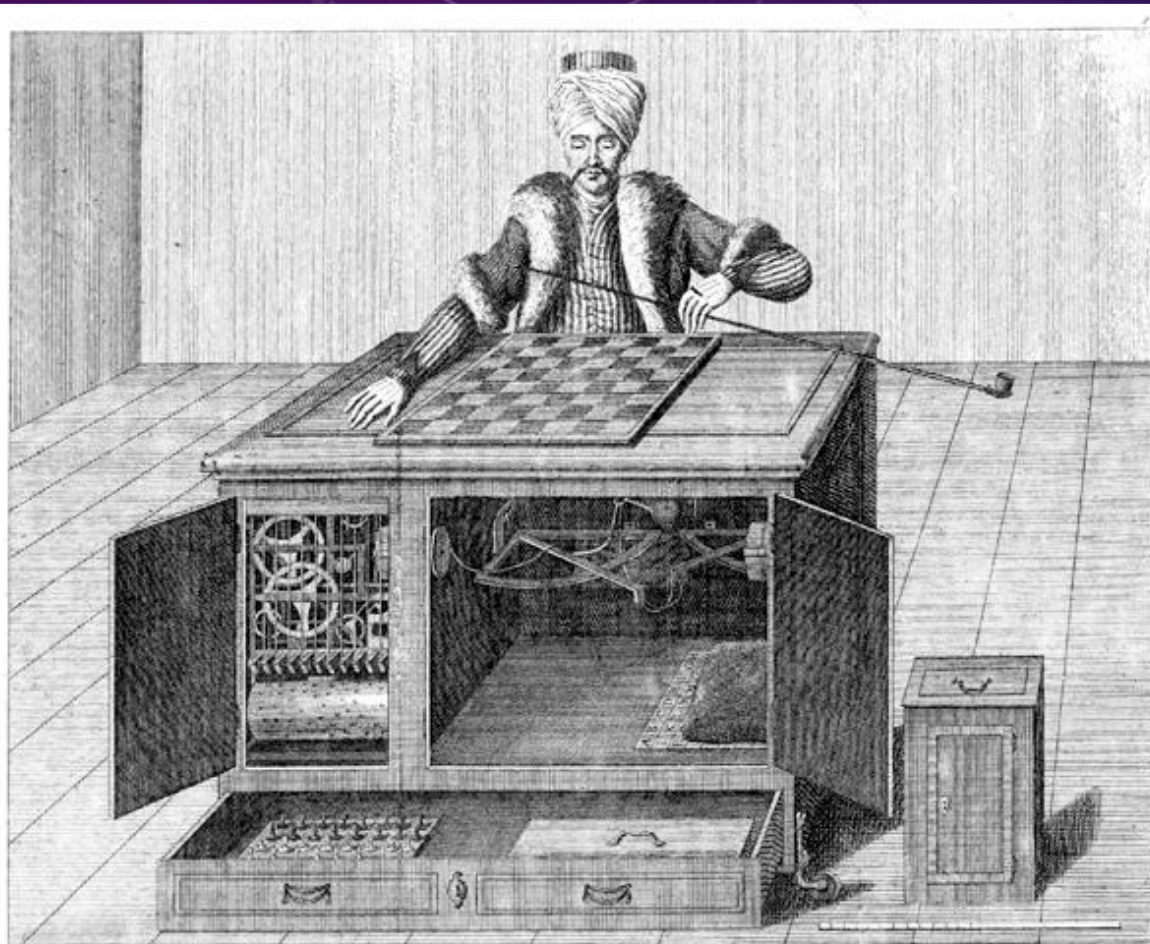


amazonmechanical turk

Artificial Artificial Intelligence



W. de Kempelen del. Che a Michel escud. Basilea. P. G. Ritz, fecit.
Der Schachspieler wie er vor dem Spiele gezeiget wird von vorn. Le Joueur d'Échecs, tel qu'on le montre avant le jeu par devant.

Outline

- Background
- Terms
- Demographics
- Workers perspective (incentives)
- Price and time
- Filtering
- Worker-Requester relationship
- Testing – Familiar problems
- Special constellations
- Pros and Cons
- Tips
- Future applications
- References

Background

- Created at 2005
- One of many...
- Population
 - Jan, 2007 – 100,000 from 100 countries
 - Jan, 2011 – 500,000 from 190 countries
 - Jan, 2015 – >1,000,000...?
- Uses
 - Capcha solving
 - Translation / Transcription
 - Picture classification
 - Social science...



Terms

- Requester, Worker
- HIT – human intelligence task
- Bonus
- Qualification
- Rejection / Success rate
- ACQ – attention check question
- API – application programming interface
 - ('artificial artificial intelligence')
- Bot
 - ('artificial artificial artificial intelligence')
- Sweatshops

How does it look for workers

Amazon Mechanical Turk - All HITs

93,352 HITs available now

Search for **HITs** containing [] that pay at least \$ **0.00** for which you are qualified **GO**

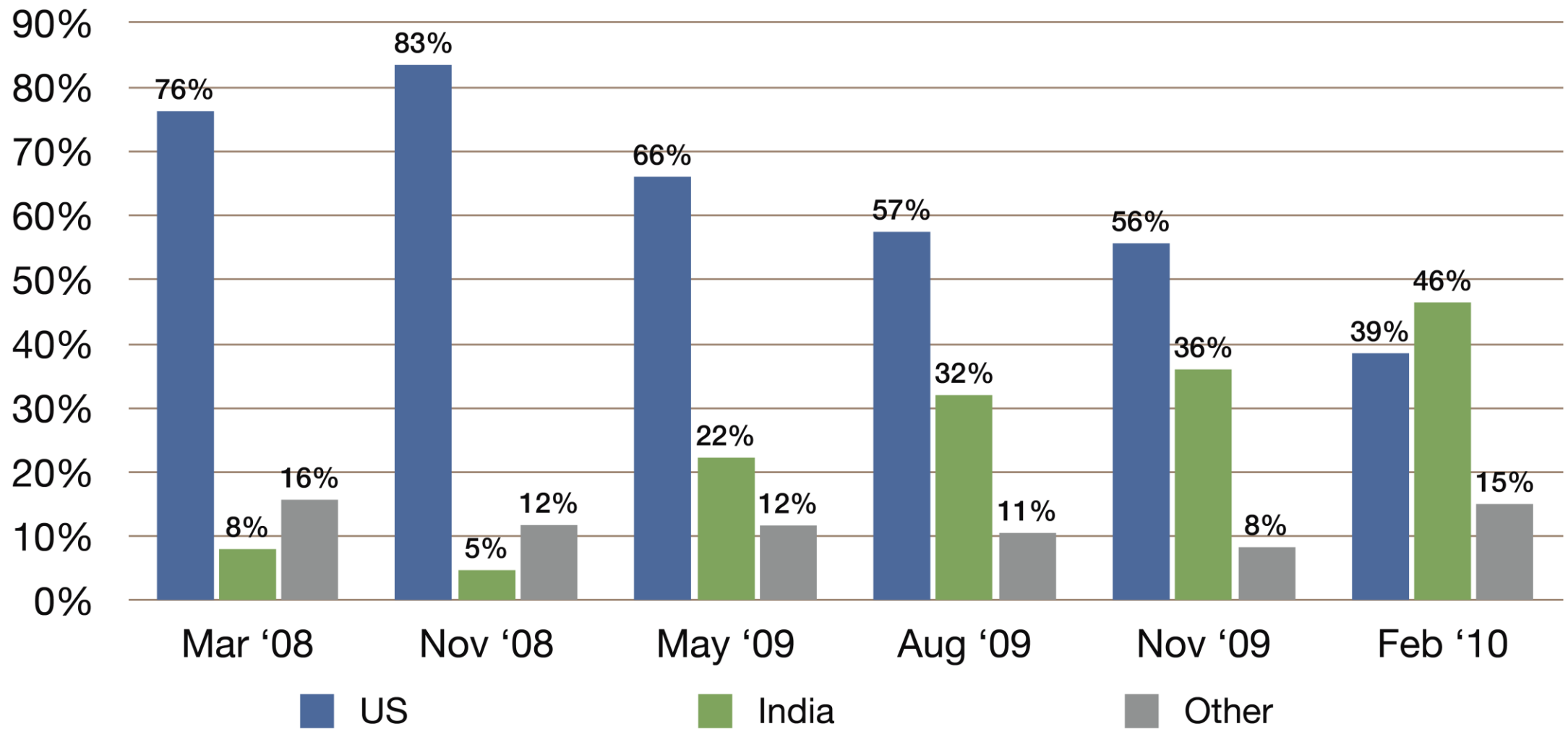
All HITs
51-60 of 1966 Results

Sort by: **HIT Creation Date (newest first)** **GO** Show all details | Hide all details [First](#) << [Previous](#) < [4](#) [5](#) [6](#) [7](#) [8](#) > [Next](#) >> [Last](#)

Grade Snippet of Audio Transcription View a HIT in this group		
Requester: CastingWords	HIT Expiration Date: Nov 19, 2010 (7 hours 46 minutes)	Reward: \$0.03
	Time Allotted: 1 hour 30 minutes	HITs Available: 6
Event Moods and Modes View a HIT in this group		
Requester: Restaurant Recommender	HIT Expiration Date: Nov 26, 2010 (6 days 23 hours)	Reward: \$0.10
	Time Allotted: 10 minutes	HITs Available: 676
Review & Proof this short article View a HIT in this group		
Requester: Todd Dickerson	HIT Expiration Date: Nov 20, 2010 (23 hours 46 minutes)	Reward: \$0.15
	Time Allotted: 60 minutes	HITs Available: 1
Grade Audio Transcription View a HIT in this group		
Requester: CastingWords	HIT Expiration Date: Nov 20, 2010 (11 hours 46 minutes)	Reward: \$0.07
	Time Allotted: 3 days 18 hours	HITs Available: 6

Recency
Money vs time
Reputation

Demographics – workers country



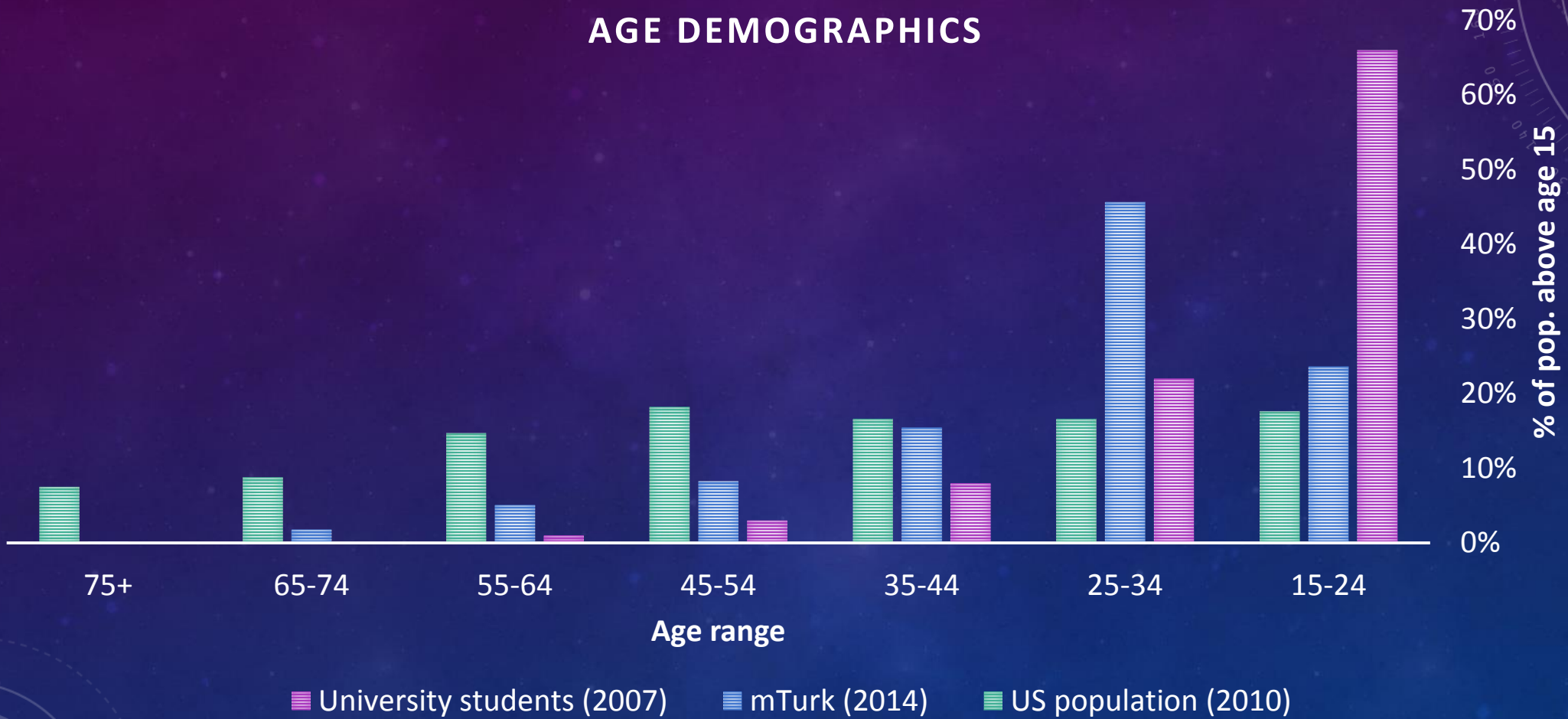
Demographics – age, gender, education, income

		Nov '08	Feb '09	May '09	Aug '09	Nov '09	Feb '10
Age (average)	Overall	32.9	33.0	32.5	31.0	31.7	30.4
	US	33.6		34.3	33.2	35.4	34.3
	India			28.8	27.6	26.4	27.8
Gender (male/female)	Overall			54% / 45%	55% / 45%	48% / 52%	55% / 45%
	India			54% / 45%	55% / 45%	48% / 52%	55% / 45%
Education (Bachelors or higher)	Overall	47%		46%	69%	66%	61%
	US	43%		46%	69%	66%	61%
	India	96%		74%	69%	66%	61%
Income (<\$10k/yr.)	Overall	10%	22%		27%	32%	39%
	US	8%		No data	12%	10%	15%
	India	35%			54%	65%	61%

Buhrmester, Kwang and Gosling (2011):
 “In short, MTurk participants were more demographically diverse than standard Internet samples and significantly more diverse than typical American college samples.”

Demographics – age

AGE DEMOGRAPHICS



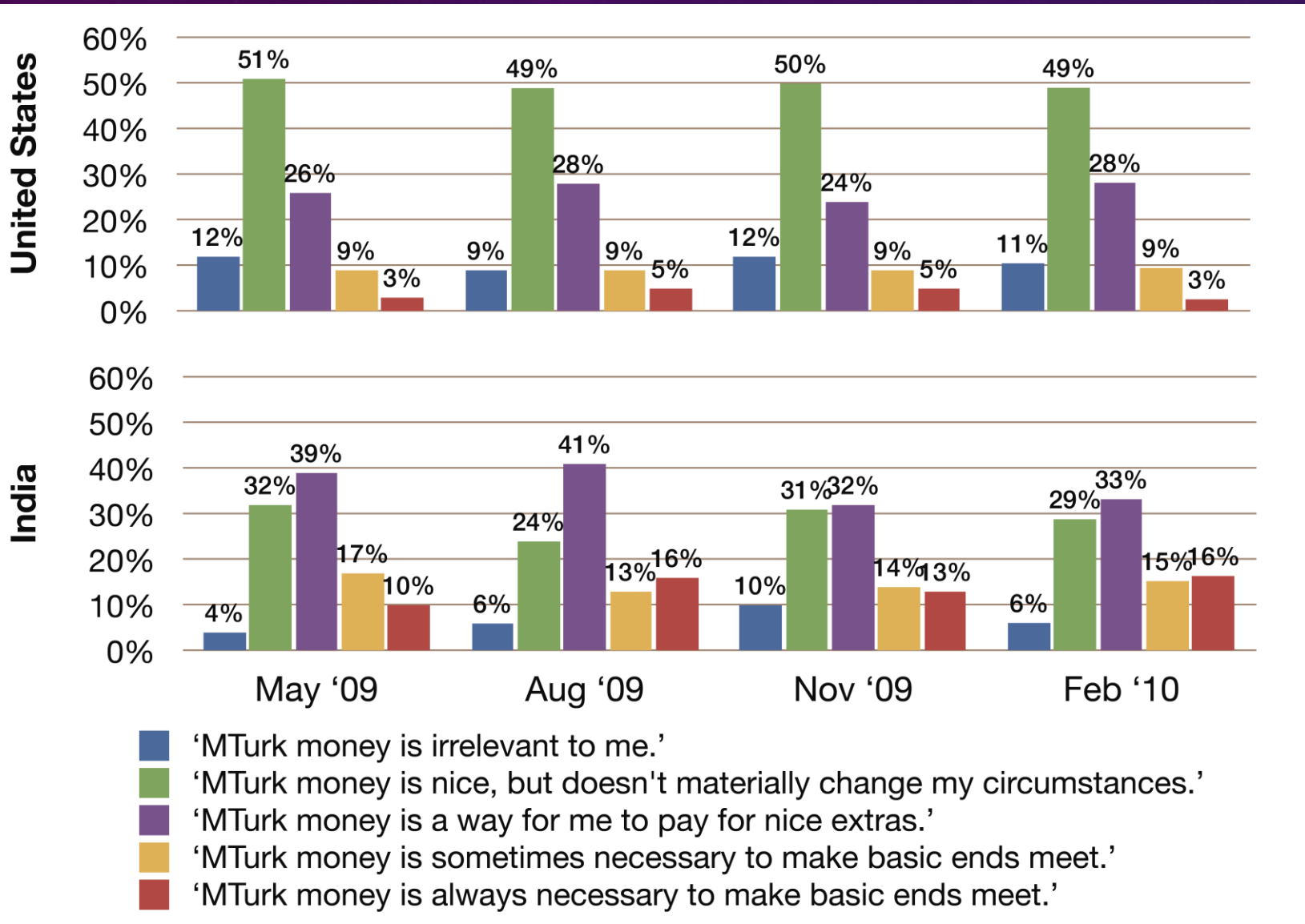
Can we trust demographic data?

- **Mason, Suri (2011) Conducting behavioral research on Amazon's Mechanical Turk**
 - Only 0.4% (1/207) changed their demographic data
- **Rand (2012) The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments**
 - I.P. address check
 - Country of residence truthfulness – 97%
 - Same participants in 2 separate studies (N~100)
 - Gender – 96%
 - Age – 93% (within 1 year)
 - Country – 98%
 - Education level – 81%
 - Yearly income – 82% (within 1 bracket out of 10)
 - Belief in god – 84% (within 1 point out of 10)

Workers' perspective

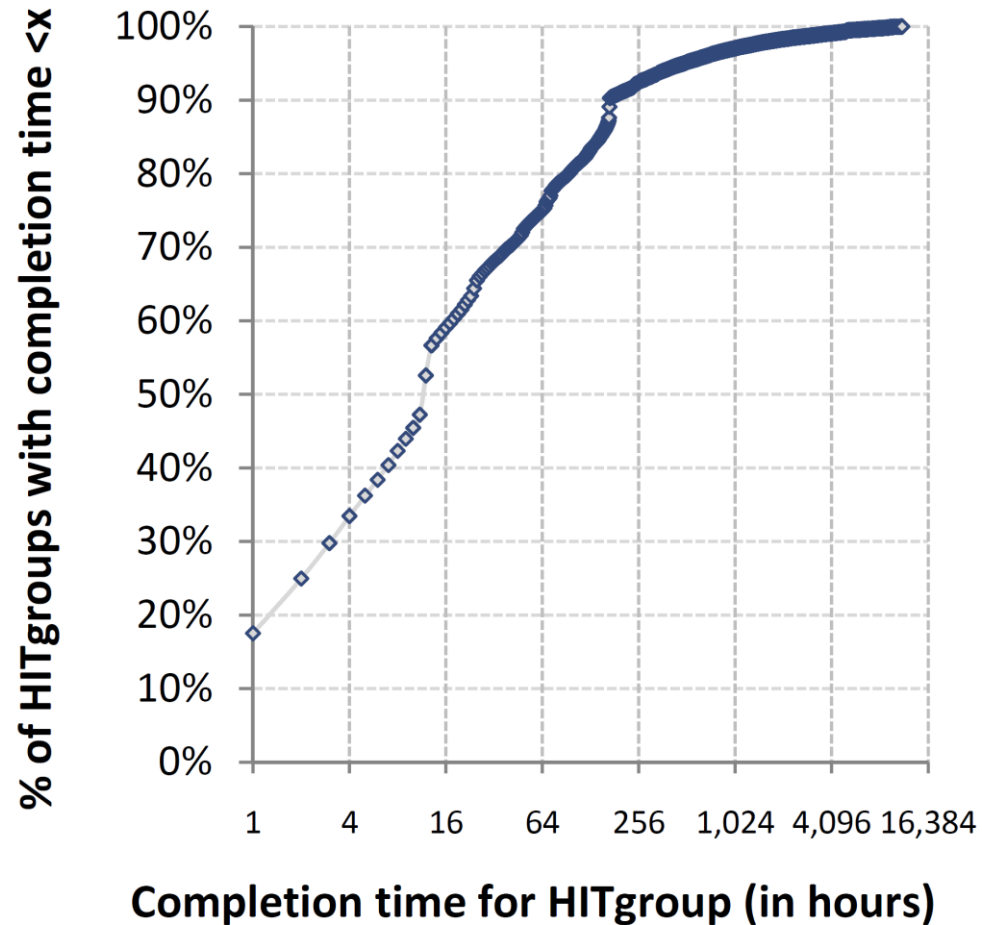
- 59% of Indian workers and 69% of U.S. workers agreed that “Mechanical Turk is a fruitful way to spend free time and get some cash” (Ipeirotis, 2010).
- “Most workers are not motivated primarily by the financial returns and genuinely care about the quality of their work” (Paolacci, Chandler, Ipeirotis, 2010).

Demographics – MTurk as a source of income

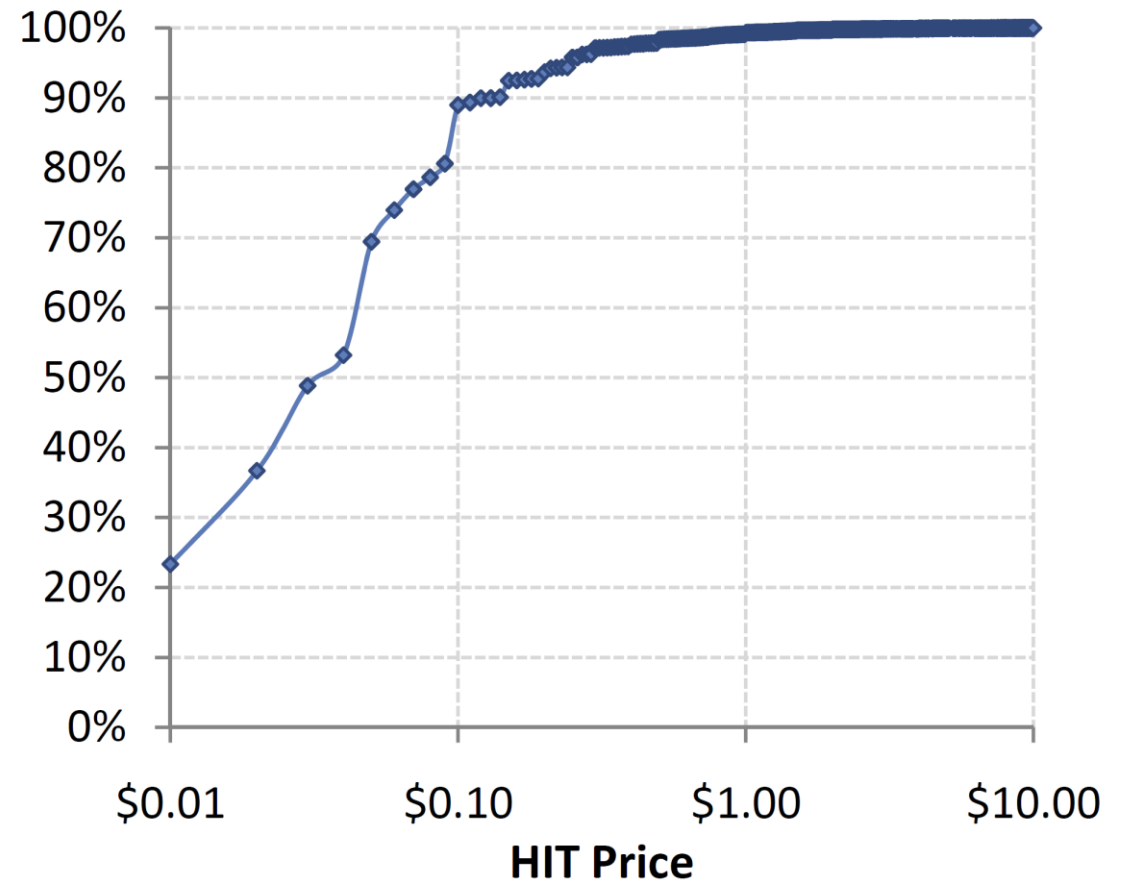


Price and Time

CDF of completion times for HIT Groups



% of HITs vs HIT price



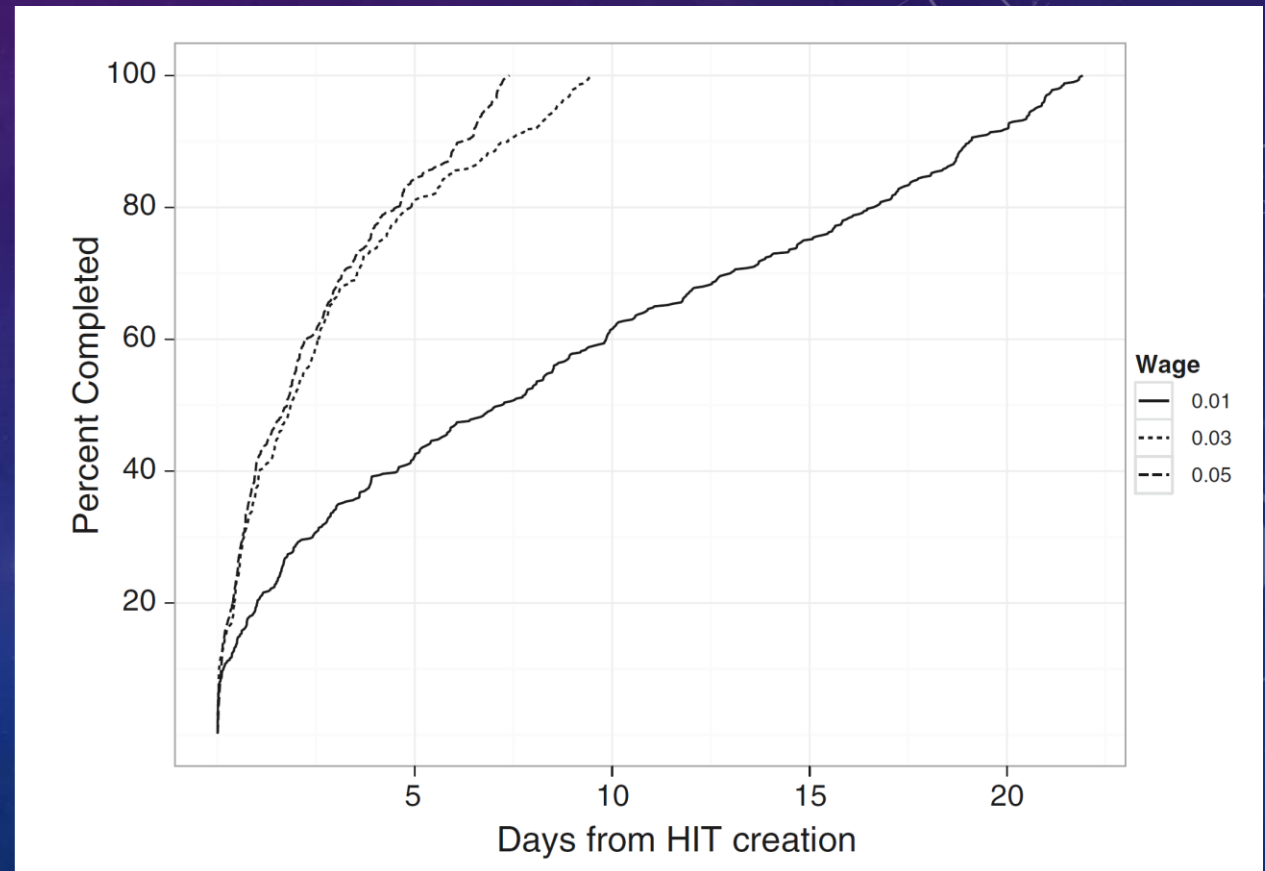
Price and Time

Buhrmester, Kwang and Gosling (2011)

HIT completed per hour

Length Compensation	5 min	10 min	30 min
2¢	5.6	5.6	5.3
10¢	25.0	14.3	6.3
50¢	40.5	31.6	16.7

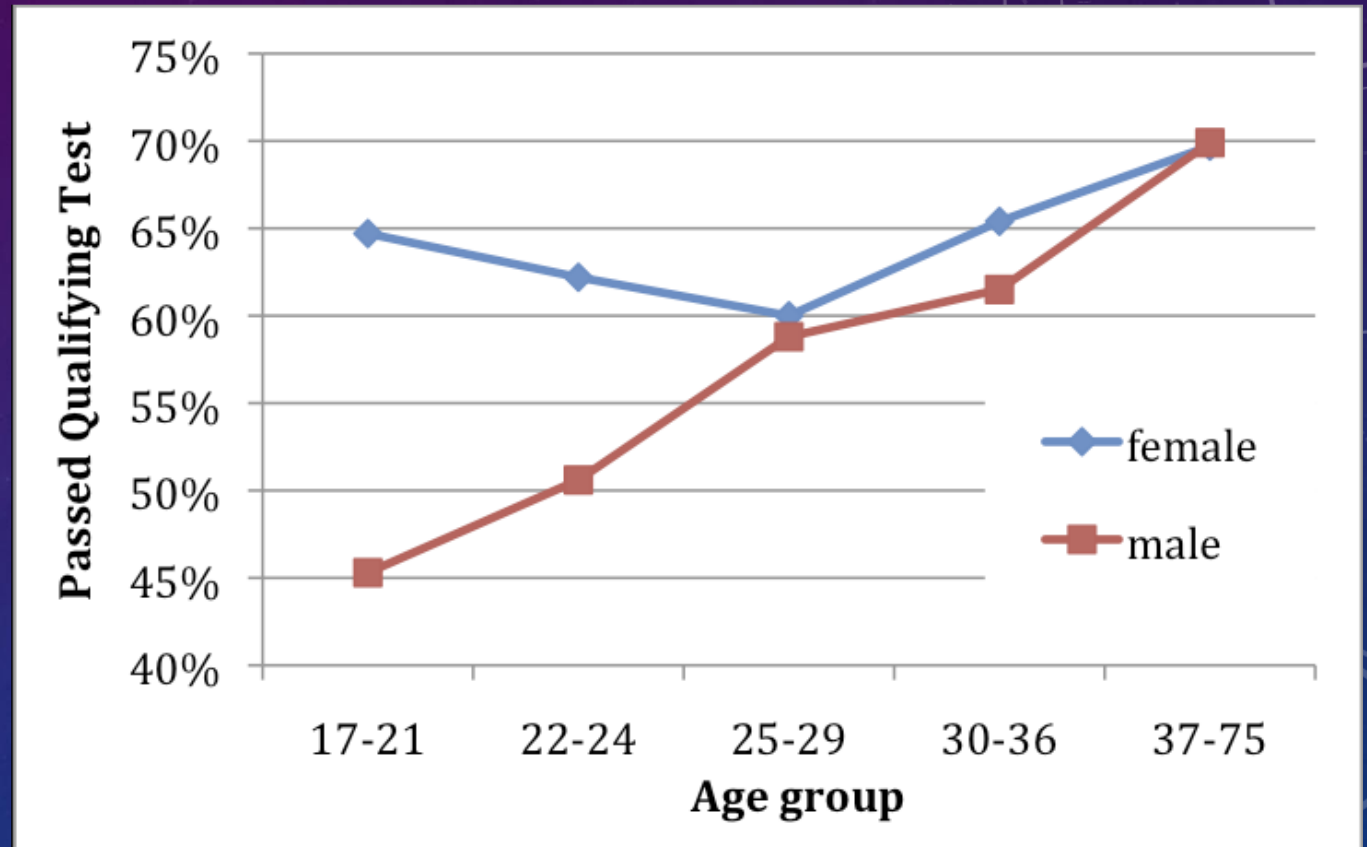
Mason, Suri (2011)



Filtering – un-rejected rate

My experience without pre-screening: ~ 50%

With pre-screening: ~ 80-90%



Filtering

- **Post**

- ACQ

- Comprehension question (manipulation check)
 - “How would you answer... if...”
 - “please select b”

- Time measurement (whole assignment/per page)

- **Pre**

- Success rate, min. number of HITs
 - **Country** (multiple choice available now)

	Invalid Comment Responses	Median duration	Duration < 1 minute
Exp 1	48.6%	1:30	30.5%
Exp 2	2.5%	4:06	6.5%

Table 1. Improvement in response quality in Experiment 2 upon the introduction of verifiable questions.

Kittur, Chi, Suh (2008)

Filtering

		High Reputation		Low Reputation		
		Passed ACQs	No ACQs	Passed ACQs	Failed ACQs	No ACQs
<i>N</i>		302	156	117	60	59
Cronbach's alpha	SDS	.629 _a	.698 _a	.471 _b	.242 _b	.557 _{ab}
	RSES	.936 _a	.934 _{ad}	.912 _{ad}	.825 _{bc}	.889 _{cd}
	NFC	.952 _a	.947 _{ad}	.891 _{ad}	.759 _{bc}	.863 _{cd}
SDS mean percent (<i>SD</i>)		44.87 (21.5)	45.71 (23.7)	48.63 (18.9)	53.0 (17.3)	49.83 (21.2)
Anchoring effect size (<i>r</i>)		.198 _a *	.183 _a *	.280 _a *	-.046 _b	.049 _b
Average percent of midpoint marked on scale items (<i>SD</i>)		19.28 (14.1)	20.78 (14.06)	25.12 (17.04)	34.21 (26.56)	27.61 (21.08)

- Peer, Vosgerau, Acquisti (working paper) **Reputation as a sufficient condition for data quality on Amazon Mechanical Turk**
 - ACQ increase data quality
 - But only for low reputation workers...
 - No significant effect in high reputation (>95%, >1000 HIT)
 - Conclusion
 - Use only high reputation workers (not 'master workers')

Reputation	Passed (All ACQs)	Failed (at Least One ACQ)	# of ACQs Failed		
			1	2	3
High	294 (97.4 %)	8 (2.6 %)	8 (2.6 %)	0	0
Low	117 (66.1 %)	60 (33.9 %)	35 (19.8 %)	14 (7.9 %)	11 (6.2 %)

Testing – Behavioral decision making

- **An Assessment of Experiments run on Amazon's Mechanical Turk, Wolfson & Bartkus, 2013**

- Endowment Effect

- Kahneman, Knetsch and Thaler (1986) – car dealer, wages



- Prospect Theory

- Kahneman and Tversky (1979) – risk averse - risk seeking



- Anchoring

- Tversky and Kahneman (1974) – wine price* (social sec. num.)



- Stewart (2009) – credit card payment (minimum payment)



Testing – Behavioral decision making

Table 3: Results on experimental tasks.

	Mechanical Turk	Midwestern university	Internet boards
<i>Asian Disease</i>			
% Risky Positive Frame	17.6%	28.1%	23.7%
% Risky Negative Frame	55.3%	67.7%	63.0%
χ^2	10.833	20.230	13.013
p	< 0.001	< 0.001	< 0.001
Effect size (w)	0.39	0.39	0.39
<i>Linda problem</i>			
% Conjunction Fallacy	72.2%	78.3%	64.4%
<i>Physician problem</i>			
Avg. Quality Success (SD)	5.93 (0.81)	5.63 (0.75)	5.73 (0.98)
Avg. Quality Failure (SD)	5.13 (1.24)	4.86 (1.29)	4.93 (1.41)
t	3.70	4.14	2.547
p	< 0.001	< 0.001	0.007
Effect size (d)	0.76	0.73	0.66

Paolacci, Chandler, Ipeirotis (2010)

Testing – Cognitive psychology

- Crump, McDonnell, Gureckis (2013) **Evaluating Amazon’s Mechanical Turk as a Tool for Experimental Behavioral Research**

- Reaction time

- Stroop

“All of the reaction time tasks chosen for validation purposes were replicated. In addition, error rates were low overall suggesting that participants took the task seriously... Overall, these replications highly recommend AMT as a tool to conduct multi-trial designs that use reaction time as a dependent measure.”

“In conclusion, AMT is a promising development for experimental cognitive science research. On balance, our investigations suggest that the data quality is reasonably high and compares well to laboratory studies... If we as scientists respect the participants and contribute to a positive experience on AMT it could turn into an invaluable tool for accelerating empirical research.”

... it used relatively ... constraints for ... mes.”

- Attention blink

- Subliminal priming

- Learning

- Category learning

“online participants generally learned more slowly [compared to lab-based studies]... the magnitude of payment does not have a strong effect on the quality of data obtained from online, crowd-sourced systems... building in checks for understanding the instructions is critical for ensuring high quality data.”

Pros and Cons

- Cons
 - Reliability?
 - Only internet
 - Random attribution (dropout)
- Pros
 - Simple
 - Fast
 - Cheap (by some studies in the US – by a factor of 6 compared to students in lab)
 - More external validity than using students
 - Diversity (age, income...)
 - International (Culture studies)
 - Big pool (>1,000,000?)

Unresolved issues

- Personality (what kind of a person is a Turker?)
- Workload (ego depletion)
- Speciality (familiarity with known manipulations)
- Environment (Alone vs. in a lab: pro or con?)
- Trust (Turker - gambles, Requester - details)

Special constellations

- **Longitudinal studies:**

- Holden, Dennie and Hicks (2013)
 - M5-120 personality scale (120 items)
 - 3 weeks test-retest: 280 1st wave, 67 2nd wave (46 relevant)
 - 0.1\$ 1st, 0.15\$ 2nd (41% india)
- Buhrmester, Kwang and Gosling (2011)
 - Big-5 personality scale (44 items)
 - 3 weeks test-retest: 60% 2nd wave
 - 0.2\$ 1st, 0.5\$ 2nd

- **Interaction tasks:**

- Mason, Suri (2012) Conducting behavioral research on Amazon's Mechanical Turk
- Suri, Watts (2011) Cooperation and Contagion in Web-Based, Networked Public Goods Experiments
- Pseudo-dyadic – Summerville and Chartier (2012)

Tips - Requesters reviews

Turkopticon (<http://microwork-dev.ucsd.edu/>)

ment and vote on an article. Easy!

requester: **Product Search** HIT Expiration

communicativity: 1.00 / 5

generosity : 2.57 / 5

fairness : 2.86 / 5

promptness : 2.00 / 5

What do these scores mean?

Scores based on 7 reviews

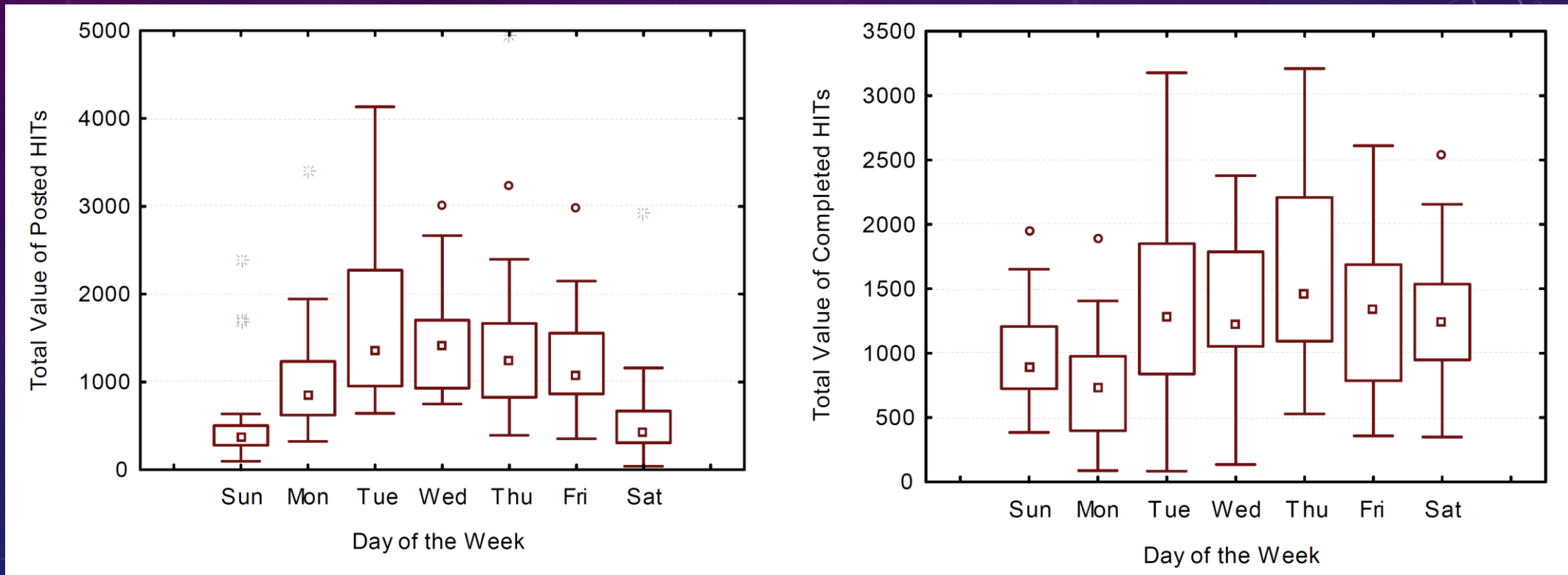
Report your experience with this requester »

requester: **MR. MOVIE QUOTE** HIT expiration

Time Alotted:

Itay Sisso A37QCE2O EZ934F Averages » HIT Group » Review Requester »	FAIR: 5 / 5 FAST: 5 / 5 PAY: 5 / 5 COMM: NO DATA	I did his survey before. he approved fast and paid well. idk about the previous comment. Jun 19 2013 Real reviews flag comment
Itay Sisso A37QCE2O EZ934F Averages » HIT Group » Review Requester »	FAIR: NO DATA FAST: NO DATA PAY: 5 / 5 COMM: NO DATA	very short survey \$0.50 TIME: 2m 4s \$14.44/h Oct 28 2013 ptosis flag comment
Itay Sisso A37QCE2O EZ934F Averages » HIT Group » Review Requester »	FAIR: 5 / 5 FAST: 5 / 5 PAY: 5 / 5 COMM: NO DATA	paid next day, good money for the amount of work. would work for again. no need to contact Oct 29 2013 dedsi...@g... flag comment
Itay Sisso A37QCE2O EZ934F Averages » HIT Group » Review Requester »	FAIR: 5 / 5 FAST: 5 / 5 PAY: 5 / 5 COMM: NO DATA	10/29, very short survey for .50, ~2 minutes and paid 10/30. Oct 30 2013 rubyr...@m... flag comment
Itay Sisso A37QCE2O EZ934F Averages » HIT Group » Review Requester »	FAIR: 5 / 5 FAST: 5 / 5 PAY: 5 / 5 COMM: NO DATA	no problems Oct 31 2013 absin...@y... flag comment

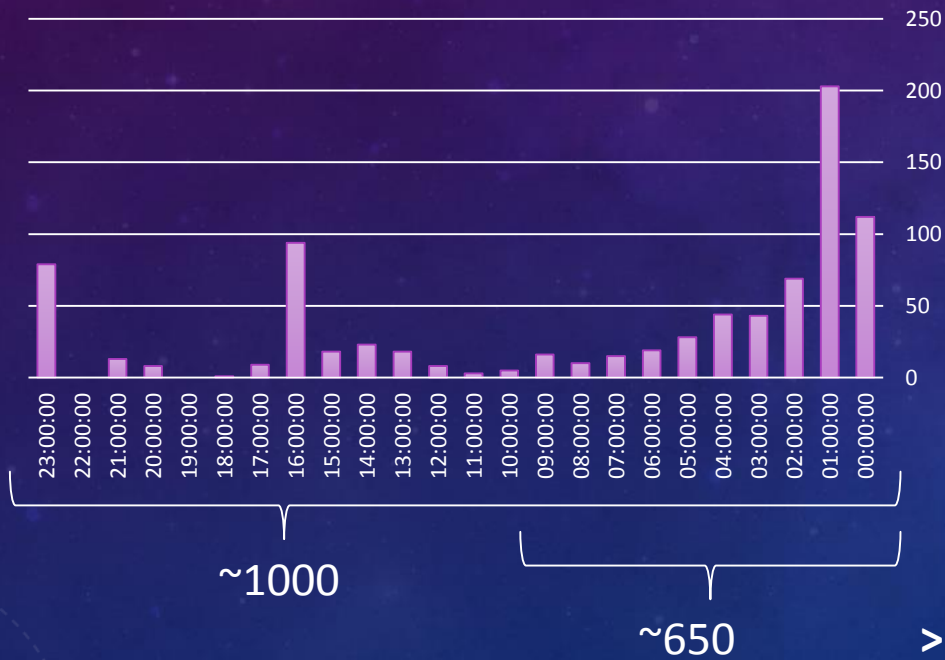
Tips – when to post?



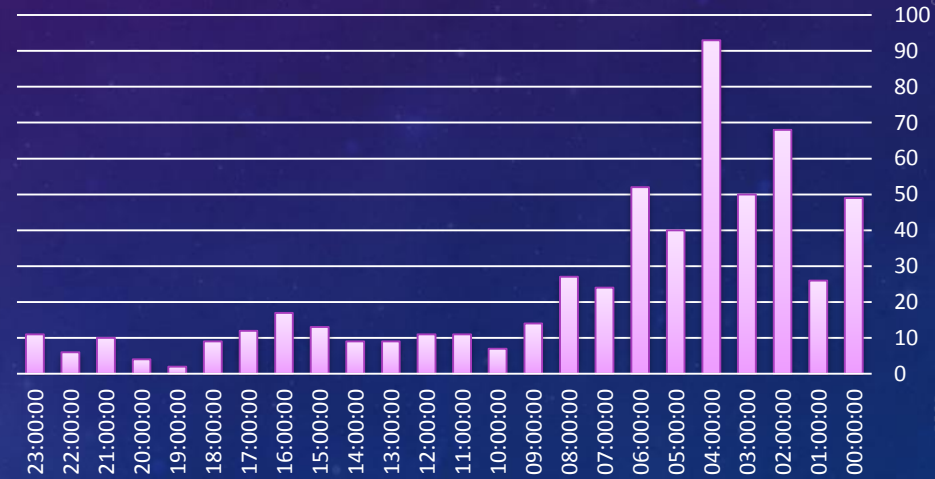
April 2010

Tips – When to post?

Respond in time of day (Israel)



Respond in time of day (Israel)



**>1000 HIT, >95%
US only
1\$, 10min**

Other tips

- Filtering by MturkID (Qualification / script+Qualtrics)
- Differential payement (Bonus)
- Test – Retest (e-mail)