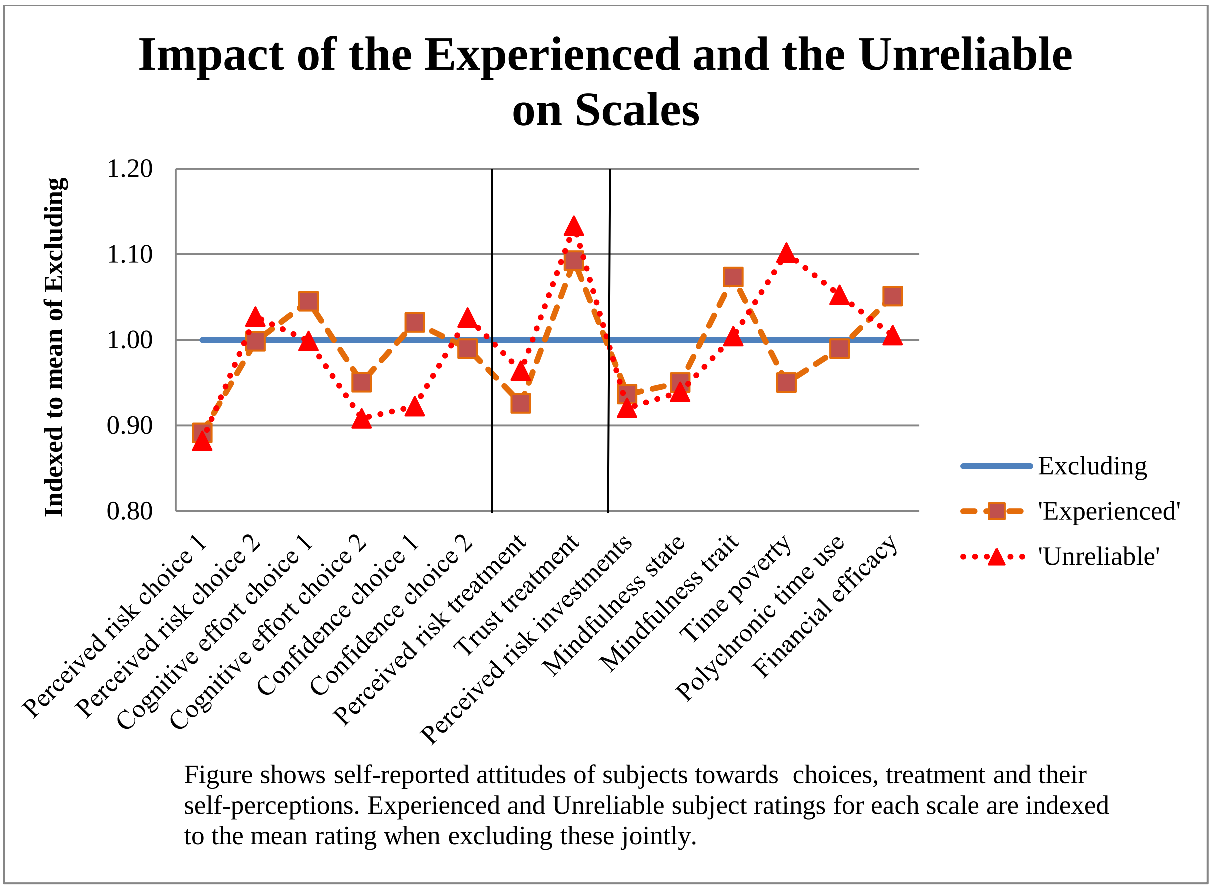


Figure shows self-reported attitudes of subjects towards choices, treatment and their self-perceptions. Experienced and Unreliable subject ratings for each scale are indexed to the mean rating when excluding these jointly.

Next we compare the alpha scores taken from the laboratory study with those taken from the MTurk study. (In addition to the scales used in studies S4 and S5, the Index of Autonomous Functioning (IAF) scale of Weinstein et al. (2012) was included to determine whether subjects with differing perceptions of autonomy responded differently to the treatment.)

Figure A5.2: Comparison of Cronbach alpha

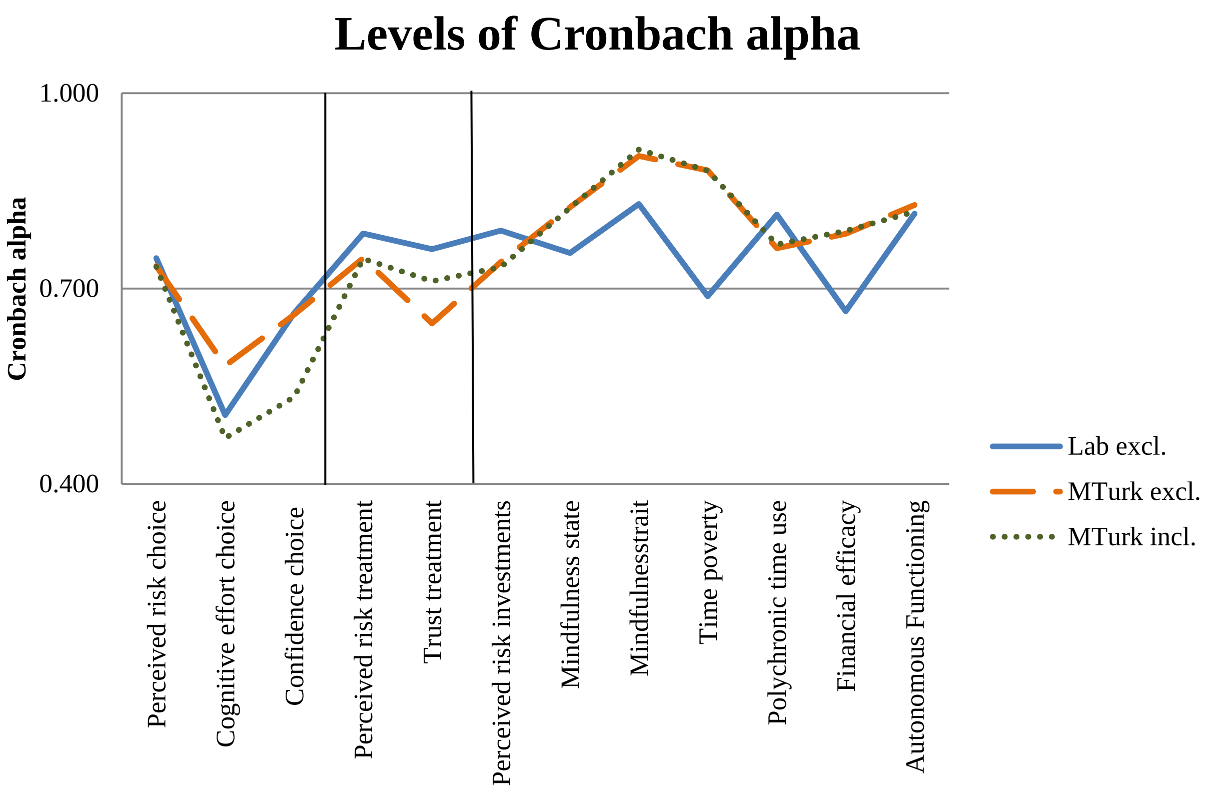


Figure shows Cronbach alpha for scales taken from Lab or MTurk results. Lab and MTurk results excluding Experienced and Unreliable are fairly similar except for self-perception scales, but are all still better than or near 0.700 (Hair et al.,2010).

Unlike the MTurk subjects, laboratory subjects had time restrictions imposed on the scales to discourage them from rushing through the scales. This was not done on MTurk since imposing forced time constraints would remove the ability to identify unreliable responses post-hoc and could also increase the attrition rate. Despite this difference as shown in Figure A5.2, the reliability of the scales corresponds favourably, with the difference between the laboratory and MTurk results on the self-perception scales in the third panel of lesser importance since the scales are still above or very close to the 0.700 recommended cut-off level (Hair et al. 2010).