Tricks for making your MTurk experience a more successful one

Dan Navarro
First, an overview of my workflow...
code experiment

"type type type..."
code experiment

post on Google

“Woohoo, I’m on the internet”
code experiment

post on Google

create MTurk HIT

“Hey Turkers, go to my webpage and I will give you moneys!”
code experiment
post on Google
create MTurk HIT
run pilot

A few people actually do go and I do give them the moneys...
code experiment

post on Google

create MTurk HIT

run pilot

run the full study

Yeah, that’s the same thing but with bigger N
code experiment

post on Google

create MTurk HIT

run pilot

run the full study

monitor email & TO

“talk, talk, talk…”
code experiment
→
post on Google
→
create MTurk HIT
→
run pilot
→
run the full study
→
monitor email & TO

(let's be honest, shall we?)

fix screw ups
The workflow of the Turker (the part that I control, anyway)
“go to this website to do the HIT”
“okay, I’m here at the website and it’s telling me what will happen”
“yeah, yeah, demographic information”
Instructions

This is a study looking at how people learn new words. During the task you’ll be shown some simple pictures and asked to guess how each picture is labelled in an alien language. To give you a feel for how the task works, we’ll walk you through an example.

The first thing you might be shown is a pair of images like this:

Kap  ???

In this example, you know that the square on the left is called a "Kap", and you’re being asked to guess whether the triangle on the right is also called a "Kap", or whether it has a different label (i.e., a "New" word).

After you make your guess, you’ll usually be given some feedback telling you the right answer:

Kap  Bom

In this case, the triangle is called a "Bom." If you’d guessed that it was a "Kap" you were wrong. If you’d guessed that it had a "New" label you were right.
Correctly answer a few questions about the study…

Check Your Knowledge!

What am I supposed to do?
- Guess what category each picture belongs to.
- Click some random buttons until the task ends.

How often will new pictures always belong to the same category as old ones?
- Sometimes. That’s what the "New" button is there for.
- Never. Every new picture belongs to the same category as an old one.
- Always. Every new picture will belong to a completely new category.

How many tasks do I have to do?
- 3 very long tasks
- 12 very long tasks
- 12 very short tasks

Submit Answers
Do some super-exciting category learning!

Which category does this belong to?

Fep  Led  New
Your completion code is **3214**. To receive payment for the HIT, return to the Amazon Mechanical Turk page and enter this code. Please contact us if something goes wrong and we’ll fix it as quickly as possible.

(Also... if you want to know what this was all about, click here after copying your completion code)
What was that about, anyway?

All of the studies posted on Turk by the “Adelaide Computational Cognitive Science” group are associated with the CCS lab at the University of Adelaide. This particular study is part of joint project with Carnegie Mellon University. We’ve been running quite a few of these little studies with the help of all you lovely folks on Amazon Mechanical Turk, and we often get people wanting to know more about the work. Like all scientific research it's a bit complicated (and often boring), but here's the basic idea...

What’s the big picture?

We’re a bunch of psychologists interested in fundamental questions about human reasoning, learning and decision making. Although we’re psychologists, a lot of our work is informed by thinking about some of the big issues in artificial intelligence and philosophy. One of the big puzzles that we’re trying to resolve is how people – for all our shortfalls – manage to learn so much from so little.

There’s a pretty big puzzle here: on the one hand, research in psychology and economics has shown that people often make very basic mistakes on “simple” reasoning problems. Yet on the other hand, when you look at the current state of artificial intelligence research, it’s clear that there are so many things that people do superbly well, making clever guesses effortlessly with minimal guidance. If you’ve ever tried to build a smart machine you start getting a feel for just how clever actual humans can be sometimes. Along the same lines, you can look at the philosophical literature and you realize that, again, so many of the things we do automatically (e.g., make “common sense” inferences) are just riddled with logical puzzles, yet people seem to do pretty well at these tasks anyway.

To put it crudely: if we’re so stupid, then why are we so smart?
Go back to MTurk and type the completion code

Instructions
1. Please open a new window in your browser and go to the webpage: http://species-count.appspot.com/
2. Follow the instructions on the screen there
3. At the end of the task, you will be given a code. Please paste it in the space below to successfully complete this HIT.

If you would like, provide feedback. (Filling this out is not necessary to get the HIT approved).
Tell the researchers how much I love them
Click here and wait for the moneys to appear
Some assorted observations about making this process work for you
(1) Ethics helps!

- Not surprisingly, people will do better work if you have earned their trust
- Ethical behaviour pays off in the form of good data :-)

<table>
<thead>
<tr>
<th>AMT Requester Name &amp; ID</th>
<th>Ratings [?] (averaged)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Adelaide CompCogSci</strong></td>
<td>FAST: 4.55 / 5</td>
</tr>
<tr>
<td>A2UKL52R7I6BX7</td>
<td>PAY: 4.42 / 5</td>
</tr>
<tr>
<td>HIT Group »</td>
<td>FAIR: 4.57 / 5</td>
</tr>
<tr>
<td></td>
<td>COMM: 4.07 / 5</td>
</tr>
</tbody>
</table>
Thanks for accepting the HIT. "How many species?" is a short psychological study investigating how people make guesses on the basis of limited information. It involves the following steps:

1. We ask for demographic information (not connected to your Amazon ID).
2. The study then explains the guessing game. You will need to pass a short test to check that you understand how the study works.
3. Next comes the experiment itself.
4. At the end, we'll give you the completion code you need to get paid for the HIT.

Total time taken should be about 15 minutes. Please don't use the "back" button on your browser or close the window until you reach the end and receive your completion code. This is very likely to break the experiment and may make it difficult for you to get paid. However, if something does go wrong, please contact us! When you're ready to begin, click on the "start" button below.
(3) Provide a field for Turkers to give generic comments
It’s **really** useful in the pilot study

The page with the code seems to have a next but...

The next button on the last page doesn’t work.
(4) Try to make your experiments *fun*

- This shouldn’t be too hard. You’re competing with marketing firms and image tagging exercises
- Almost anything is more fun than image tagging…
What was that about, anyway?

All of the studies posted on Turk by the “Adelaide Computational Cognitive Science” group are associated with the CCS lab at the University of Adelaide. This particular study is part of joint project with Carnegie Mellon University. We’ve been running quite a few of these little studies with the help of all you lovely folks on Amazon Mechanical Turk, and we often get people wanting to know more about the work. Like all scientific research it’s a bit complicated (and often boring), but here’s the basic idea...

What’s the big picture?

We’re a bunch of psychologists interested in fundamental questions about human reasoning, learning and decision making. Although we’re psychologists, a lot of our work is informed by thinking about some of the big issues in artificial intelligence and philosophy. One of the big puzzles that we’re trying to resolve is how people – for all our shortfalls – manage to learn so much from so little.

There’s a pretty big puzzle here: on the one hand, research in psychology and economics has shown that people often make very basic mistakes on “simple” reasoning problems. Yet on the other hand, when you look at the current state of artificial intelligence research, it’s clear that there are so many things that people do superbly well, making clever guesses effortlessly with minimal guidance. If you’ve ever tried to build a smart machine you start getting a feel for just how clever actual humans can be sometimes. Along the same lines, you can look at the philosophical literature and you realize that, again, so many of the things we do automatically (e.g., make “common sense” inferences) are just riddled with logical puzzles, yet people seem to do pretty well at these tasks anyway.

To put it crudely: if we’re so stupid, then why are we so smart?
(6) Make sure the site is hosted by someone competent like Google…

The sudden spike from 200 Turkers hitting the server at the same time broke our old hosting service.
Rest of the talk…

• Don’t screw up completion codes
• Do use instruction checks, but be smart about how you use attention checks
• Turkers aren’t naive subjects
• Turkers self-select into your study
• The testing environment is uncontrolled
Don’t screw up your completion codes
Your completion code is 3214. To receive payment for the HIT, return to the Amazon Mechanical Turk page and enter this code. Please contact us if something goes wrong and we'll fix it as quickly as possible.

(Also... if you want to know what this was all about, click here after copying your completion code)

Completion code is a random number, highlighted to make it obvious to the Turker

Link to debrief screen opens in new tab so that the Turker doesn’t lose the completion code!
The prevalence of 0192 is due to (a) a bug in my code, and (b) the fact that I wrote my experiment defensively so that if the Javascript broke, the number 0192 was written into the HTML. So there was always something the Turker could submit.

<table>
<thead>
<tr>
<th>HIT ID</th>
<th>Worker ID</th>
<th>Lifetime/Approval Rate</th>
<th>Q2</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (10/10)</td>
<td>8645</td>
<td>None.</td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (5/5)</td>
<td>3832</td>
<td>Thank you and good luck with your research.</td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (2/2)</td>
<td>4881</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (5/5)</td>
<td>0192</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (11/11)</td>
<td>9266</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (2/2)</td>
<td>2794</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (4/4)</td>
<td>8648</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (7/7)</td>
<td>4012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (4/4)</td>
<td>0192</td>
<td>good experience</td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (4/4)</td>
<td>0192</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (1/1)</td>
<td>3387</td>
<td>Thank you.</td>
<td></td>
</tr>
<tr>
<td>JVRD13MFWMQ9MNK5D</td>
<td>100% (3/3)</td>
<td>0192</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
There it is.
Paranoia pays off

Your completion code is
0192.
To receive payment for the HIT, return to the Amazon Mechanical Turk page and enter this code. Please contact us if something goes wrong and we'll fix it as quickly as possible.
Instruction checks & attention checks
Instruction checks

- Instruction checks:
  - Questions at the end of the instructions intended to verify that participants understood the task
  - If they get it wrong they aren’t excluded, they’re just asked to read the instructions again

- We use these a lot
  - They make the instructions more interactive
  - Call attention to the important aspects of the task
  - Helps stop participants “skimming” the instructions
  - *(In hindsight, all our lab studies should have done this)*
Check Your Knowledge!

What am I supposed to do?
- Guess what category each picture belongs to.
- Click some random buttons until the task ends.

How often will new pictures always belong to the same category as old ones?
- Sometimes. That's what the "New" button is there for.
- Never. Every new picture belongs to the same category as an old one.
- Always. Every new picture will belong to a completely new category.

How many tasks do I have to do?
- 3 very long tasks
- 12 very long tasks
- 12 very short tasks

Submit Answers
Most people get it right first time, or maybe make an error once, but it seems to help
Attention checks

• Instruction checks are, in some sense, part of the cover story for the task.

• An attention check (AC) is different

• An AC is a question / stimuli added during the task to check that participants are “paying attention

• If they fail the attention checks the data are excluded from analysis
I have mixed feelings about AC items

• Good example:
  • In a 2AFC task, include a few items that are very easy
  • Participants will notice, but it doesn’t feel out of place
  • It’s just part of the task

• Bad example:
  • In a survey about racism, ask people what 2+2 equals
  • It’s so obviously an attention check item. Sophisticated Turkers will spot it a mile away even if they’re not really paying much attention to the task
  • It’s a much weaker check than people think
Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys

Adam J. Berinsky  Massachusetts Institute of Technology
Michele F. Margolis  Massachusetts Institute of Technology
Michael W. Sances  Massachusetts Institute of Technology

Good survey and experimental research requires subjects to pay attention to questions and treatments, but many subjects do not. In this article, we discuss “Screeners” as a potential solution to this problem. We first demonstrate Screeners’ power to reveal inattentive respondents and reduce noise. We then examine important but understudied questions about Screeners. We show that using a single Screener is not the most effective way to improve data quality. Instead, we recommend using multiple items to measure attention. We also show that Screener passage correlates with politically relevant characteristics, which limits the generalizability of studies that exclude failers. We conclude that attention is best measured using multiple Screener questions and that studies using Screeners can balance the goals of internal and external validity by presenting results conditional on different levels of attention.
Here is a typical example of an attention check “screener” question...

When a big news story breaks people often go online to get up-to-the-minute details on what is going on. We want to know which websites people trust to get this information. We also want to know if people are paying attention to the question. To show that you’ve read this much, please ignore the question and select ABC News and The Drudge Report as your two answers.

When there is a big news story, which is the one news website would you visit first? (Please only choose one)

- New York Times website
- Huffington Post
- Washington Post website
- CNN.com
- FoxNews.com
- MSNBC.com
- The Drudge Report
- Google News
- ABC News website
- CBS News website
- NBC News website
- Yahoo! News
- The Associated Press (AP) website
- Reuters website
- National Public Radio (NPR) website
- USA Today website
- New York Post Online
- None of these websites

Berinsky et al (2014)
Here is a typical example of an attention check “screener” question...

When a big news story breaks people often go online to get up-to-the-minute details on what is going on. We want to know which websites people trust to get this information. **We also want to know if people are paying attention to the question.** To show that you’ve read this much, please ignore the question and select ABC News and The Drudge Report as your two answers.

When there is a big news story, which is the one news website would you visit first? (Please only choose one)

- [ ] New York Times website
- [ ] Huffington Post
- [ ] Washington Post website
- [ ] CNN.com
- [ ] FoxNews.com
- [ ] MSNBC.com
- [ ] The Drudge Report
- [ ] Google News
- [ ] ABC News website
- [ ] CBS News website
- [ ] NBC News website
- [ ] Yahoo! News
- [ ] The Associated Press (AP) website
- [ ] Reuters website
- [ ] National Public Radio (NPR) website
- [ ] USA Today website
- [ ] New York Post Online
- [ ] None of these websites

Berinsky et al (2014)
Tversky & Kahneman’s “gains vs losses” framing effect

<table>
<thead>
<tr>
<th>Passed Screener</th>
<th>Mortality Frame</th>
<th>Save Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic</td>
<td>64</td>
<td>37</td>
</tr>
<tr>
<td>“Sure”</td>
<td>36</td>
<td>63</td>
</tr>
</tbody>
</table>

Berinsky et al (2014)
Random responses by people who failed to pass the screener test

<table>
<thead>
<tr>
<th>Passed Screener</th>
<th>Mortality Frame</th>
<th>Save Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic</td>
<td>64</td>
<td>37</td>
</tr>
<tr>
<td>“Sure”</td>
<td>36</td>
<td>63</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Failed Screener</th>
<th>Mortality Frame</th>
<th>Save Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic</td>
<td>53</td>
<td>49</td>
</tr>
<tr>
<td>“Sure”</td>
<td>47</td>
<td>51</td>
</tr>
</tbody>
</table>
But because most people pass the screener, the overall effect on the data is pretty minimal

<table>
<thead>
<tr>
<th>All Respondents</th>
<th>Mortality Frame</th>
<th>Save Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic</td>
<td>62</td>
<td>39</td>
</tr>
<tr>
<td>“Sure”</td>
<td>38</td>
<td>61</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Passed Screener</th>
<th>Mortality Frame</th>
<th>Save Frame</th>
<th>Failed Screener</th>
<th>Mortality Frame</th>
<th>Save Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probabilistic</td>
<td>64</td>
<td>37</td>
<td>Probabilistic</td>
<td>53</td>
<td>49</td>
</tr>
<tr>
<td>“Sure”</td>
<td>36</td>
<td>63</td>
<td>“Sure”</td>
<td>47</td>
<td>51</td>
</tr>
</tbody>
</table>

Notes: Cell entries are column percentages. N=376 for all respondents, N=301 for passed Screener, and N=75 for failed Screener. Data Source: Study 1.
My experience… the ubiquitous “80-10-10” split

- 80% of people give great data (yellow)
- 10% of people give data that make sense in hindsight, but are weird and you’d never have predicted it (red)
- 10% of people give answers that appear random at every level (orange) … which is usually easy to detect!
Turkers are not naive subjects, so some studies are not well-suited to MTurk
If a baseball and a bat cost $1.10 together, and the bat costs $1.00 more than the ball, how much does the ball cost?
Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers

Jesse Chandler · Pam Mueller · Gabriele Paolacci

© Psychonomic Society, Inc. 2013

Abstract Crowdsourcing services—particularly Amazon Mechanical Turk—have made it easy for behavioral scientists to recruit research participants. However, researchers have overlooked crucial differences between crowdsourcing and traditional recruitment methods that provide unique opportunities and challenges. We show that crowdsourced workers are human by asking them to transcribe distorted images of words, while simultaneously digitizing illegible portions of books (von Ahn, Maurer, McMillen, Abraham, & Blum, 2008), Galaxy Zoo (www.galaxyzoo.org) solicits “citizen-scientists” to view and classify astronomical images (Lintott et al., 2008), and Foldit (www.fold.it) harnesses the
Turkers aren’t naive subjects

- There are lots of psych studies on MTurk
- Don’t run it on MTurk if:
  - The task is very common on MTurk already; and
  - The task is subject to practice effects

- Example:
  - There’s zero point in running the Cognitive Reflection Test on MTurk: experienced Turkers will have seen the “bat and ball” problem more often than you have
Chandler et al (2013)

<table>
<thead>
<tr>
<th>Game</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prisoner’s dilemma</td>
<td>56%</td>
</tr>
<tr>
<td>Ultimatum game</td>
<td>52%</td>
</tr>
<tr>
<td>Trolley problem</td>
<td>30%</td>
</tr>
<tr>
<td>P-beauty contest</td>
<td>7%</td>
</tr>
<tr>
<td>Dictator game</td>
<td>0%</td>
</tr>
</tbody>
</table>
Turkers aren’t randomly assigned to your study. They will be following you! (If they like you)
Turkers track requesters and tasks

<table>
<thead>
<tr>
<th>Activity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>read MTurk blogs</td>
<td>28%</td>
</tr>
<tr>
<td>follow requesters</td>
<td>55%</td>
</tr>
<tr>
<td>follow academic requesters</td>
<td>33%</td>
</tr>
</tbody>
</table>

Chandler et al (2013)
They have the technology...

Turk Assist is a Chrome extension that makes working on Amazon’s Mechanical Turk a little easier.

It’ll keep track of the HITs that you’ve worked on, let you block Requesters and HITs that you don’t want to see, and it can even notify you when new HITs are posted by your favorite requesters.
One person reviewing us for four separate HITS

HIT(s):
Learn to categorize patterns (~15 minutes): Make some guesses about the next observation (15 mins): A fishing game where you need to maximize the amount of fish you catch (5-10 minutes): Learning alien words (5-10 mins)

took ~11 minutes
payment is $1
approved in 2 hours

----------------------

HIT "Make some guesses about the next observation (15 mins)"
did on October 7
took ~9 minutes
payment is $1.5
approved in 2 hours

----------------------

HIT A fishing game where you need to maximize the amount of fish you catch (5-10 minutes).
did on November 7
took ~5 minutes
payment is $0.75
approved same day

----------------------

HIT Learning alien words (5-10 mins)
did on November 23
took 5 minutes
payment is $0.75
approved within 1 hour

This review was edited by the author Wed Jan 21 05:50 PST.
This review was edited by the author Sun Nov 23 18:15 PST.
This review was edited by the author Sun Nov 23 17:15 PST.
This review was edited by the author Fri Nov 07 16:39 PST.
This review was edited by the author Fri Nov 07 13:25 PST.
This review was edited by the author Tue Oct 07 18:13 PDT.
This review was edited by the author Tue Oct 07 17:02 PDT.
This review was edited by the author Sat Sep 20 16:30 PDT.

Sep 20 2014 flag | comment
It’s very common: at this point, most people doing our HITs have seen us before.
What does that mean?

- If your study is one where you can’t have repeat participants then you need to track this
What does that mean?

• If your study is one where you can’t have repeat participants then you need to track this.

• High tech solution:
  • Install the command line tools, use them to define a “qualification”. It’s an advanced trick. See Chandler et al (2013) for details.
What does that mean?

• If your study is one where you can’t have repeat participants then you need to track this

• High tech solution:
  • Install the command line tools, use them to define a “qualification”. It’s an advanced trick. See Chandler et al (2013) for details

• Low tech solution:
  • Ask them not to repeat, but be prepared for many of them to (honestly) forget what they did before
  • Inspect the worker IDs of those that complete the second task and exclude them (after paying, of course)
The environment is not controlled, and the testing machine is not controlled (sometimes this matters)
Turkers aren’t in a lab, so if you require a controlled environment then be careful*

<table>
<thead>
<tr>
<th>Activity</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>with other people</td>
<td>27%</td>
</tr>
<tr>
<td>listening to music</td>
<td>14%</td>
</tr>
<tr>
<td>watching TV</td>
<td>18%</td>
</tr>
<tr>
<td>chatting online</td>
<td>6%</td>
</tr>
</tbody>
</table>

Chandler et al (2013)

* For my studies I usually don’t want a controlled environment. The distribution above looks a lot like the conditions under which people usually have to learn, think or reason. Ecological validity is a thing that matters…
Collecting response times using Amazon Mechanical Turk and Adobe Flash

Travis Simcox • Julie A. Fiez

Published online: 14 May 2013
© Psychonomic Society, Inc. 2013

Abstract Crowdsourcing systems like Amazon Mechanical Turk (AMT) allow data to be collected from a large pool of people in a short amount of time. This use has garnered considerable interest from behavioral scientists. Several experiments conducted on AMT have focused on the timing instruments because of difficulties inherent to experimental paradigms over the Internet. This study investigated the viability of presenting stimuli and collecting response times using Adobe Flash to run ActionScript in conjunction with AMT. First, the timing properties of a phototransistor and two different microprocessors were investigated using a phototransistor and two microprocessors running under several conditions mimicking those that may be present in research using AMT. This experiment revealed some strengths and weaknesses of the timing capabilities of this method. Next, a flanker task and a lexical decision task were administered to participants in Adobe Flash. The experiment used in this paper was designed to assess the feasibility of using Adobe Flash to perform experimentation using AMT. First, AMT provides a stable pool of participants, in contrast to undergraduate participant pools in which the supply of participants varies over the course of the semester or may be very limited at smaller universities. Second, AMT workers are diverse, as compared to...
Measuring the actual length of a 1000ms ISI on MTurk

Almost always close to 1000ms. (Yay!)
Measuring the actual length of a 1000ms ISI on MTurk

But a few Turkers are running machines that can’t be trusted!
But so far we’ve had surprisingly good experiences* with RT data on MTurk

(*in tasks where RTs are expected to be fairly long)

I’d have lost money on this. I totally expected the this effect to be impossible to detect on MTurk
And that’s the workshop, folks! We’re around to chat if you like.