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

Jeanette Deetlefs

M. Chylinski, A. Ortmann

Motivation Research Results Discussion

Amazon’s Mechanical Turk

✓ Low-cost

✓ Fast turnaround

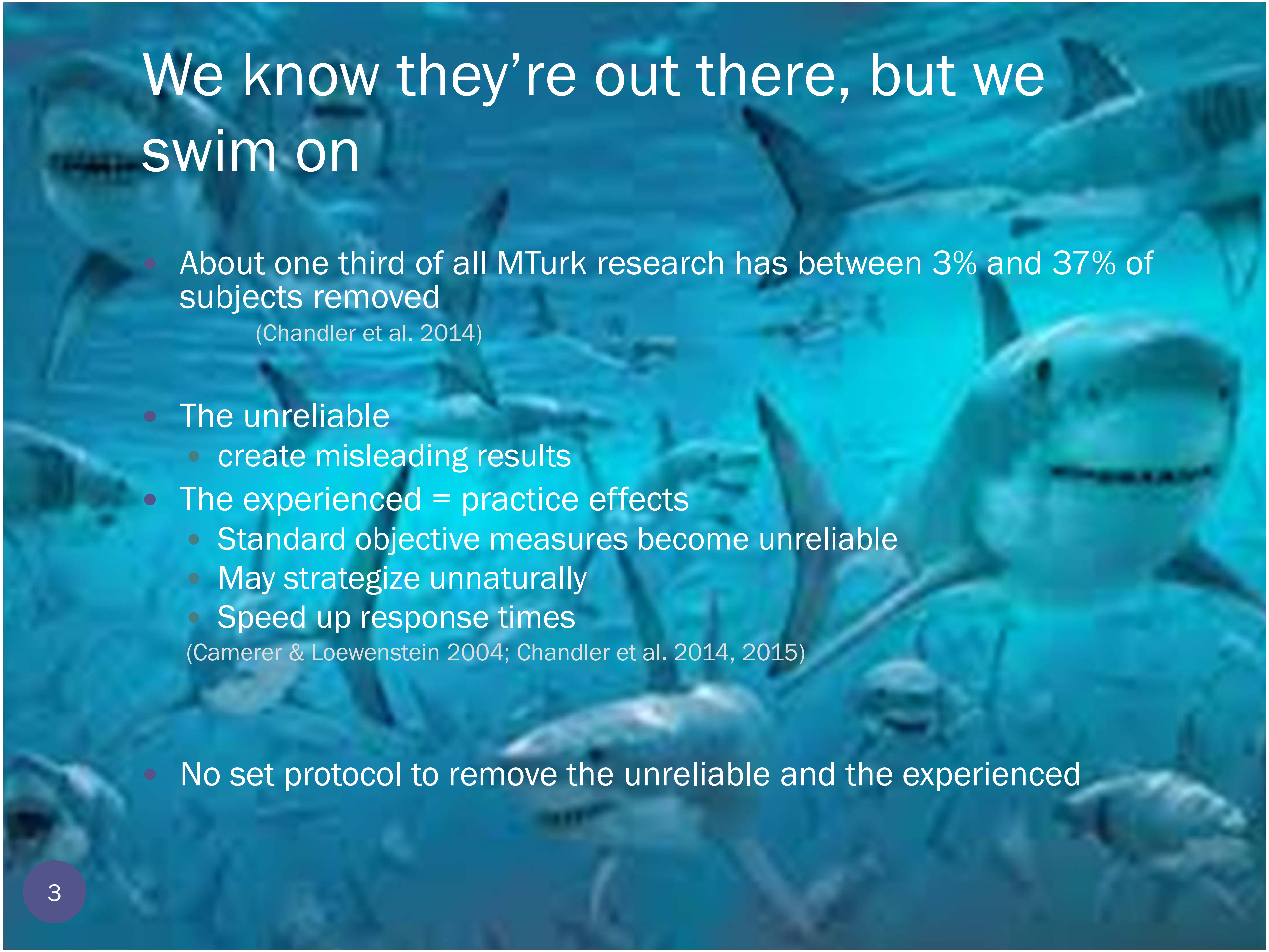
✓ Acceptable validity

But….

✗ Super-Turkers (the experienced)

&

✗ Spammers (the unreliable)



We know they're out there, but we swim on

- About one third of all MTurk research has between 3% and 37% of subjects removed

  (Chandler et al. 2014)

- The unreliable
  - create misleading results
- The experienced = practice effects
  - Standard objective measures become unreliable
  - May strategize unnaturally
  - Speed up response times

  (Camerer & Loewenstein 2004; Chandler et al. 2014, 2015)

- No set protocol to remove the unreliable and the experienced

Our research…

- 12 studies with 2736 subjects
  - 9% are experienced with our risk-type experiment (Super-Turkers)
  - 11% are unreliable (Spammers) with faster response times and poorer completion
- Detailed analysis at overall (n=505) and sub-sample level (n=17 to n=42)
- Comparison of a Bizlab (n=149) and MTurk (n=154) study

What we found…

- Objective measures are most influenced e.g.,
  - the experienced have response times that are 38% faster
  - the unreliable score 10% lower on financial literacy measures

What we found…

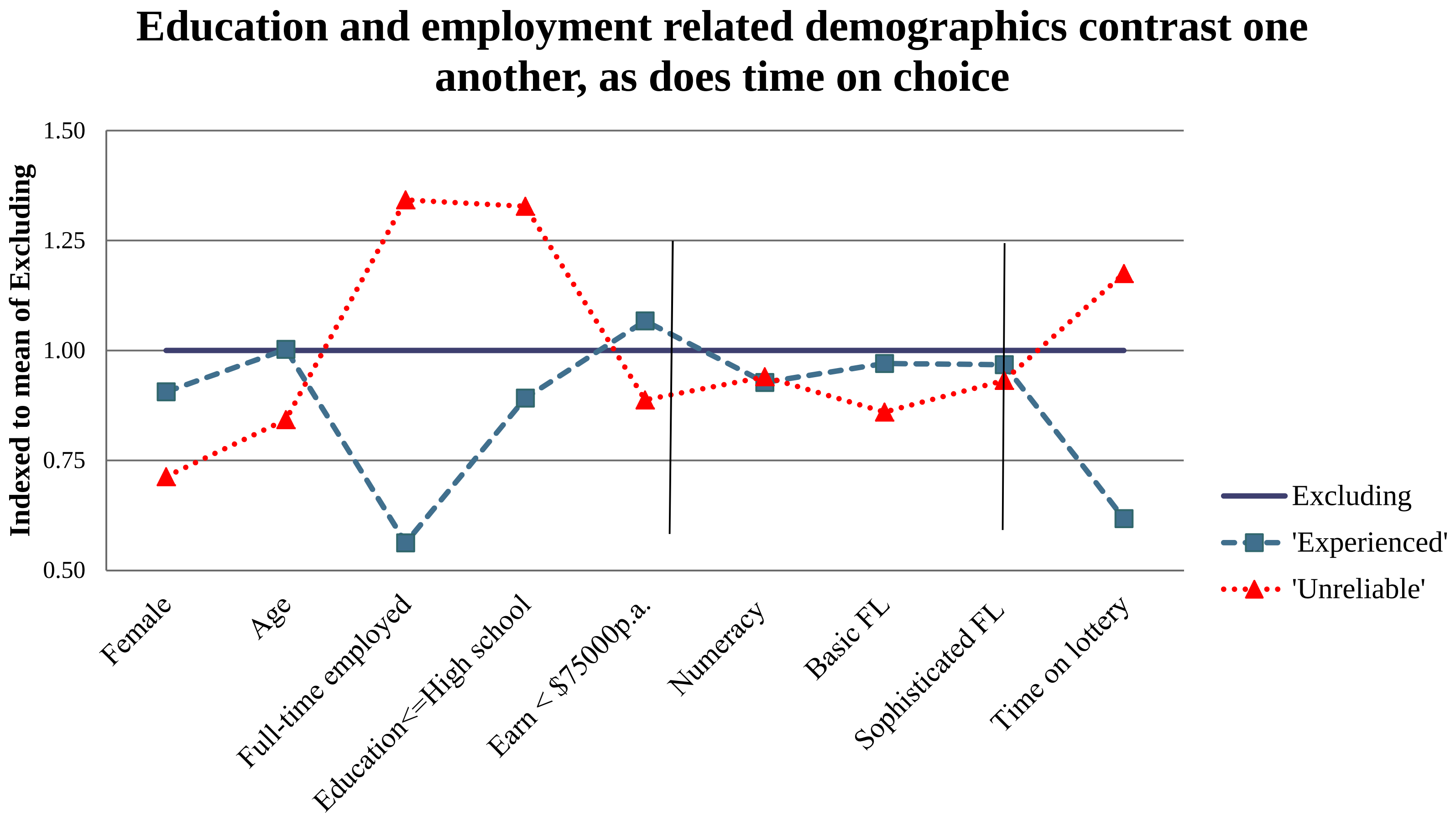


Figure shows Experienced and Unreliable means indexed to mean of 'Excluding'. For demographics: female=1, full-time employment=1, highest education is high school=1, earn <$75000p.a.=1. Financial-literacy (FL) indexed mean of correct responses.

What we found ctd…

- Objective measures are most influenced e.g.,
  - the experienced have response times that are 38% faster
  - the unreliable score 10% lower on financial literacy measures

- Little difference in outcomes when both are included

BUT …

- Exclusion doubles our effect sizes

	MTurk excl.	**MTurk incl.**
F	23.90	14.80
Obs	104	135
Adj R-squared	0.395	0.236
	Coefficient	Coefficient
(time on choice^L-1)/L	(std. err)	(std. err)
	eta-squared	*eta-squared*
treatment	0.342	0.349
	(0.271)	(0.254)
	0.01	*0.01*
prime	-1.459***	-0.956***
	(0.257)	(0.243)
	0.19	*0.09*
treatment x prime	-0.335	-0.522
	[illegible]	[illegible]

Implications

- The problem is probably larger than we found
  - Our participation hurdle was high
    - 99% acceptance rate for Turkers
    - Not rewarded if participated more than once
  - Lotteries are possibly less common
- This problem will grow
  - Academic preference for the tried and tested
  - No way to track subjects collectively
  - 55% of Turkers report that they follow particular Requesters (Chandler et al. 2014)

Staying safe…

Include a bonus

Add time-limited instructions at the start of the experiment to eliminate Spammers or ‘bots’

Record the Turker id number and IP address

Maintain a master database of Turker identity numbers and IP addresses

Stringently clean the data using a multi-pronged approach

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.

Over-sample

Thank you – Questions?