Sampling and Research Online

Psych 251

10/16/17

Course updates

- Problem set #2
- Project comments coming
  - Once you are approved, assess material and draft author contact (from template)
  - This week - begin choosing implementation format
- 10/30 midterm presentations
  - 5min + 2min questions
  - Mock up of experiment for comments
  - More info on website

Outline

- **Measures in R**
- **Sampling**
  - Basics of sampling theory
  - Stratification
- **Research Online**

Stevens (1946) classification

Scale Type	Permissible Statistics	Admissible Scale Transformation	Mathematical structure
nominal / categorical	mode, Chi-squared	substitution	unordered
ordinal	median, percentile	monotonic increasing	totally ordered set
interval	mean, standard deviation, correlation, regression, analysis of variance	positive linear	affine line
ratio	above plus geometric mean, harmonic mean, coefficient of variation, logarithms	Positive similarities (multiplication)	one-dimensional vector space

Variables in R

- Numbers
  - Generalize to arrays
  - `x <- 1` or `x <- c(1,2,3,4)`
  - `x[1]` or `x[1:10]`
- Characters
  - `x <- “hello”` or `x <- c(“hi”,”hey”,”hello”)`
- Factors:
  - nominal or ordinal variable type
  - `abc <- factor(c("x","y","x","z"))`
  - `[1] x y x z`
  - `Levels: x y z`
  - `ordered(ses, levels = c("low", "middle", "high"))`
  - `Levels: low < middle < high`

In R (continued)

- Nominal and ordinal variables are both factors
  - The levels have an ordering, but this ordering can be arbitrary or fixed
  - Up to you to make sure it's reasonable
  - `levels(x) <- c("agree","disagree","neutral")`
  - `levels(x) <- c("disagree","neutral","agree")`
  - Make sure you're not renaming variables though!
- Interval and ratio variables are both numbers
  - It's up to you to decide whether zero is meaningful
- But bad things can happen...
  - E.g. mean of a factor?

forcats

forcats

CRAN 0.1.1 | build passing | coverage 77%

Overview

R uses **factors** to handle categorical variables, variables that have a fixed and known set of possible values. Historically, factors were much easier to work with than character vectors, so many base R functions automatically convert character vectors to factors. (For more historical context, I recommend *stringsAsFactors: An unauthorized biography* by Roger Peng, and *stringsAsFactors = <sigh>* by Thomas Lumley.) These days, making factors automatically is no longer so helpful, so packages in the tidyverse never create them automatically.

- `fct_relevel()` is similar to `stats::relevel()` but allows you to move any number of levels to the front.
- `fct_inorder()` orders according to the first appearance of each level.
- `fct_infreq()` orders from most common to rarest.
- `fct_rev()` reverses the order of levels.

Outline

- **Measures in R**
- **Sampling**
  - Basics of sampling theory
  - Stratification
- **Research Online**

Sampling

- Want to estimate some measure(s) for a large or infinite population, so test some subset of that population
- **Sample frame**: list to sample from
- Major concerns in sampling
  - **Power**: do you have a large enough sample to detect an effect?
  - **Bias**: does your method of sampling asymptotically converge to the correct estimate?
  - **Independence**: do you modify your sampling scheme after beginning data collection?

Power

- **Power** is the probability of detecting an effect of known size given some sample
- **Power analysis** (later in the quarter) is how to find this out
  - A priori: how big a sample do we need
  - Post hoc: how much power did they have
- Intuitive power analysis (for when you don't know how big the effect is)
  - How many height measurements would we need to get a reliable difference between men and women?
  - How about a reliable measure of colorblindness among genders? (~2%)

Bias: Non-probability sampling

- Not all members of the population have a chance of being included
- **Convenience sample**
  - Cheap, easy, but can have issues in generalizability
- **Volunteer sample**
  - Selection bias likely: who are the people who would want to be in this study?
- Key concept here is to ask about ways in which sample bias could affect results
  - This may not generalize from study to study: concerns vary

What if you have the whole population?

- Then you're not sampling.
  - You're not trying to generalize to another population
- If you have *most* of the population, you will need to make a *finite population correction*
  - Standard error goes down as you get more and more of the sample

Random sampling

- All members of population have a chance to be included
- **Simple random sampling**
  - Easy when you have some method for doing this
- **Systematic sampling**
  - E.g. take every *k*th person
- **Stratified sampling**
  - Break sample into groups

Two sampling problems

- Nationally-representative survey
  - Want to make sure that minority groups are appropriately represented
  - But if you sample randomly, may end up undersampling minorities by accident
- School-based intervention
  - If you sample students randomly, you will end up visiting 100 schools
  - You will also underweight small schools

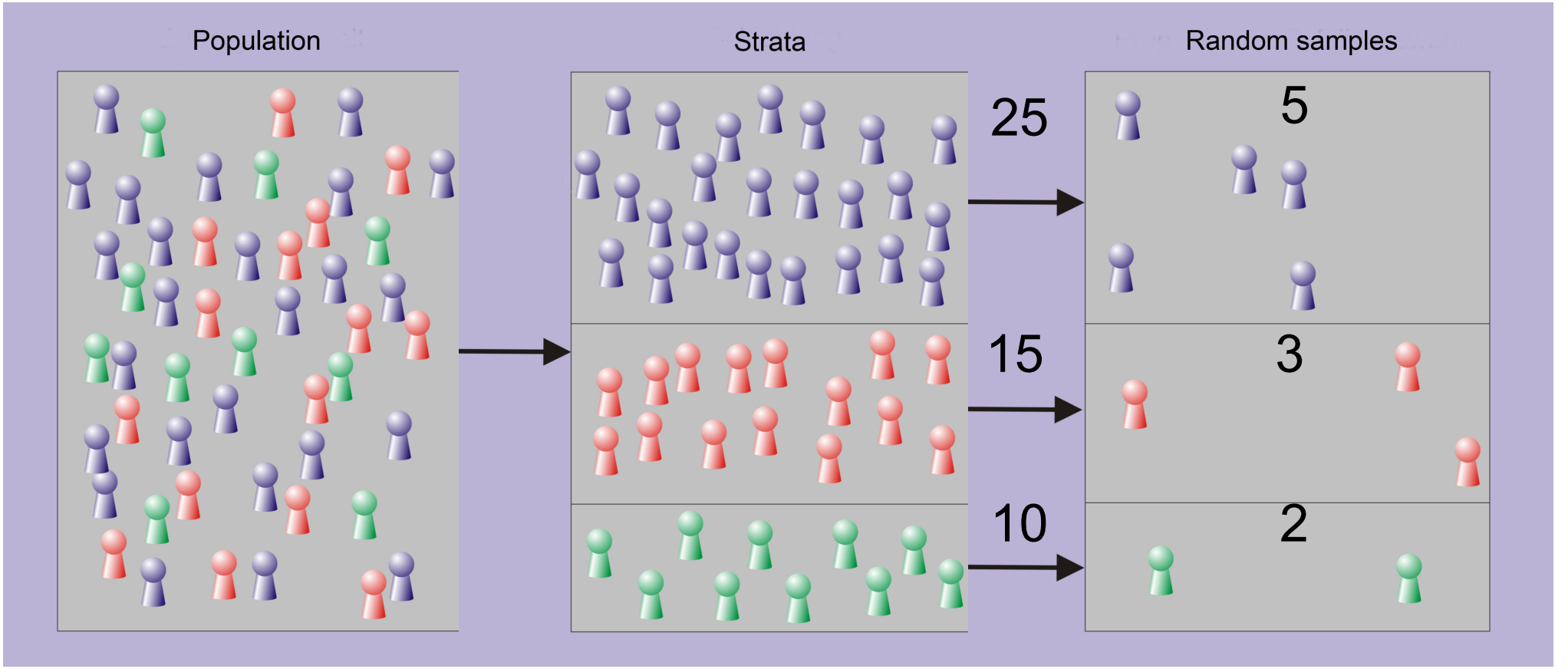
Stratified/hierarchical sampling

- Primary sampling units (psu) are selected with a probability proportional to size of target population psu
  - E.g.: minority populations, schools
- Secondary sampling units (ssu) is selected at random from each of the psu.
  - E.g.: individuals, students in schools

Two stratification strategies

- Proportional
  - Sample from psu proportional to fraction of the total population
  - E.g., make sure you have the same proportion of the minority group as the national proportion
- Disproportionate/optimal
  - Sample from psus relative to their variability or their interest
  - E.g. get more samples from minority groups so you can appropriately estimate their views

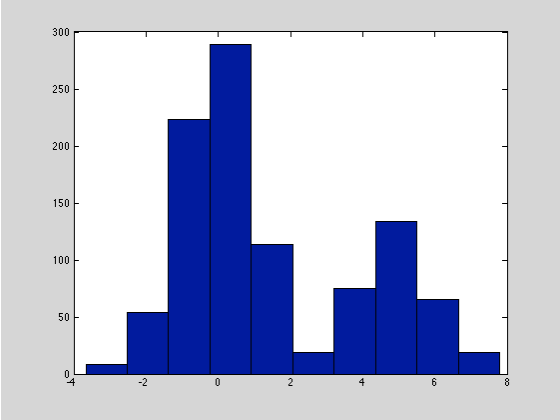
Stratified sampling visualized



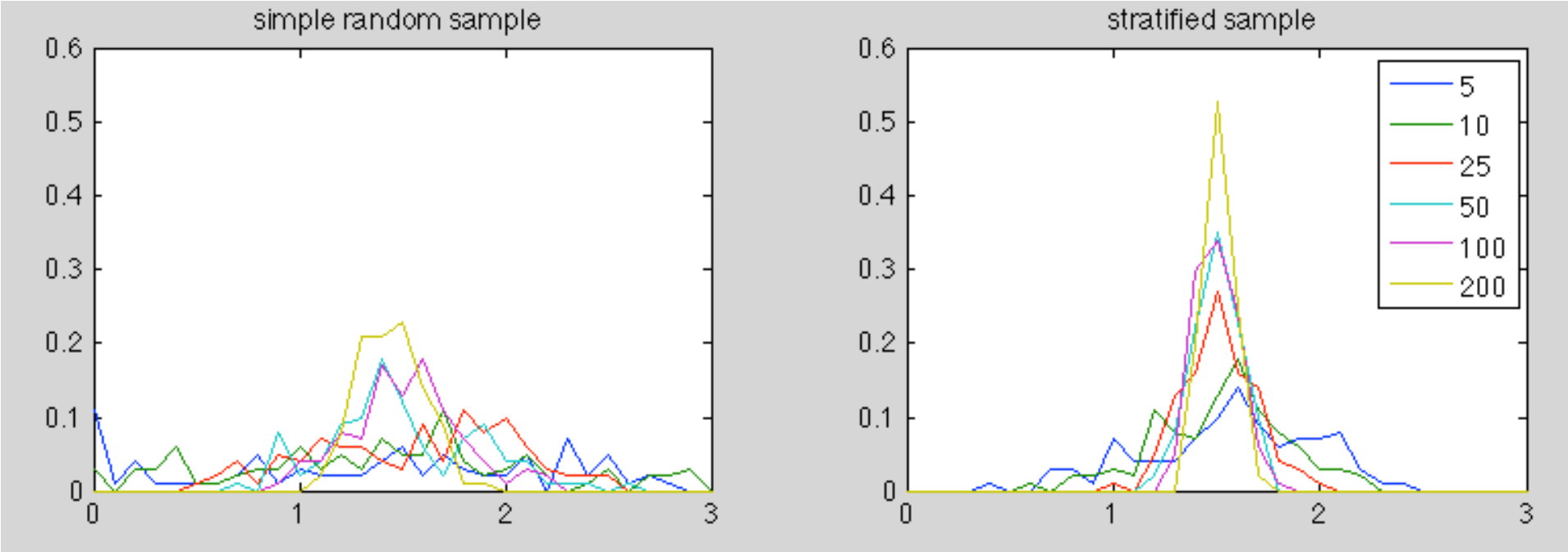
When would you stratify?

- If you want to be sure of precision for each subgroup
  - E.g. nationally representative surveys
- Because of some practicality
  - E.g. students are in schools
- When sampling problems are different for different populations
  - E.g. survey of prisoners vs. jobless adults vs. employed adults
- **Gains in precision of result** when
  - E.g. measure of interest might correlate with strata (e.g. learning outcomes & school size)
  - Hence random sample of students without stratification may undersample kids from small schools

Stratified sampling: simulation



Intuitive explanation:
Sometimes you don't get enough members of the subgroups!



Stratified sample reduces variability for a better estimate of mean

Pre-specification of sampling

- Make a plan and write it down
  - Why?
  - Avoid contingent data collection
    - Pre-specify data collection plan
    - Analyze as much as you want
- Alternative
  - Correct for contingent data collection
    - Sequential analysis
    - Bayesian methods
    - When would you do this?

General Article

aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science
22(11) 1359–1366

DOI: 10.1177/0956797611417632
http://pss.sagepub.com
SAGE

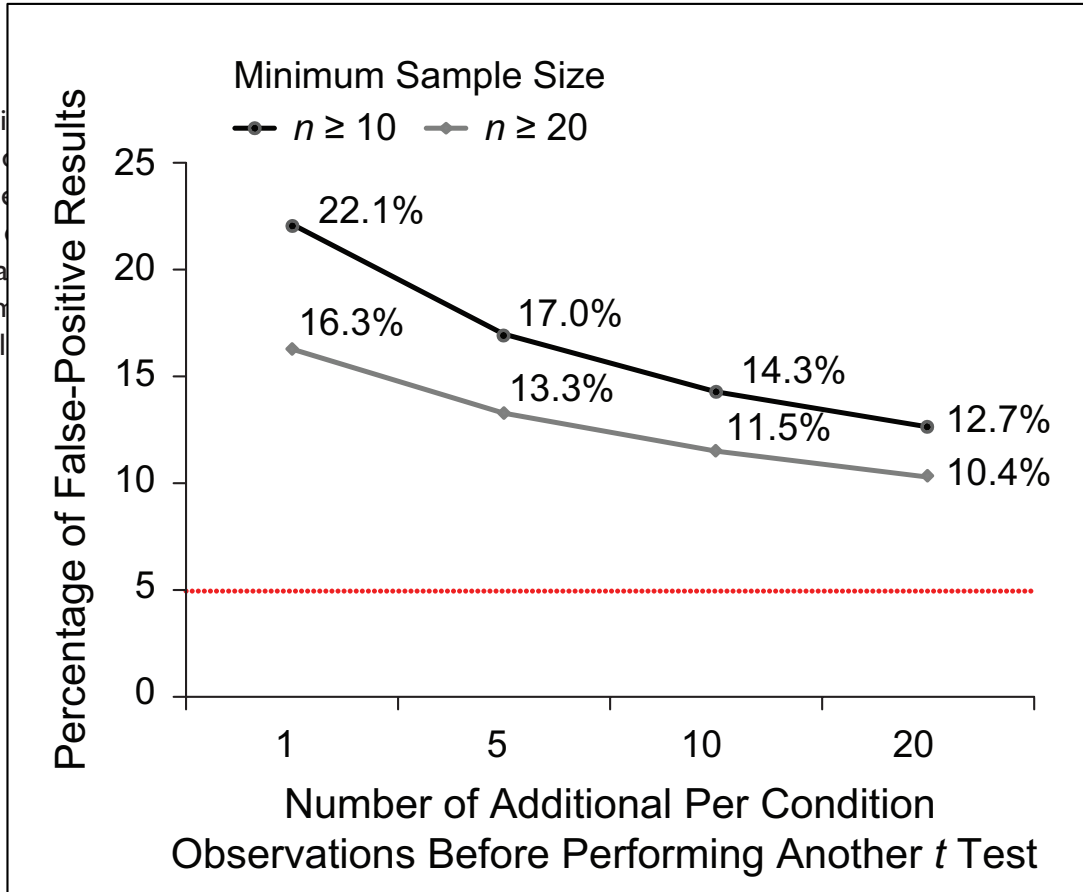
False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

Abstract

In this article, we accomplish two things. First, we show that despite empi
of false-positive findings (≤ .05), flexibility in data collection, analysis, an
rates. In many cases, a researcher is more likely to falsely find evidence
that it does not. We present computer simulations and a pair of actual
it is to accumulate (and report) statistically significant evidence for a fa
and straightforwardly effective disclosure-based solution to this problem
authors and four guidelines for reviewers, all of which impose a minimal

"P-value sniffing": changing sampling scheme non-independently



Sampling: the bottom line

- Think about whether bias introduced by sampling method could have had something to do with observed effect
  - No mechanistic procedure for this
  - Depends on the claim you want to test
  - And crucially, the generalization that you want to make from it

#overlyhonestmethods

- A sample size of 150 patients gave 90% power, and by a remarkable coincidence, was what the sponsor had budgeted for #overlyhonestmethods
- Sample size was smaller than planned because I had been in grad school for 10 yrs & my advisor wanted me to graduate. #overlyhonestmethods
- No SES data were available for the sample, as the testers forgot to ask. #overlyhonestmethods
- We surveyed a sample of American college undergrads and extrapolated to all humans, because... #evolutionarypsychology #overlyhonestmethods

Outline

- **Measures in R**
- **Sampling**
  - Basics of sampling theory
  - Stratification
- **Research Online**

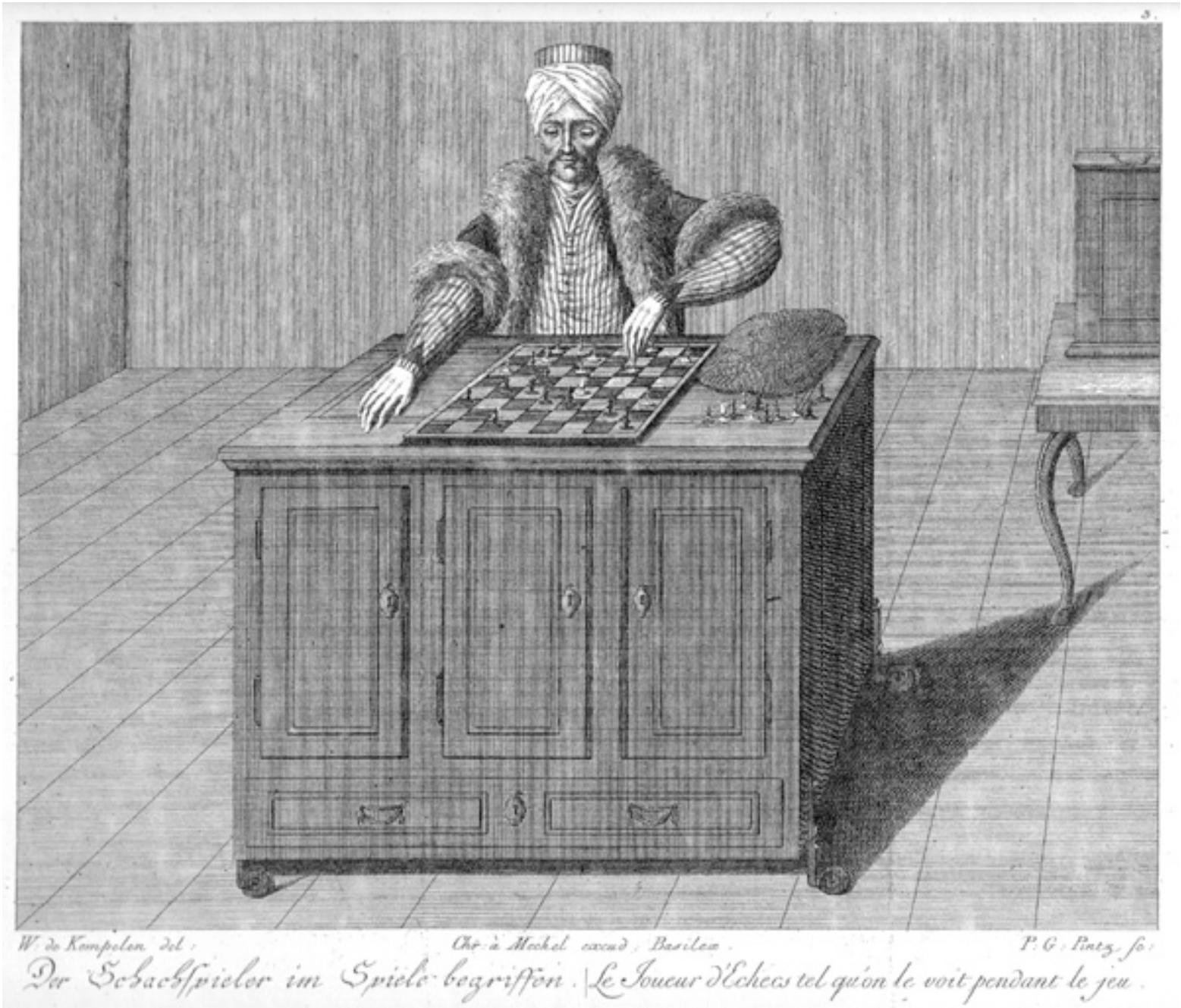
Why do research on the Internet?

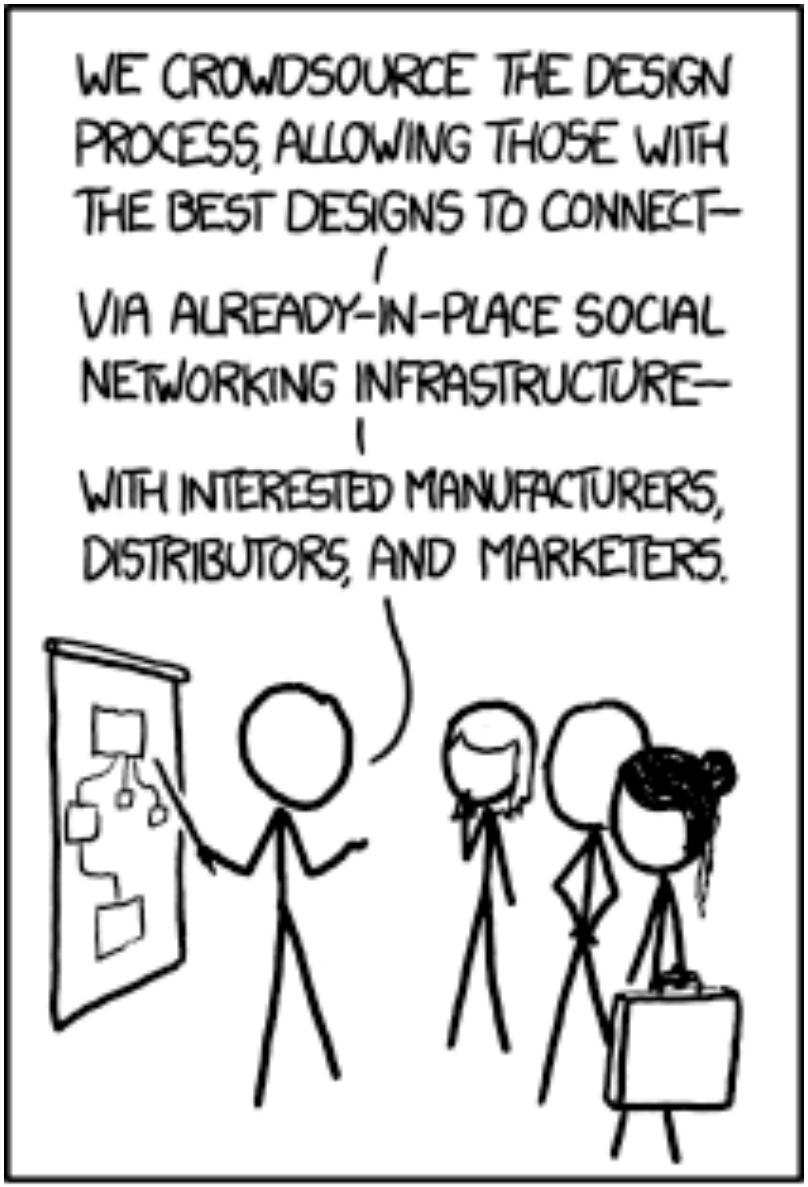
- More data, faster collection
- Hassle-free iteration
- No RA required
- Code can be reused
- Standardized instructions
- No expectancy effects
- Tasks can be replicated in any browser (no funny business or “gotchas”)
- Often cheaper (but ethics?)

What is crowdsourcing?

- Any opportunity to get many people to do small bits of work
  - Sometimes but not always in exchange for pay
- Many different types
  - Microwork: murk.com, crowdflower.com
  - Crowdvoting: contest sites like 99designs
  - Crowdfunding: kickstarter.com, indiegogo.com, etc.
- Novel applications (there are hundreds!)
  - ManyEyes for data visualization
  - Crowdsourced medical research (finding patients etc.)
  - Folding@home etc. etc.

The Mechanical Turk (mturk.com)




WE CROWDSOURCE THE DESIGN PROCESS, ALLOWING THOSE WITH THE BEST DESIGNS TO CONNECT—
VIA ALREADY-IN-PLACE SOCIAL NETWORKING INFRASTRUCTURE—
WITH INTERESTED MANUFACTURERS, DISTRIBUTORS, AND MARKETERS.
NOBODY CAUGHT ON THAT OUR BUSINESS PLAN DIDN'T INVOLVE *US* IN ANY WAY—IT WAS JUST A DESCRIPTION OF OTHER PEOPLE MAKING AND SELLING PRODUCTS.

Luis von Ahn

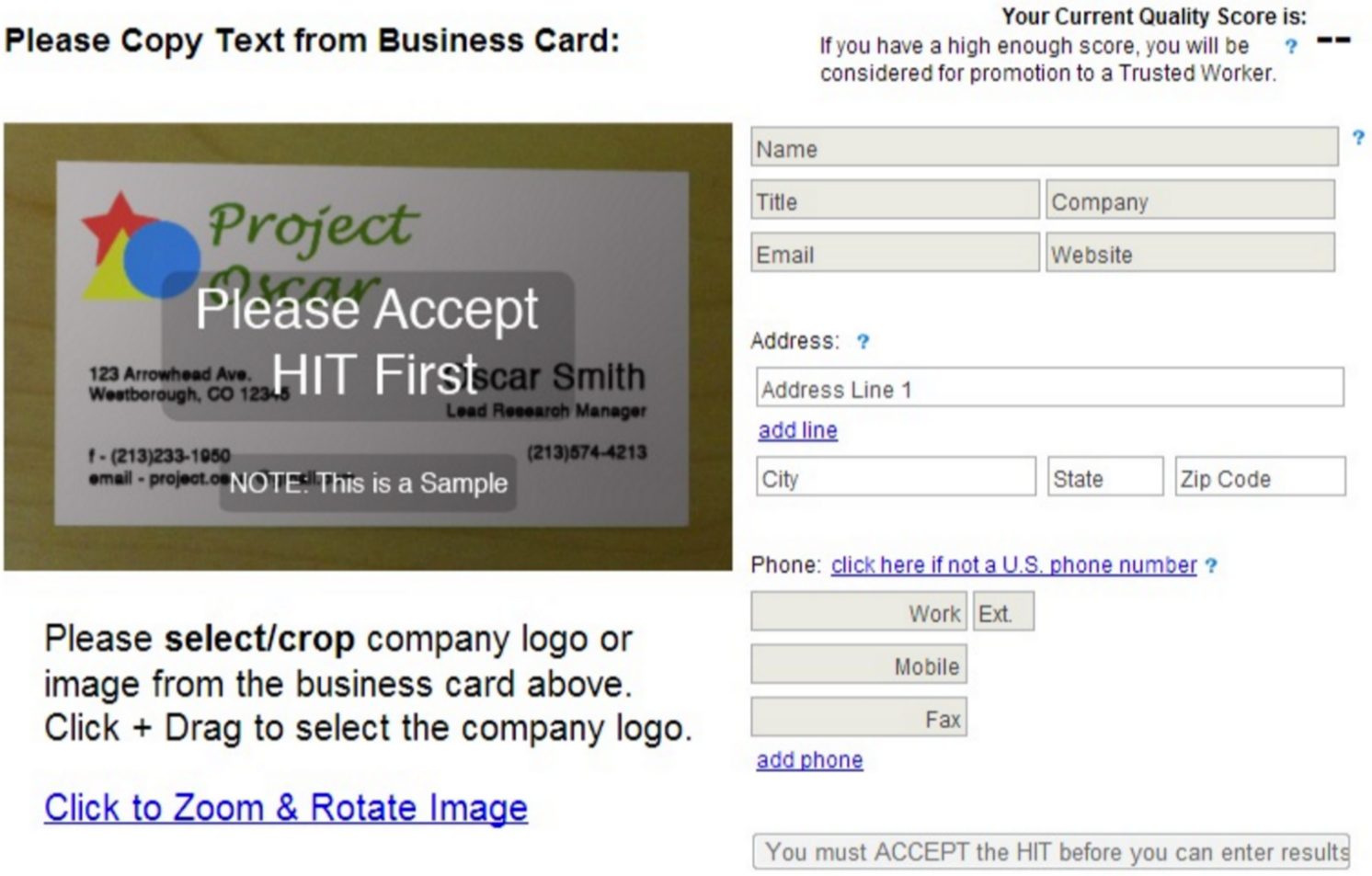

433

Business

11 absurd Mechanical Turk tasks you can do for pennies

By Kris Holt

Jan 14, 2013, 10:49am CT | Last updated Jan 14, 2013, 11:21am CT

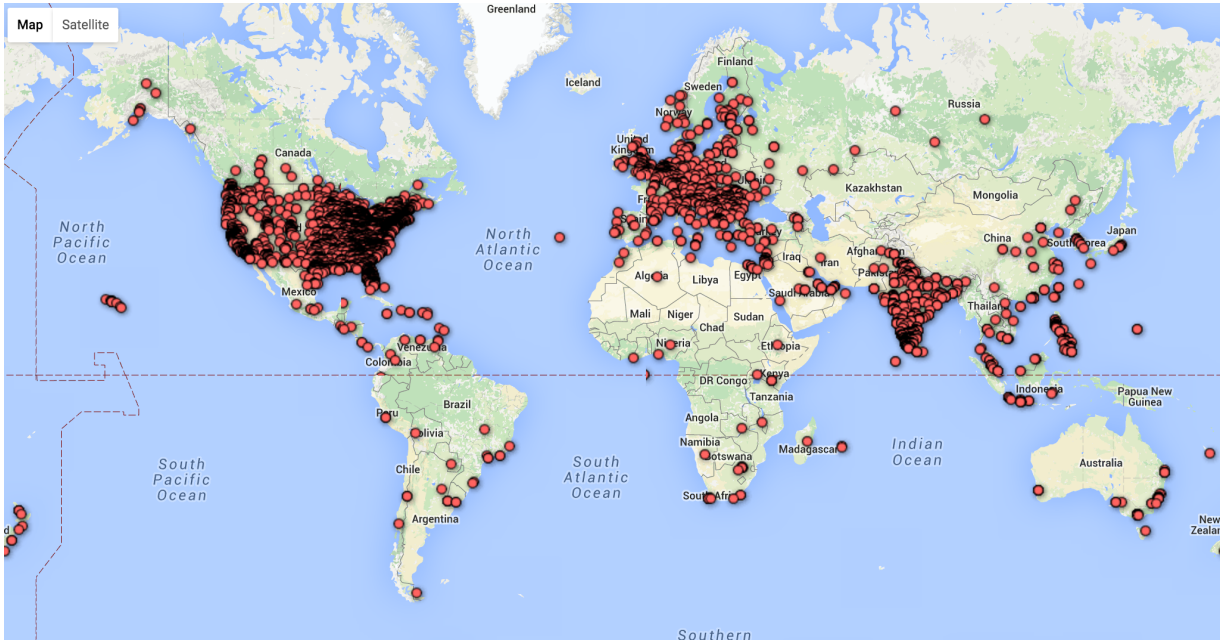


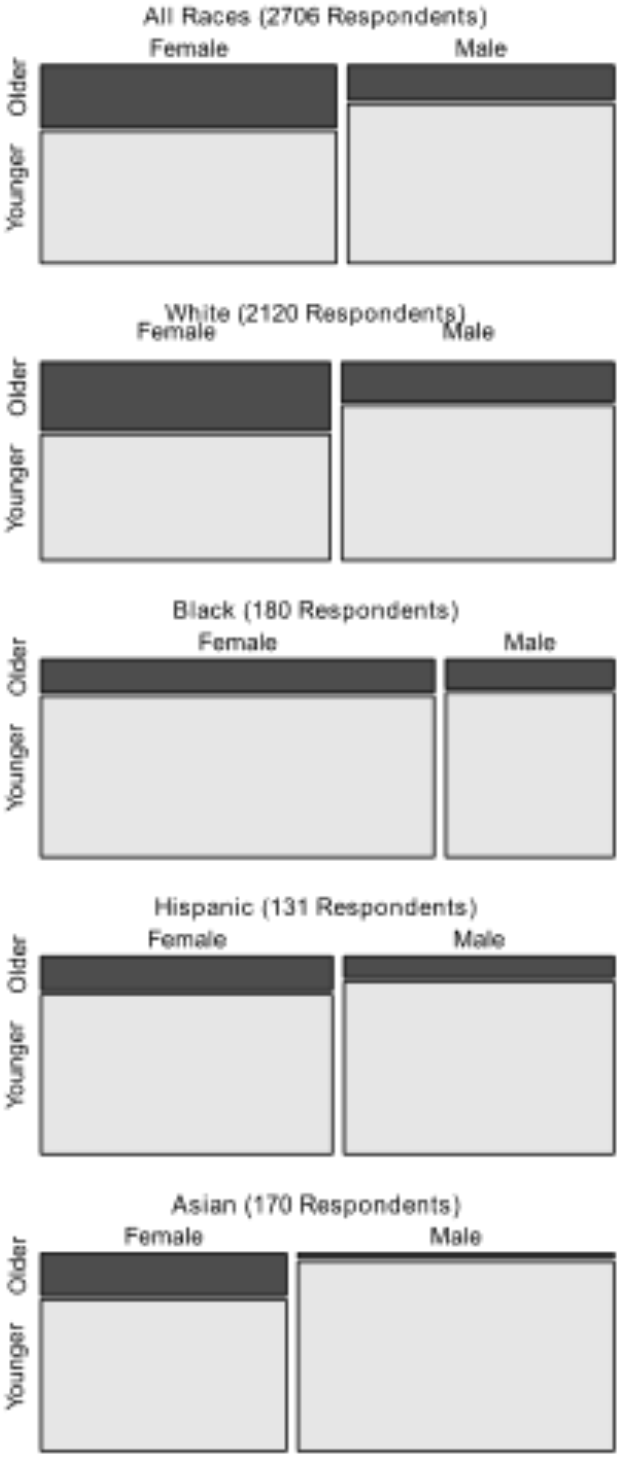

I turk for to turk brings CHANGE
I TURK FOR EXTRA CASH
I TURK TO EARN $ FOR BOOKS!
I Turk to Battle Insomnia
I TURK for CASH!
I turk for incredibly small quantities of cash!
I Turk for the LULZ ;-P
I TURK FOR CHRISTMAS
I TURK FOR MAKING MONEY IN MY LEISURE TIME
I TURK FOR DRUG MONEY
JUST KIDDING
iTURK to support my shopping habit
I TURK FOR CASH
I Turk for CHRISTMAS!
I TURK BECAUSE I LIKE TO DO WEIRD THINGS

Readings

- Crump et al. (2013)
  - Gorgeous set of replications of classic cogpsych findings on turk, nuanced discussion
- Optional: Buhrmester et al. (2011)
  - Very good intro to this topic, very short!
  - Norming data in the supplemental materials
- Optional: Mason & Suri (2012)
  - More in-depth article
  - Lots of important information about how to use the site
  - Mturk hasn't changed in years (Amazon doesn't care about it)
- Please read these if you are new to crowdsourcing!

Demographics





Geography:

- Limited to US *requesters* (for tax purposes)
- Workers from anywhere (mostly US & India)
  - Limited by tax verification etc.
  - Most studies choose US-only

Newer: Huff & Tingley (2014)

http://www.behind-the-enemy-lines.com/2010/03/new-demographics-of-mechanical-turk.html

Superturkers

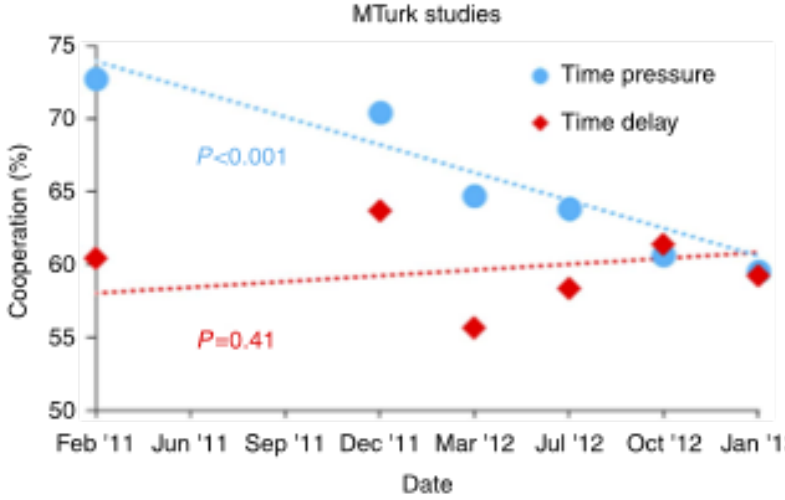


Figure 3 | The effect of time pressure on cooperation decayed as the MTurk subject pool became more experienced. Percentage of endowment contributed to the public good is shown, for subjects that obeyed the time constraint. For visualization periods, date is rounded down into 100-day bins. To make cooperation levels comparable across studies, only data from US residents is plotted, as there are different baseline average levels of cooperation across subject pools from different countries, and later studies restricted to the United States only. Note that the first bin includes RGN Study 6 as well as two other MTurk studies run shortly thereafter. Trend lines and *P*-values generated by linear regression with robust s.e. values clustered on IP address, not including controls.

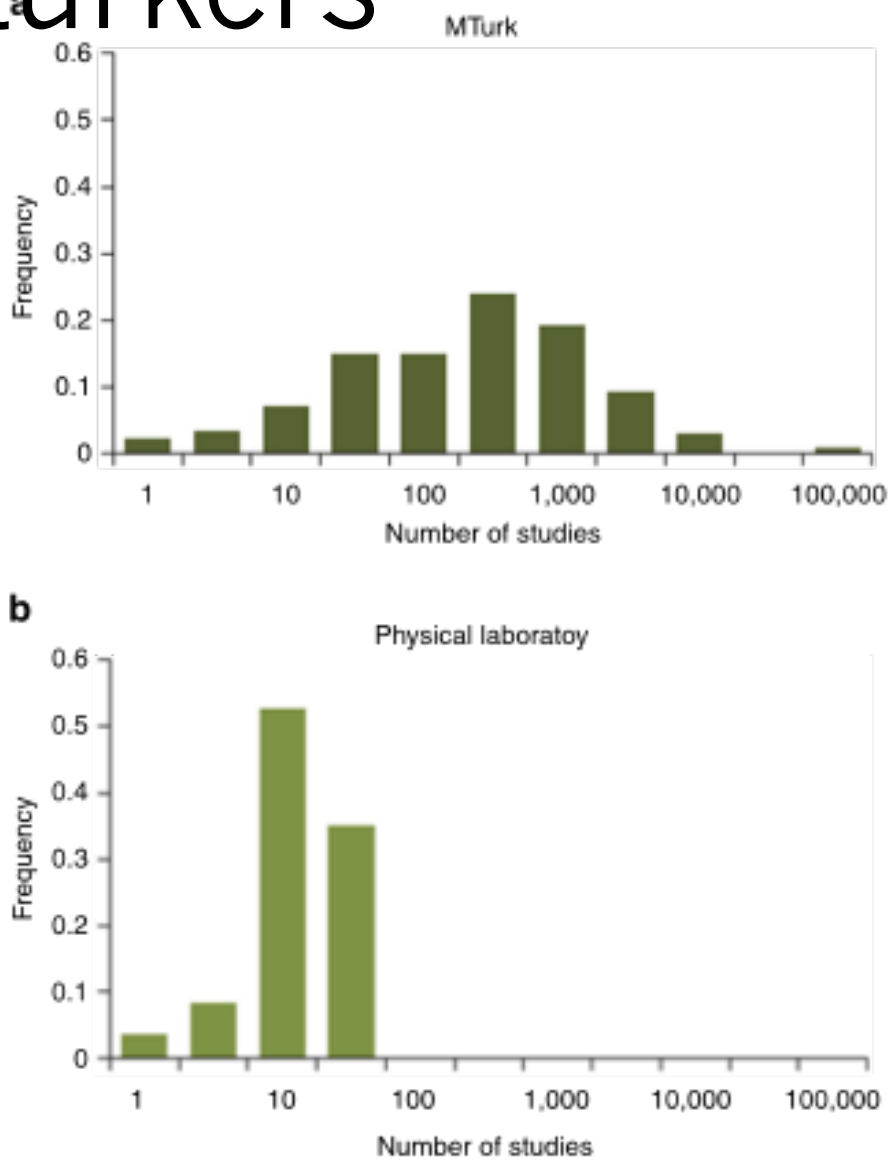


Figure 2 | Subjects on MTurk are vastly more experienced than subjects in conventional physical laboratories. Shown is the distribution of self-reported total number of studies completed by a sample of 291 MTurk subjects (**a**) and 118 subjects from the Harvard Decision Sciences Laboratory subject pool (**b**).

Rand et al. (2013)

Taking the worker's perspective

amazonmechanical turk
Artificial Artificial Intelligence

MTURK | hot | new | rising | controversial | top | gilded | wiki | promoted

1 9 **Watercooler** First 500 (self.mturk)
submitted 12 hours ago by Uhfgood
6 comments share pocket

2 0 **Help/Advice** Best times to check /r/HitsWorthTurkingFor (self.mturk)
submitted 3 hours ago by Uhfgood
2 comments share pocket

3 1 **Help/Advice** Technical problem with surveys? (Pictures/Choices not showing.) (self.mturk)
submitted 4 hours ago by Office_Zombie
3 comments share pocket

4 1 **Help/Advice** Taxes... (self.mturk)
submitted 4 hours ago by crg_26
8 comments share pocket

5 2 **Qual/HIT Question** Sergey Schmidt hits broken for anyone else? (self.mturk)
submitted 7 hours ago by GodofBoom
4 comments share pocket

6 89  **Watercooler** woke up, checked my email, and... HOLY FUCK (i.imgur.com)
submitted 1 day ago by fredrickthegreat
18 comments share pocket

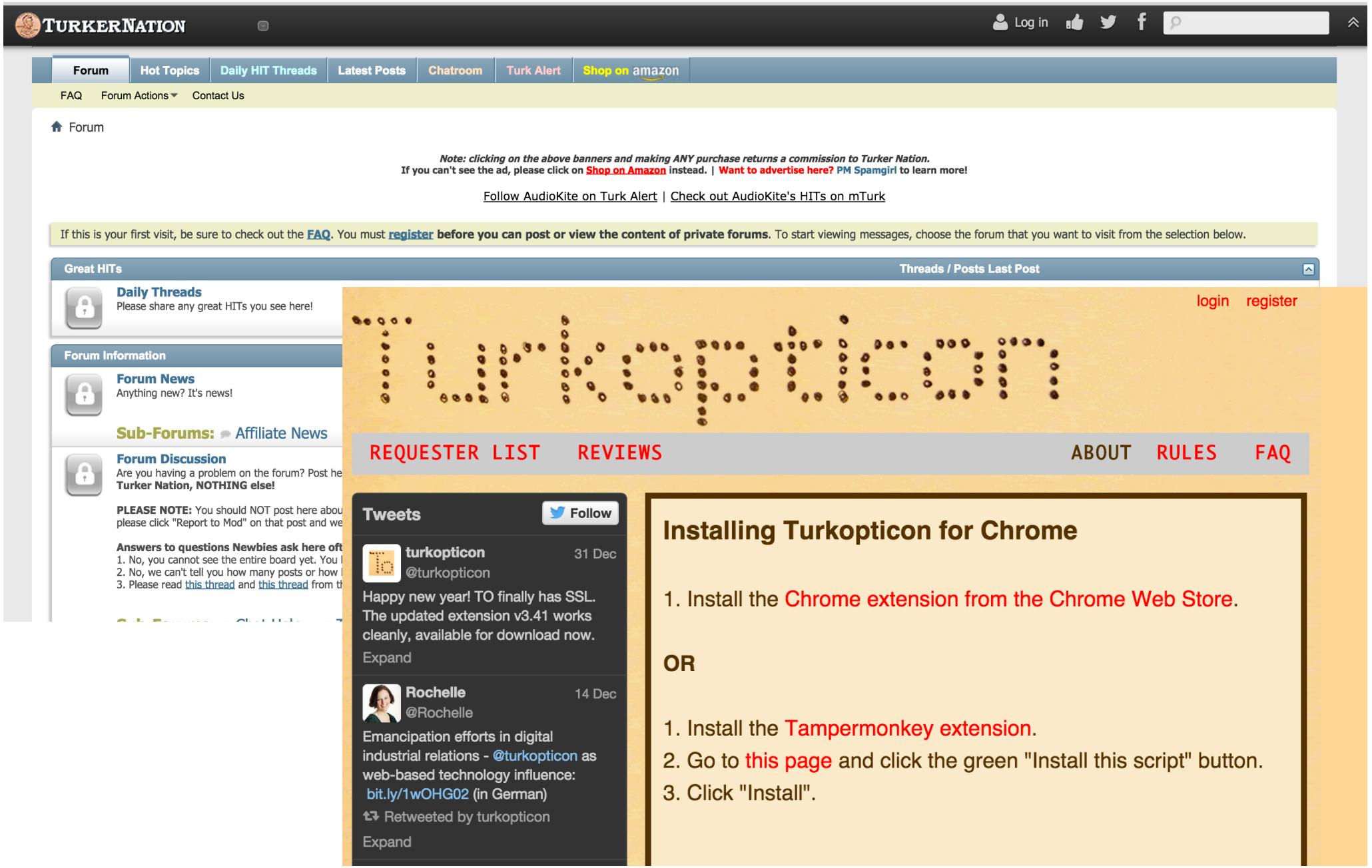
7 1 **Qual/HIT Question** Any issues with 'We-Pay-You-fast' HITs? (self.mturk)
submitted 16 hours ago by sambu123
3 comments share pocket

8 8 **Watercooler** Mturkgrind has switched to Xenforo! (self.mturk)
submitted 1 day ago by zingyMTG
3 comments share pocket

9 5  **Scripts/Software** New mturk Tool - Great MTurk Launcer AHK (more info in comments. (i.imgur.com)
submitted 1 day ago by lotkrotan
4 comments share pocket

10 1 **Account Issues** "Worker Account Halted for 5 minutes" (self.mturk)
submitted 1 day ago by joncholliday

Taking the worker's perspective



Ethics of working with turk

DYNAMO Wiki

Page | Discussion | Read | View source | View history | Search

Guidelines for Academic Requesters

DYNAMO

Home | Vote on new ideas! | How it works | Forum | Sign in/Register

Powering change on MTurk

We are a community of 542 Turkers and growing...!

We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers

Niloufar Salehi[1], Lilly C. Irani[2], Michael S. Bernstein[1], Ali Alkhatib[1], Eva Ogbe[1], Kristy Milland[3, 4], Clickhappier[4, 5]
[1]Stanford University, [2]UC San Diego, [3]Ryerson University, [4]We Are Dynamo, [5]MTurkGrind
niloufar@cs.stanford.edu, lirani@ucsd.edu, {msb, ali.alkhatib, eoogbe}@cs.stanford.edu,
kmilland@ryerson.ca, clickhappier@wearedynamo.org

Project of Stanford CS HCI Lab (Salehi et al. 2015)

Norming

Table 3. *Reliability Alphas Between Samples*

	MTurk				Standard Internet
Scale	2 cents	10 cents	50 cents	Average	
SDO	.93	.89	.93	.92	.91
RSES	.90	.90	.91	.90	.91
BFI Extraversion	.86	.88	.85	.86	.87
BFI Agreeableness	.76	.73	.82	.77	.77
BFI Conscientiousness	.86	.86	.82	.85	.77
BFI Emotional Stability	.89	.89	.87	.88	.85
BFI Openness	.80	.90	.80	.83	.79
Clarity	.87	.92	.93	.91	.90
Avoidant	.81	.85	.81	.82	.84
Anxious	.85	.81	.81	.82	.81
SLCS	.93	.92	.93	.93	.92
Mean	.87	.88	.87	.87	.86

Note. For MTurk data, N = 160 for all scales except N = 74 for Clarity, Avoidant, Anxious, and SLCS. All MTurk data were collected over a two-week period from January through February 2010. BFI = Big Five Inventory, SDO = Social Dominance Orientation, RSES = Rosenberg Self-Esteem Scale, SLCS = Self-Liking and Competence Scale, AAQ= Adult Attachment Questionaire, Comparison samples for the BFI, SDO, and RSES came from a large sample of college undergraduates (N = 1822) collected by Gosling, Rentfrow, and Swann (2003). Comparison samples for self-concept clarity and AAQ consisted of 116 participants collected online by the second author through posting ads on Yahoo groups, Facebook, and Craigslist during February 2009.

Buhrmester et al. (2011)

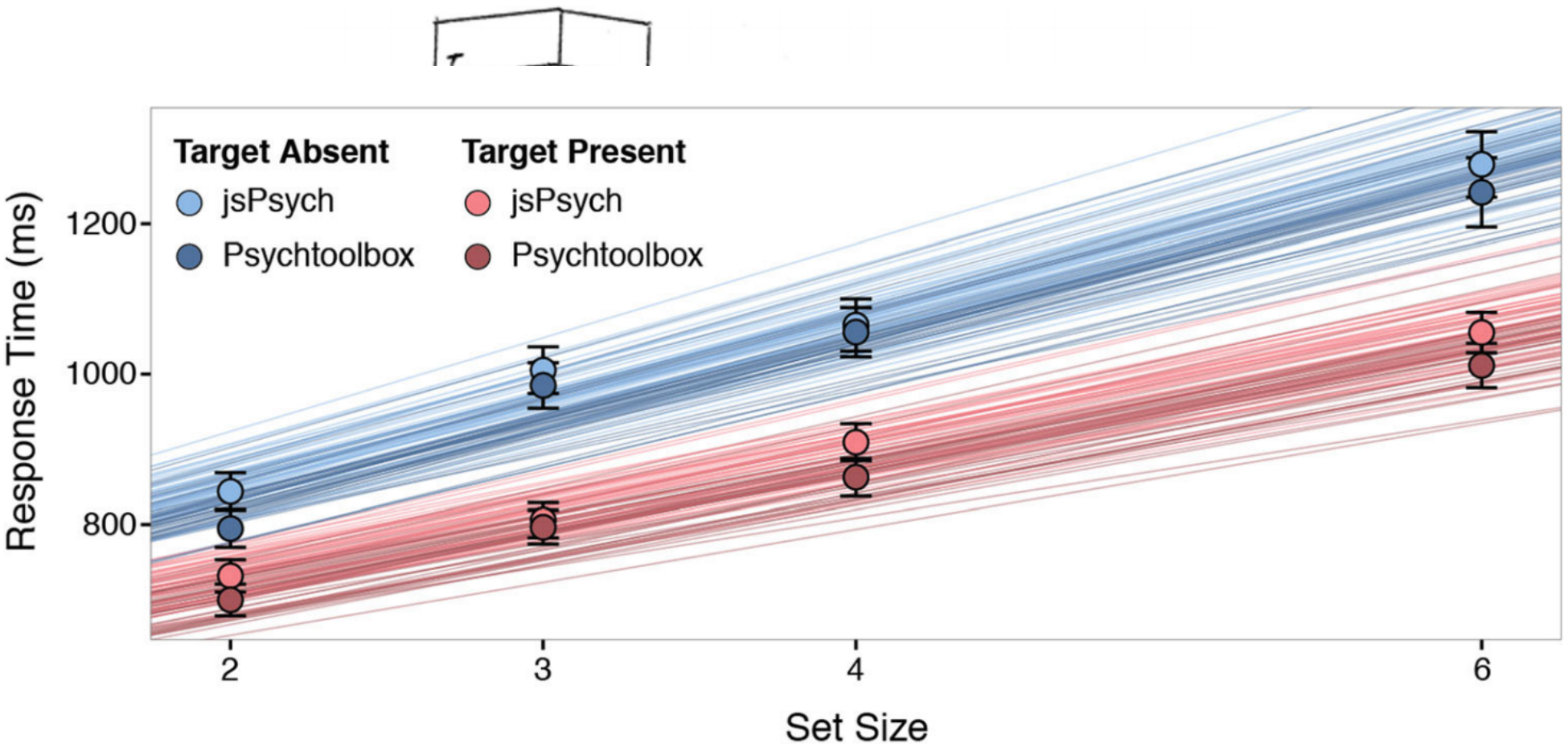
Norming

Judgment & Decision-Making Tasks

	Mechanical Turk	Midwestern university	Internet boards
Asian Disease			
% 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
Linda problem			
% Conjunction Fallacy	72.2%	78.3%	64.4%
Physician problem			
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

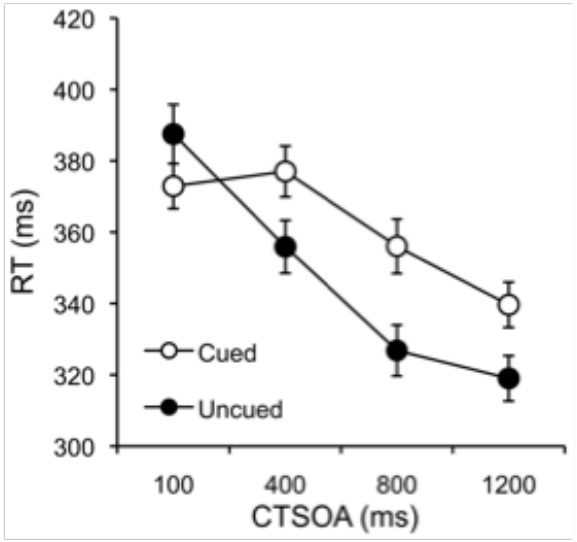
Paolacci et al. (2010)

De Leeuw & Motz (2015)

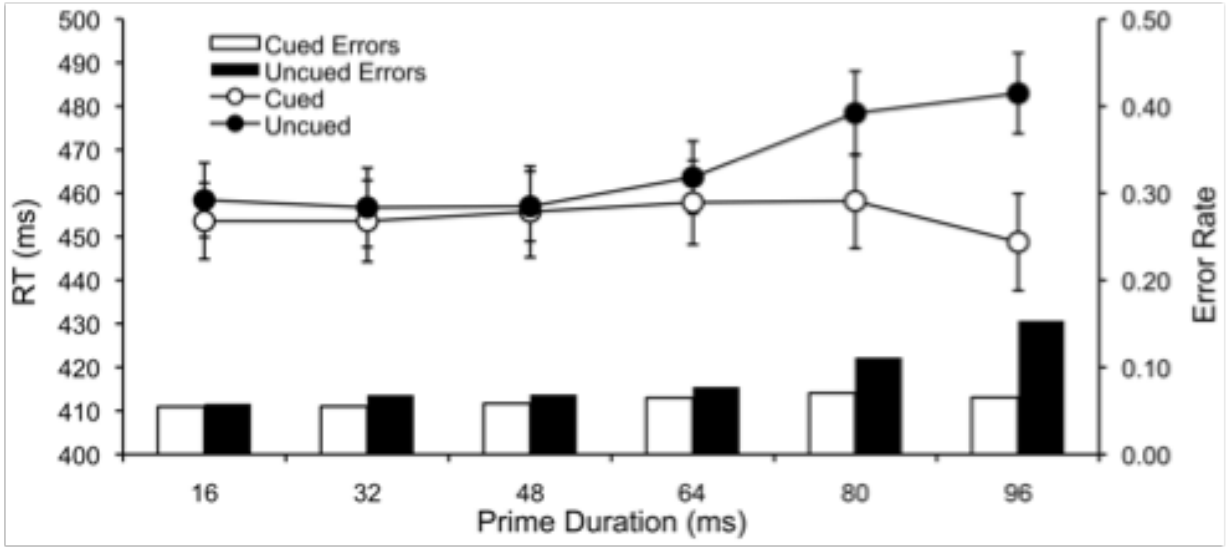


Crump et al. (2013)

Visual cueing with tiny RT differences



Masked priming doesn't work with short



Recommendations:

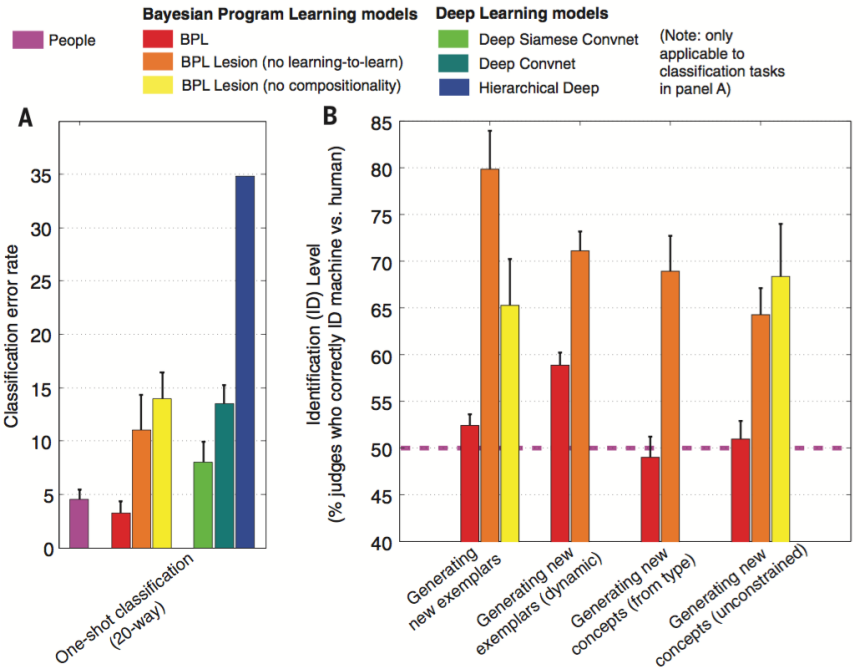
- Choose tasks appropriately
- Pay well
- Make tasks fun
- Make sure participants understand instructions

RESEARCH ARTICLES

COGNITIVE SCIENCE

Human-level concept learning through probabilistic program induction

Brenden M. Lake,[1]* Ruslan Salakhutdinov,[2] Joshua B. Tenenbaum[3]



4 + 40 + 18 + 150 + 60 + 150 + 36 + 120 + 21 + 125

Generating one-shot learning stimuli

one-shot classification

New drawing production

Visual Turing test highly trained model

Visual Turing test less trained

Visual Turing test (dynamic)

Drawing new exemplars

Judging new exemplars

Drawing new concepts

Judging new concepts

Outstanding issues?