experimental design for linguists - pt 3

HPSG 2012

PHILIP HOFMEISTER
UNIVERSITY OF ESSEX

Acceptability Judgments

ACCEPTABILITY JUDGMENTS

- Numerous ways of eliciting 'grammaticality' judgments

ACCEPTABILITY JUDGMENTS

- Numerous ways of eliciting 'grammaticality' judgments
  - Binary judgments (YES/NO)

ACCEPTABILITY JUDGMENTS

- Numerous ways of eliciting 'grammaticality' judgments
  - Binary judgments (YES/NO)
  - Forced choice

ACCEPTABILITY JUDGMENTS

- Numerous ways of eliciting 'grammaticality' judgments
  - Binary judgments (YES/NO)
  - Forced choice
  - Likert scales (1-5 / 1-7 / etc.)

ACCEPTABILITY JUDGMENTS

- Numerous ways of eliciting 'grammaticality' judgments
  - Binary judgments (YES/NO)
  - Forced choice
  - Likert scales (1-5 / 1-7 / etc.)
  - Magnitude Estimation

ACCEPTABILITY JUDGMENTS

- Numerous ways of eliciting 'grammaticality' judgments
  - Binary judgments (YES/NO)
  - Forced choice
  - Likert scales (1-5 / 1-7 / etc.)
  - Magnitude Estimation
  - Thermometer Judgments

ACCEPTABILITY JUDGMENTS

- Numerous ways of eliciting 'grammaticality' judgments
  - Binary judgments (YES/NO)
  - Forced choice
  - Likert scales (1-5 / 1-7 / etc.)
  - Magnitude Estimation
  - Thermometer Judgments
  - Speeded vs. non-speeded

ACCEPTABILITY JUDGMENTS

- Which method to use?
- What instructions to use?
- How to analyze and treat the data?

INSTRUCTIONS

- Instructions should be in the simplest language possible and should lack any technical jargon
  - e.g. grammatical, noun, verb, phrase, semantics, syntax

SAMPLE INSTRUCTIONS

Please read each sentence, then answer the question immediately following, and rate the sentence for naturalness on a scale of 1 to 7, 1 being extremely unnatural and 7 being extremely natural. Assign higher numbers to sentences you find more natural, and lower numbers to sentences you find less natural.

SAMPLE INSTRUCTIONS

Please read each sentence, then answer the question immediately following, and rate the sentence for naturalness on a scale of 1 to 7, 1 being extremely unnatural and 7 being extremely natural. Assign higher numbers to sentences you find more natural, and lower numbers to sentences you find less natural.

We are interested in how natural you think the structure of the sentences below sound, not how plausible the meanings are. For example, "The man bit the dog" describes something less likely to happen than "The dog bit the man", but both are natural English sentences---there's nothing identifiably wrong with either sentence. So you should give them the same rating. You should provide ratings that match up with what would sound natural to you in a conversation or in reading a text, but you should NOT rely on what grammar books may have said is the right way to talk. There are no right or wrong ratings . . . we are exclusively interested in what your opinion is.

INSTRUCTIONS

- It's still not well-understood what participants are basing their ratings on
  - Participants perform similarly when asked to judge on the basis of meaningfulness vs. grammaticality (Maclay & Sleator 1960)
  - Syntactic well-formedness & interpretability are deeply intertwined

FORCED CHOICE TASKS

- Which of these is better?
  - *Which book did who write?*
  - *What did who write?*

ACCEPTABILITY JUDGMENTS

- Magnitude Estimation
  - Adapted from psychophysics research (Stevens 1975)

= 10	= ?

ACCEPTABILITY JUDGMENTS

- Magnitude Estimation
- Adapted from psychophysics research (Stevens 1975)

Gail seems to Gail like fishing	No one likes fishing
= 10	= ?

ACCEPTABILITY JUDGMENTS

- Magnitude Estimation
- Adapted from psychophysics research (Stevens 1975)

Gail seems to Gail like fishing	No one likes fishing
= 10	= 30

ACCEPTABILITY JUDGMENTS

- Magnitude Estimation
  - Potential advantages
    - larger space of judgments
    - gradience in judgments
    - each item can receive a unique score

- Magnitude Estimation
  - Participants don't use magnitudes
    - Instructions are ignored or participants are incapable of following them in linguistic judgment tasks
    - Featherston (2008): there is no zero point as in psychophysics

THERMOMETER JUDGMENTS (FEATHERSTON 2009)

- Judgments made on a linear scale with respect to 2 reference points
  - The way that the project was approaching to the deadline everyone wondered. = 20
  - The architect told his assistant to bring the new plans to the foreman's office. = 30

ACCEPTABILITY JUDGMENTS

- Any difference between different acceptability measures?
  - Weskott & Fanselow (2011)
    - *. . . dass der Mönch dem Jäger geholfen hat*
      - *that the monk-NOM the hunter-DAT helped AUX*
    - *. . . dass dem Jäger der Mönch geholfen hat.*
      - *that the hunter-DAT the monk-NOM helped AUX*

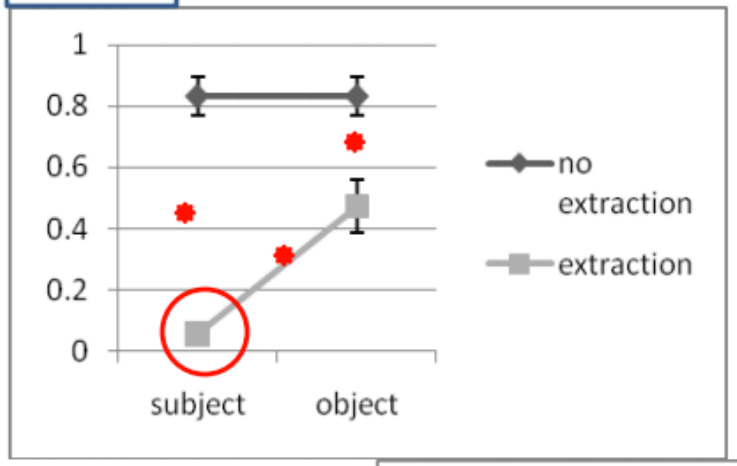
ACCEPTABILITY JUDGMENTS

- Weskott & Fanselow (2011): no notable differences between binary judgments, n-point judgments, and ME
  - Same participants rated the same items using binary judgments & ME or n-point & ME methods

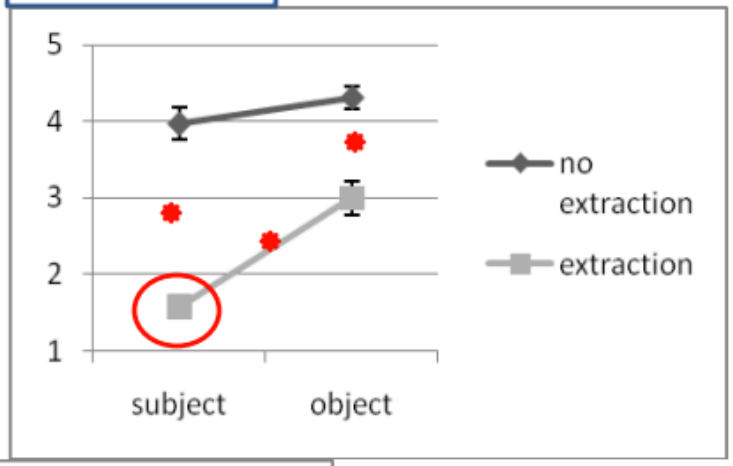
FUKUDA, MICHEL, BEECHER, & GOODALL (2010)

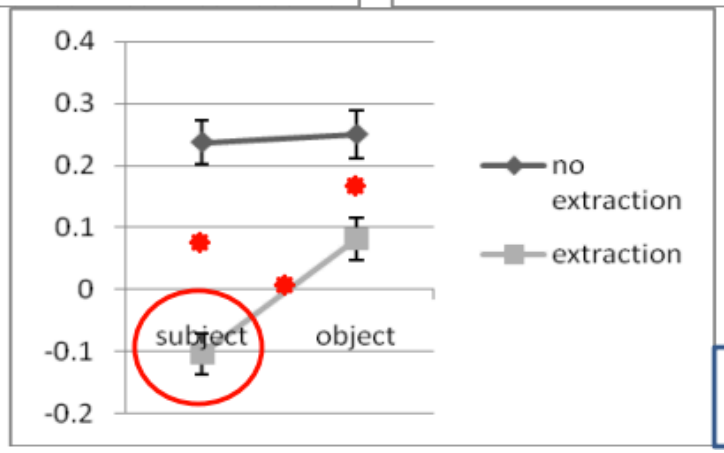
- What do you think [pictures of __] will be on the website?
- What do you think the website will post [pictures of ___]?
- Do you think the pictures of the new car will be on the website?
- Do you think the website will post pictures of the new car?

Y/N



5-point





ME

✸ $p < .05$

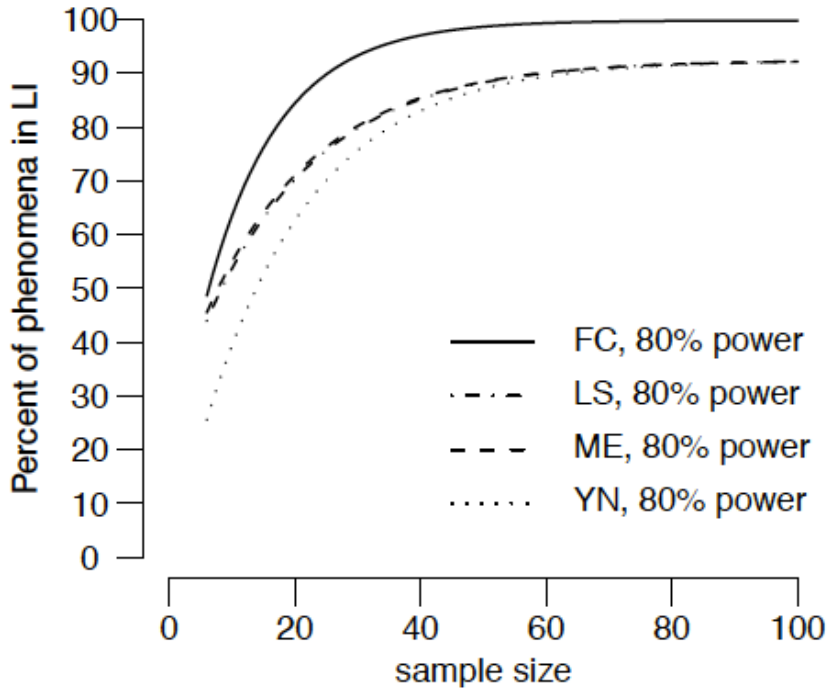
ACCEPTABILITY JUDGMENTS

- In many circumstances, it is likely true that there is little difference across methods
  - Do results hold with a large # of predicted contrasts?
  - ME creates unnecessary noise due to task demands
    - Other technical issues with ME (Sprouse 2011)

COMPARING METHODS

- What have studies to date established?
  - Significant differences between a small # of conditions are as detectable with methods that use a small scale
  - Effect sizes may even be *larger* with closed scale paradigms

COMPARING METHODS



- Note, if the question is: what methodology will maximize chances of finding an effect
  - Then, forced choice will be the best because it's design coerces participants to find a difference

COMPARING METHODS

- Let's say you find a significant effect with a YES/NO design but not Thermometer Judgments
  - Does this mean YN tasks are better?

COMPARING METHODS

- In determining what's the 'best' methodology, the better question seems to be:
  - What's the most informative & predictive method?

ANALYZING JUDGMENTS

- Experimental 'best practice' in judgment tasks is not well-established

ANALYZING JUDGMENTS

- When should participants and data points be excluded?
- How should materials be presented?
- How should ordinal scale data be treated?

PARTICIPANT & OUTLIER REMOVAL

- Easy cases
  - Participants with no variation in judgments, e.g. all judgments = 4
  - Participants with significant L2 exposure (unless that's what's being investigated)
  - Participants who cannot answer a majority of comprehension Qs correctly

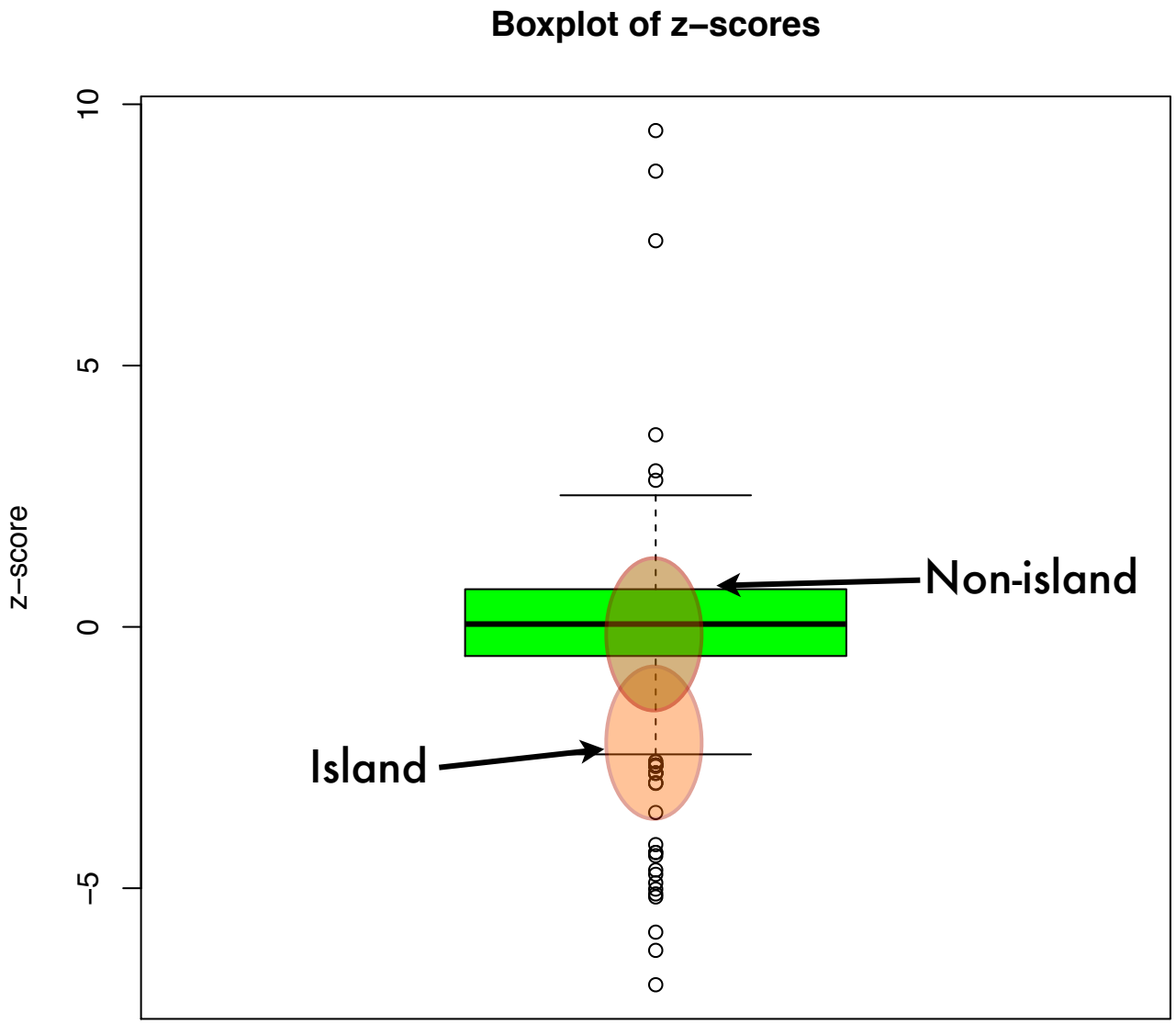
PARTICIPANT & OUTLIER REMOVAL

- Z-scores can also be used to remove outliers, e.g. > 2.5 SDs
  - Exercise caution
  - Removal of extreme outliers can reduce differences between conditions
  - Remove outliers for each condition, not for the entire dataset

PARTICIPANT & OUTLIER REMOVAL

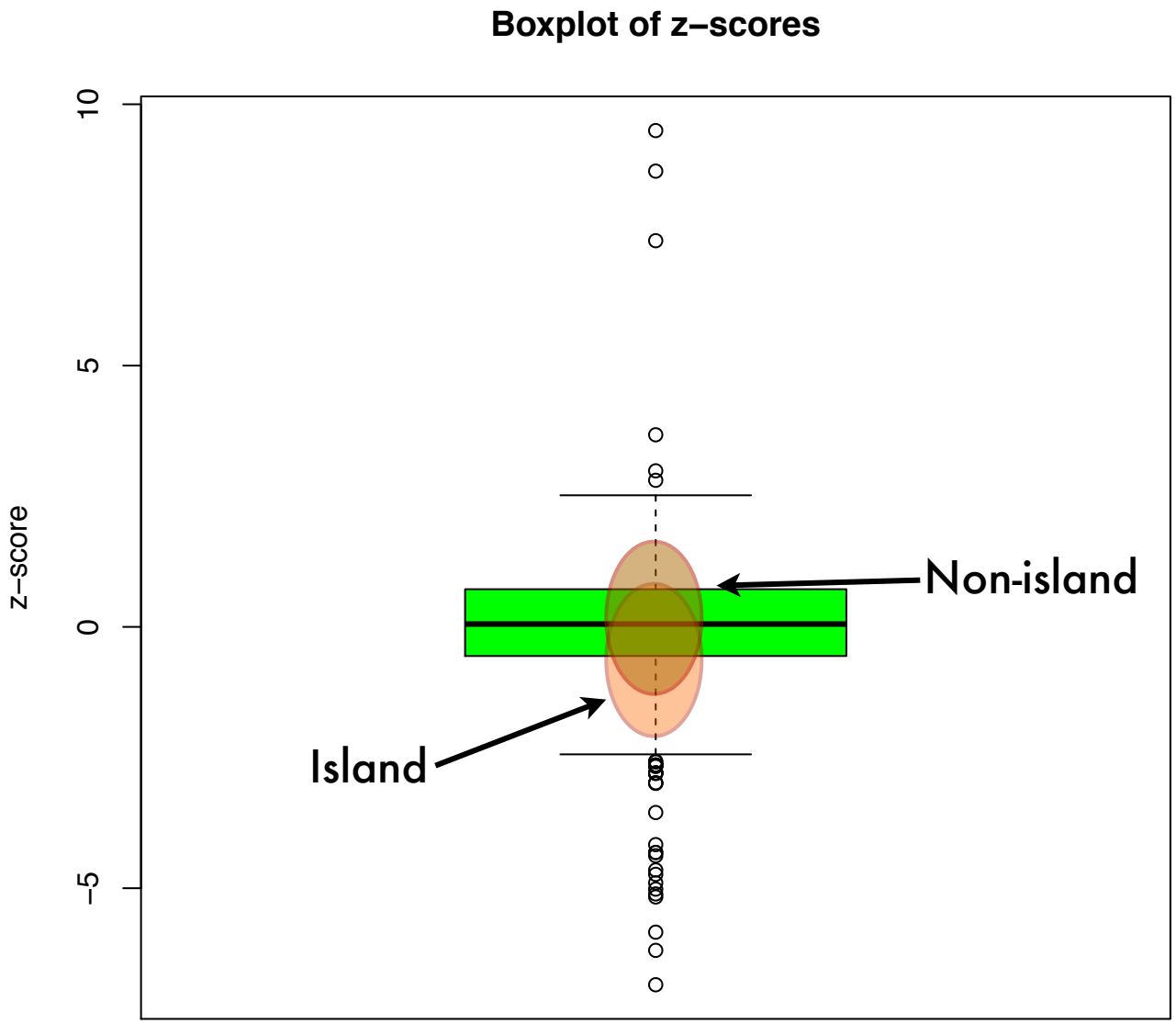
- *I saw who Emma doubted the report that we had captured in the nationwide FBI manhunt.*
- *I saw who Emma doubted that we had captured in the nationwide FBI manhunt.*

PARTICIPANT & OUTLIER REMOVAL



HPSG2012

PARTICIPANT & OUTLIER REMOVAL



HPSG2012

PRESENTATION OF MATERIALS

- Many formal acceptability experiments present sentences in their entirety and allow participants unlimited time to rate the sentences

PRESENTATION OF MATERIALS

- Excessive time for introspective allows for more and more orthogonal factors to interfere with judgments

PRESENTATION OF MATERIALS

- Alternative: Present words or sentences for a fixed period of time
  - Equalizes study time across participants

FAST OR SLOW? (SCHUTZE 1996)

> Obviously, if our goal is to examine the on-line processing of grammaticality, its effects on parsing, and so forth, then first reactions will be most useful. But if it is the status of the sentence that concerns us, it is not clear which should be preferred.

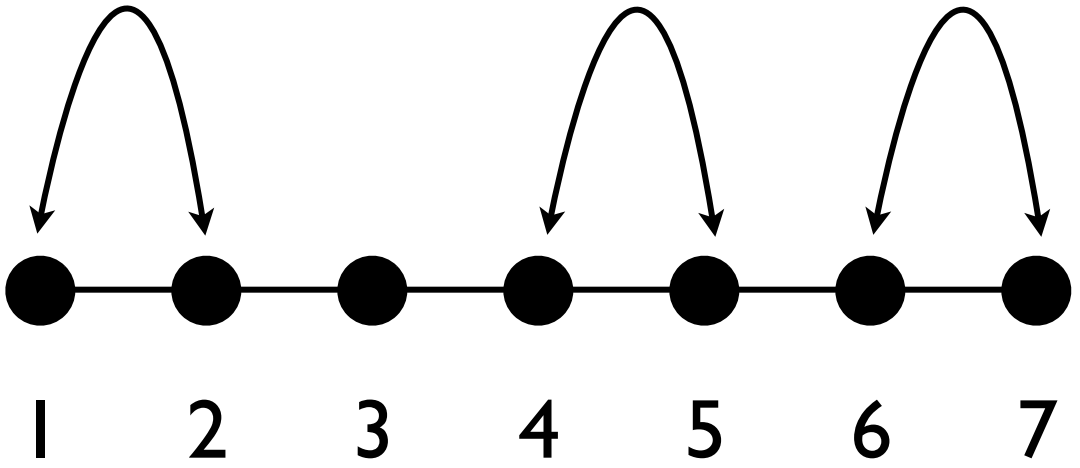
FAST OR SLOW?

- If you believe that performance factors can obscure competence factors, then time-limited judgments are undesirable

PRESENTATION OF MATERIALS

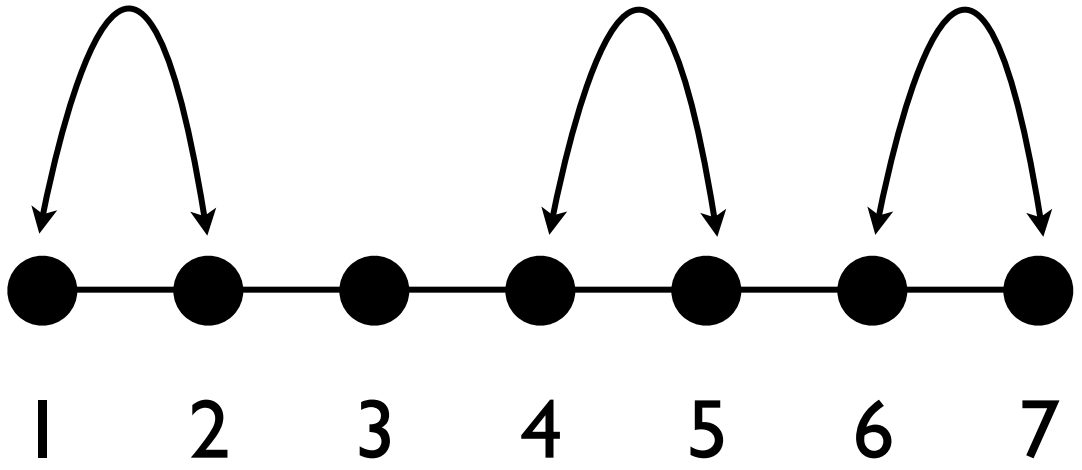
- Acceptability surveys often lack any assessment of understanding
  - Where possible, comprehension Qs provide a minimal check on reading for understanding

ORDINAL SCALE DATA



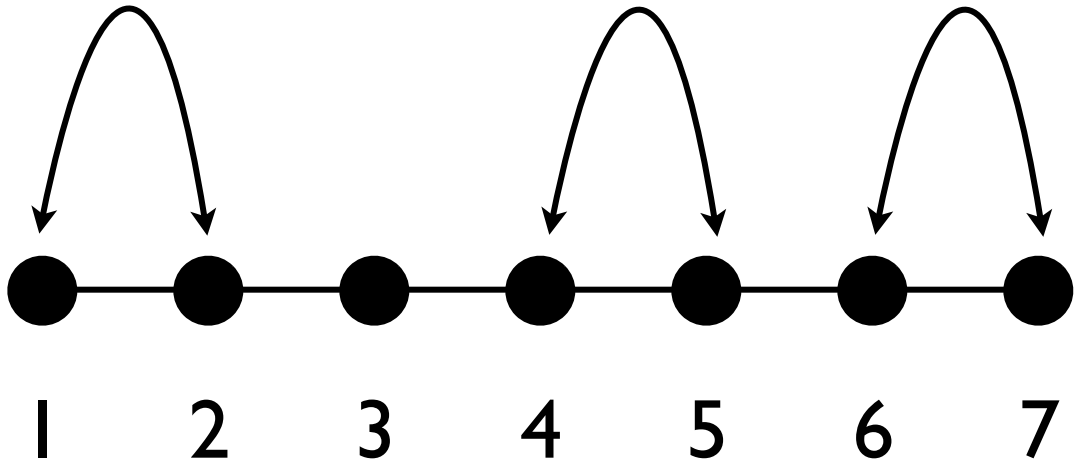
Is the distance between 1 & 2 the same as the distance between 4 & 5?

ORDINAL SCALE DATA



Most statistical methods for equation modeling assume that data are continuous

ORDINAL SCALE DATA



Stepwise logistic regression, weighted least squares, PROBIT regression

- Methodologies like ME & TJ produce linear data that is more suitable to standard statistical tests (like ANOVAs)
- YES/NO tasks can also be analyzed with well-understood methods (e.g. logistic regression)

CONCLUSION

- Much remains to be known about the judgment process and thus what the best way of eliciting and analyzing judgments is
  - But there is a considerable body of evidence to consult now . . .

RECOMMENDED RESOURCES

- Hill (1961)
- Chapman (1974)
- Greenbaum (1977)
- Chraudron (1983)
- Nagata (1988)
- Schuetze (1996)
- Cowart (1997)
- Dabrowska (2010)

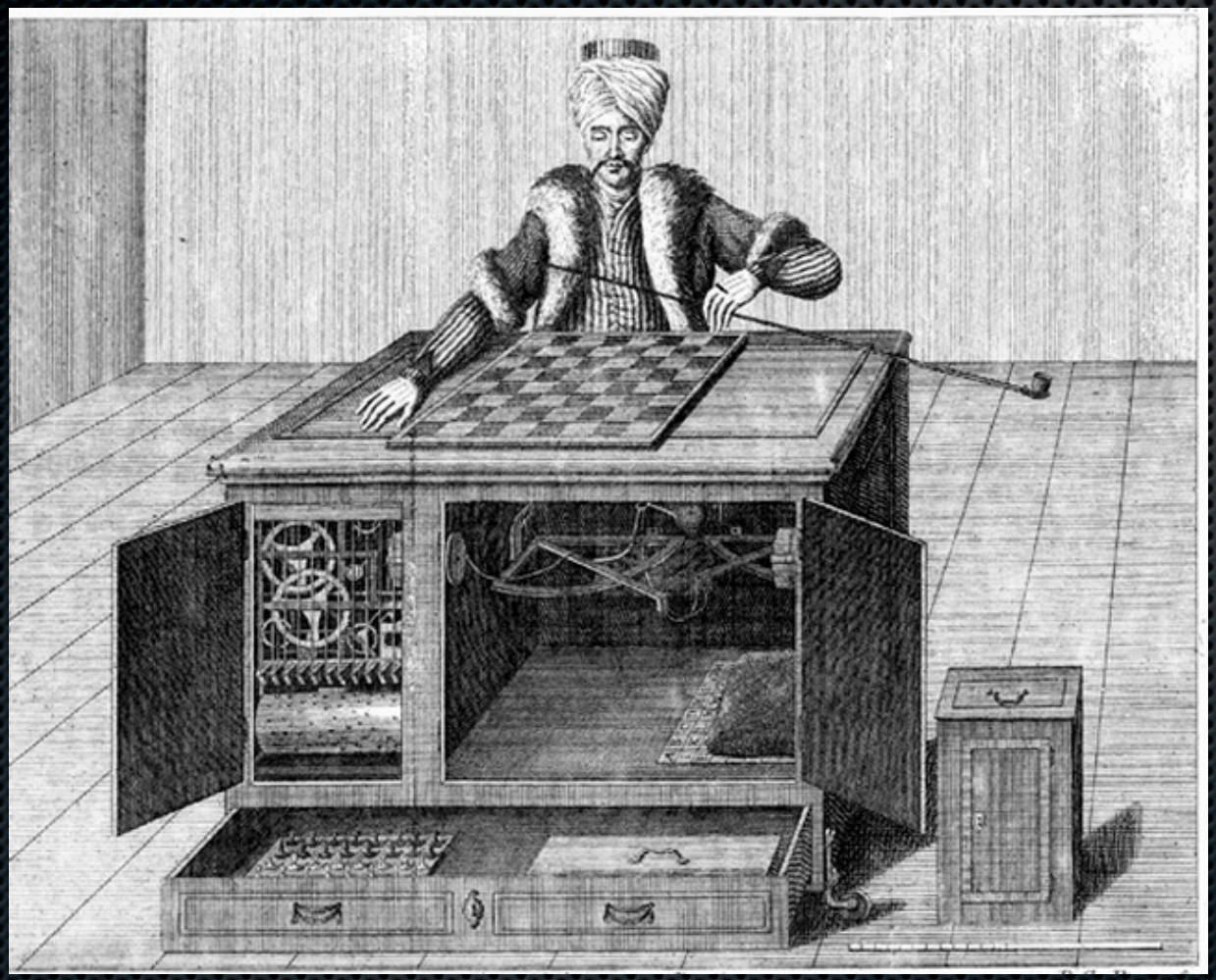
COLLECTING JUDGMENTS VIA MECHANICAL TURK

“

Rather, the main argument for the methodological status quo has always been that the benefits of formal experimentation are not (yet) offset by the decrease in convenience. [Myers 2009:418]

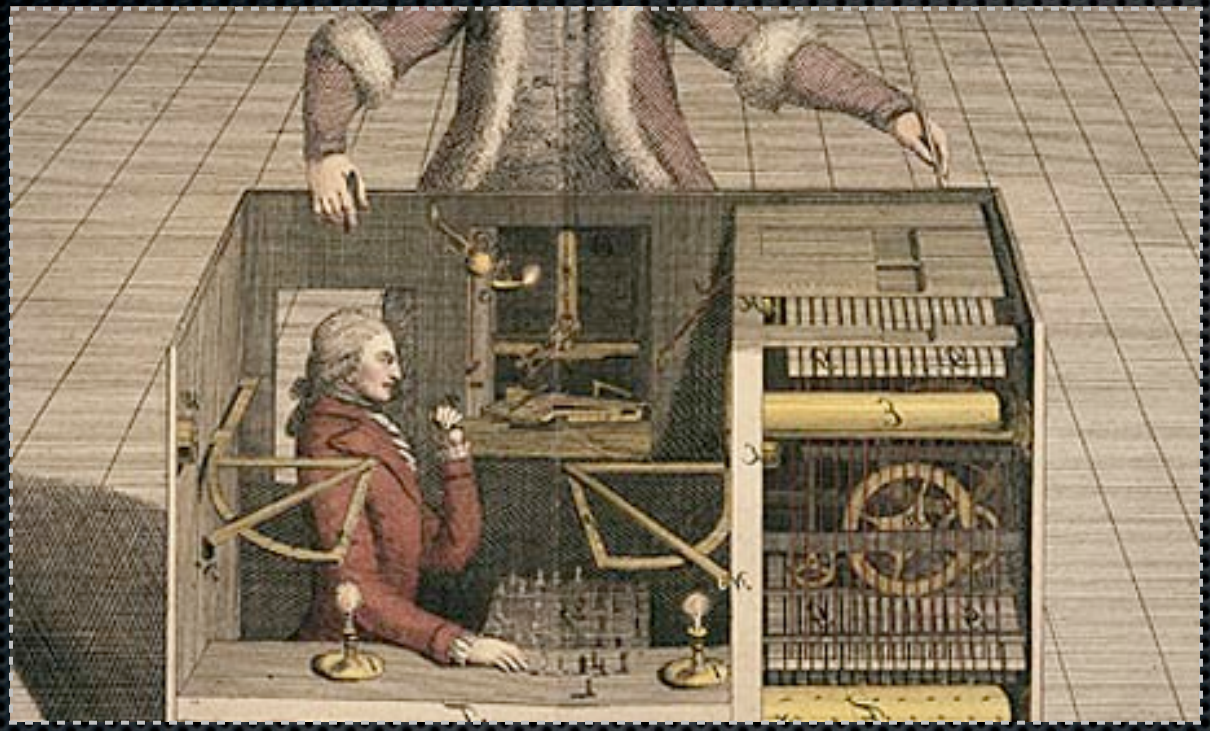
”

MECHANICAL TURK



- Amazon's Mechanical Turk is a crowd-sourcing forum that allows for quick, reliable, and easy collection of judgment data

MECHANICAL TURK



- Primarily developed as a means for eliciting human feedback where computer-baed methods are inefficient/ unreliable.

MECHANICAL TURK

- Some terminology
  - Requestor: individual/organization requesting workers to complete a task (HIT)
  - Worker: anonymous individual who completes task
  - HIT: Human Intelligence Task = Job

MECHANICAL TURK

- How does it work?

MECHANICAL TURK

- How does it work?
  - A requestor posts a HIT

MECHANICAL TURK

- How does it work?
  - A requestor posts a HIT
    - Specifies geographical restrictions, worker qualifications, e.g. 95% approval, pay rate

MECHANICAL TURK

- How does it work?
  - A requestor posts a HIT
    - Specifies geographical restrictions, worker qualifications, e.g. 95% approval, pay rate
  - Workers accept job, submit work

MECHANICAL TURK

- How does it work?
  - A requestor posts a HIT
    - Specifies geographical restrictions, worker qualifications, e.g. 95% approval, pay rate
  - Workers accept job, submit work
  - MT stores data and provides spreadsheet output

MECHANICAL TURK

- How does it work?
  - A requestor posts a HIT
    - Specifies geographical restrictions, worker qualifications, e.g. 95% approval, pay rate
  - Workers accept job, submit work
  - MT stores data and provides spreadsheet output
  - Requestor approves work quality & pays worker

MECHANICAL TURK

- Advantages

MECHANICAL TURK

- Advantages
  - Speed = 60-80 English-speaking subjects/hour

MECHANICAL TURK

- Advantages
  - Speed = 60-80 English-speaking subjects/hour
  - Cost-effective = To complete a 100 item survey, prices of $1-$3/hr are normal

MECHANICAL TURK

- Advantages
  - Speed = 60-80 English-speaking subjects/hour
  - Cost-effective = To complete a 100 item survey, prices of $1-$3/hr are normal
  - Reliable = At least 3 independent research teams have confirmed that it produces results similar to laboratory investigations (Frank et al 2010; Munro et al 2010; Sprouse 2011)

MECHANICAL TURK

- Disadvantages:
  - Still new, in beta mode
  - Somewhat restricted pool of participants; repeat participants are a danger
  - Cheating is possible if you're not careful
  - Mechanical Turk output is disorganized (but there's a solution)

MECHANICAL TURK

- What do you need to get started?
  - An item file
  - turkolizer - available for free @ tedlab.mit.edu/software/turkolizer.py
  - turk-template-changer.py (optional)
  - a computer with a command prompt, e.g. Terminal (Mac), CygWin (Windows), etc.

ITEM FILES

- Item files contain your experimental materials and any fillers/distractors
  - It's simple: write your items in a word processor with the following format
    - # experiment-name item_number condition_name
    - Target Sentence
    - Comprehension Question (Optional)

SAMPLE ITEM

exex 9 extract_noextra
Kenneth finally revealed which President he overheard a nasty remark about earlier while on the subway.
? Did Kenneth hear the remark on the subway? Yes

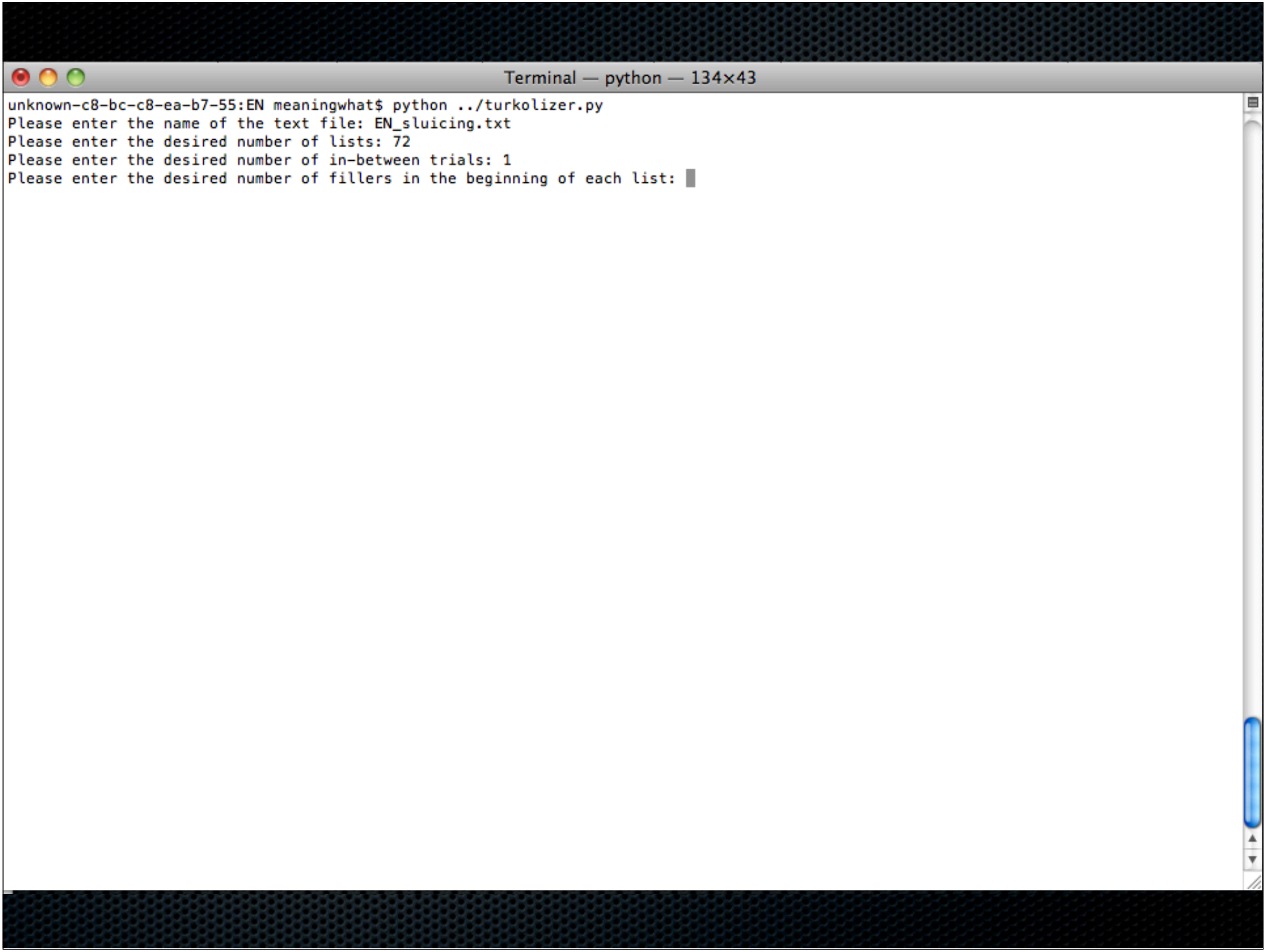
exex 9 extract_extra
Kenneth finally revealed which President he overheard a nasty remark earlier about while on the subway.
? Did Kenneth hear the remark on the subway? Yes

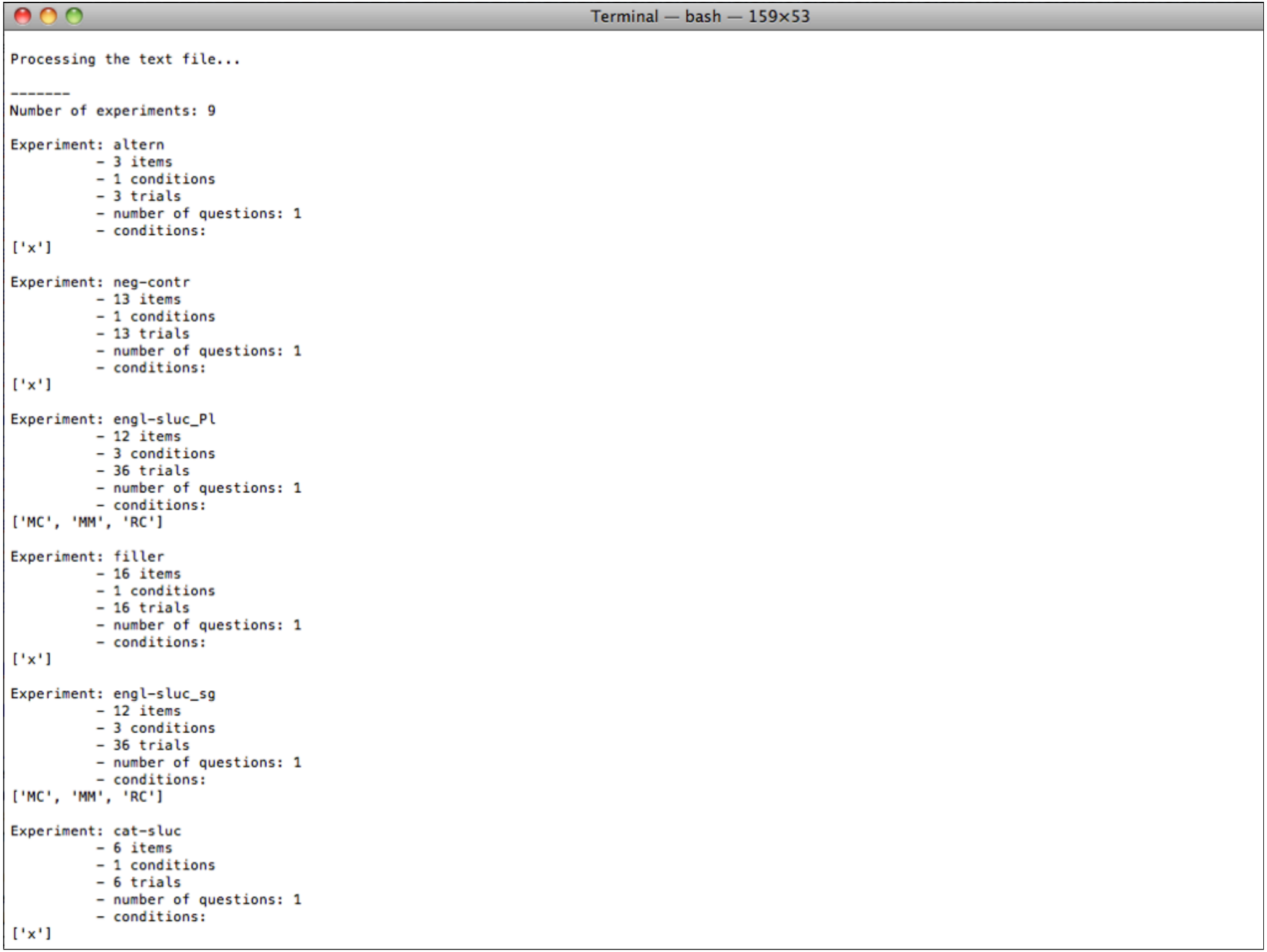
exex 9 noextract_extra
Kenneth finally revealed that he overheard a nasty remark about the President earlier while on the subway.
? Did Kenneth hear the remark on the subway? Yes

exex 9 noextract_noextra
Kenneth finally revealed that he overheard a nasty remark earlier about the President while on the subway.
? Did Kenneth hear the remark on the subway? Yes

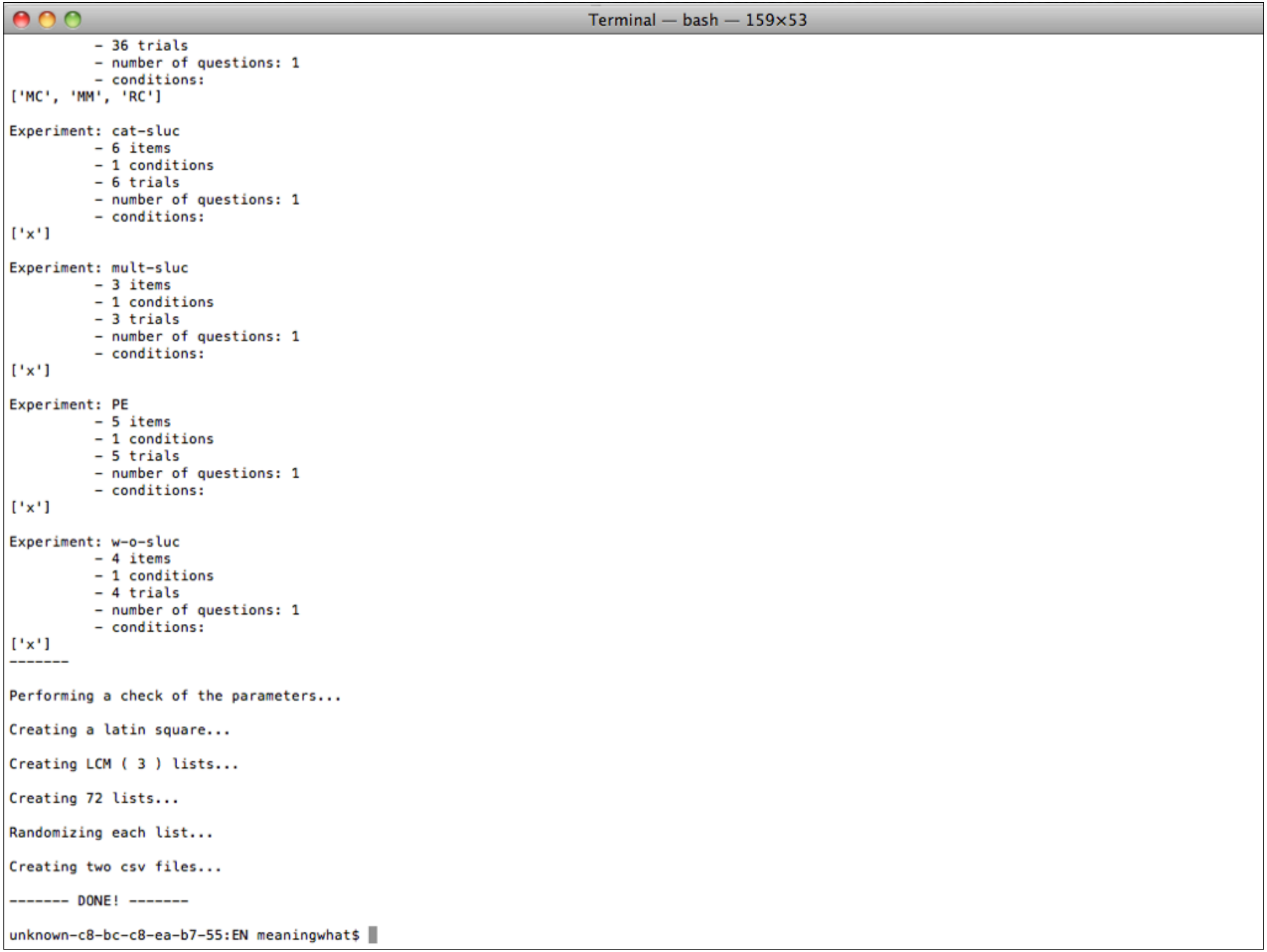
TURKOLIZER

- Software developed at TedLab @ MIT
- Takes your item file and turns into the format that Mechanical Turk needs
- Let's see an example

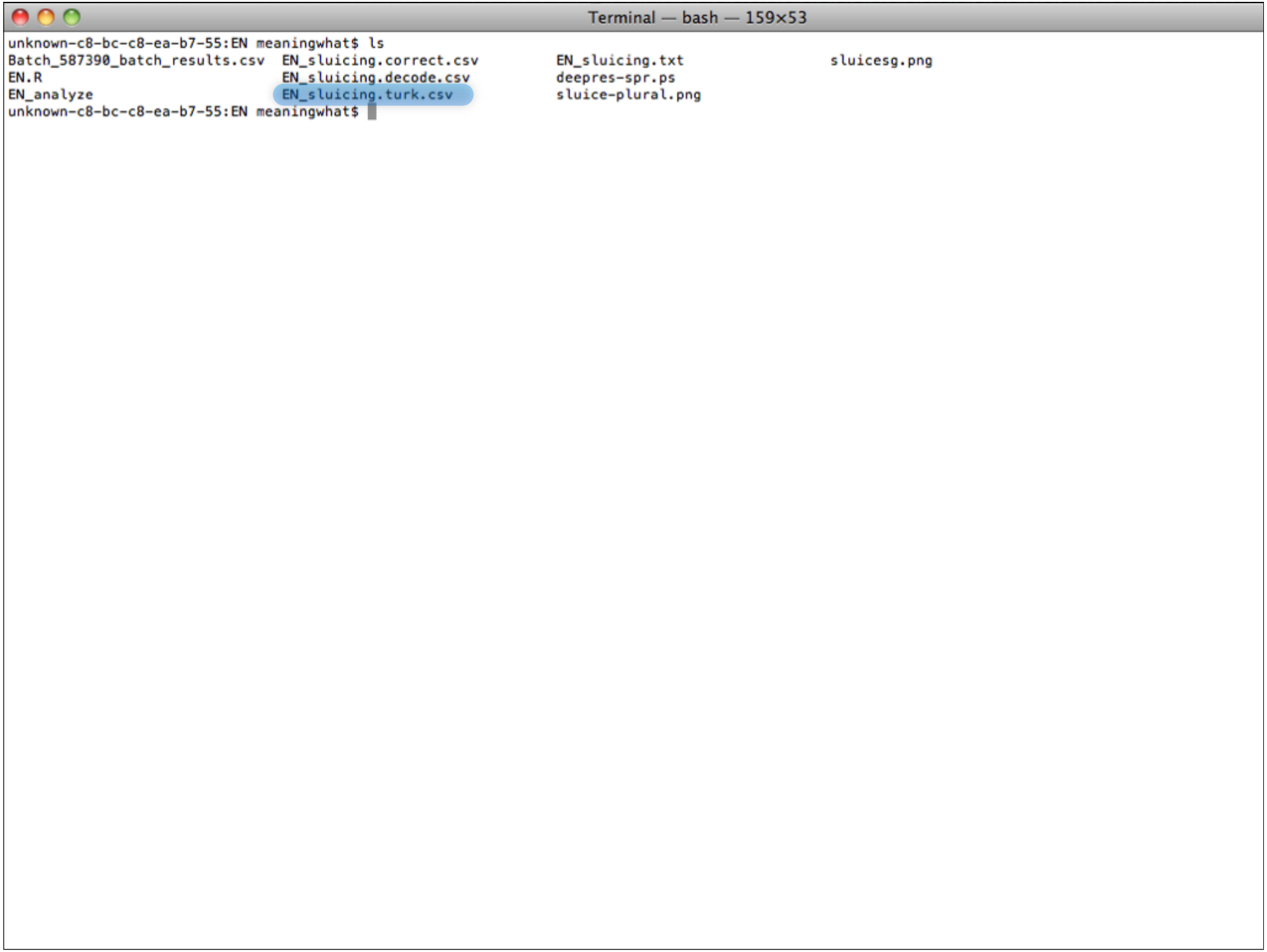

Terminal — python — 134×43
unknown-c8-bc-c8-ea-b7-55:EN meaningwhat$ python ../turkolizer.py
Please enter the name of the text file: EN_sluicing.txt
Please enter the desired number of lists: 72
Please enter the desired number of in-between trials: 1
Please enter the desired number of fillers in the beginning of each list:


Terminal — bash — 159×53
Processing the text file...

Number of experiments: 9
Experiment: altern
- 3 items
- 1 conditions
- 3 trials
- number of questions: 1
- conditions:
['x']
Experiment: neg-contr
- 13 items
- 1 conditions
- 13 trials
- number of questions: 1
- conditions:
['x']
Experiment: engl-sluc_Pl
- 12 items
- 3 conditions
- 36 trials
- number of questions: 1
- conditions:
['MC', 'MM', 'RC']
Experiment: filler
- 16 items
- 1 conditions
- 16 trials
- number of questions: 1
- conditions:
['x']
Experiment: engl-sluc_sg
- 12 items
- 3 conditions
- 36 trials
- number of questions: 1
- conditions:
['MC', 'MM', 'RC']
Experiment: cat-sluc
- 6 items
- 1 conditions
- 6 trials
- number of questions: 1
- conditions:
['x']

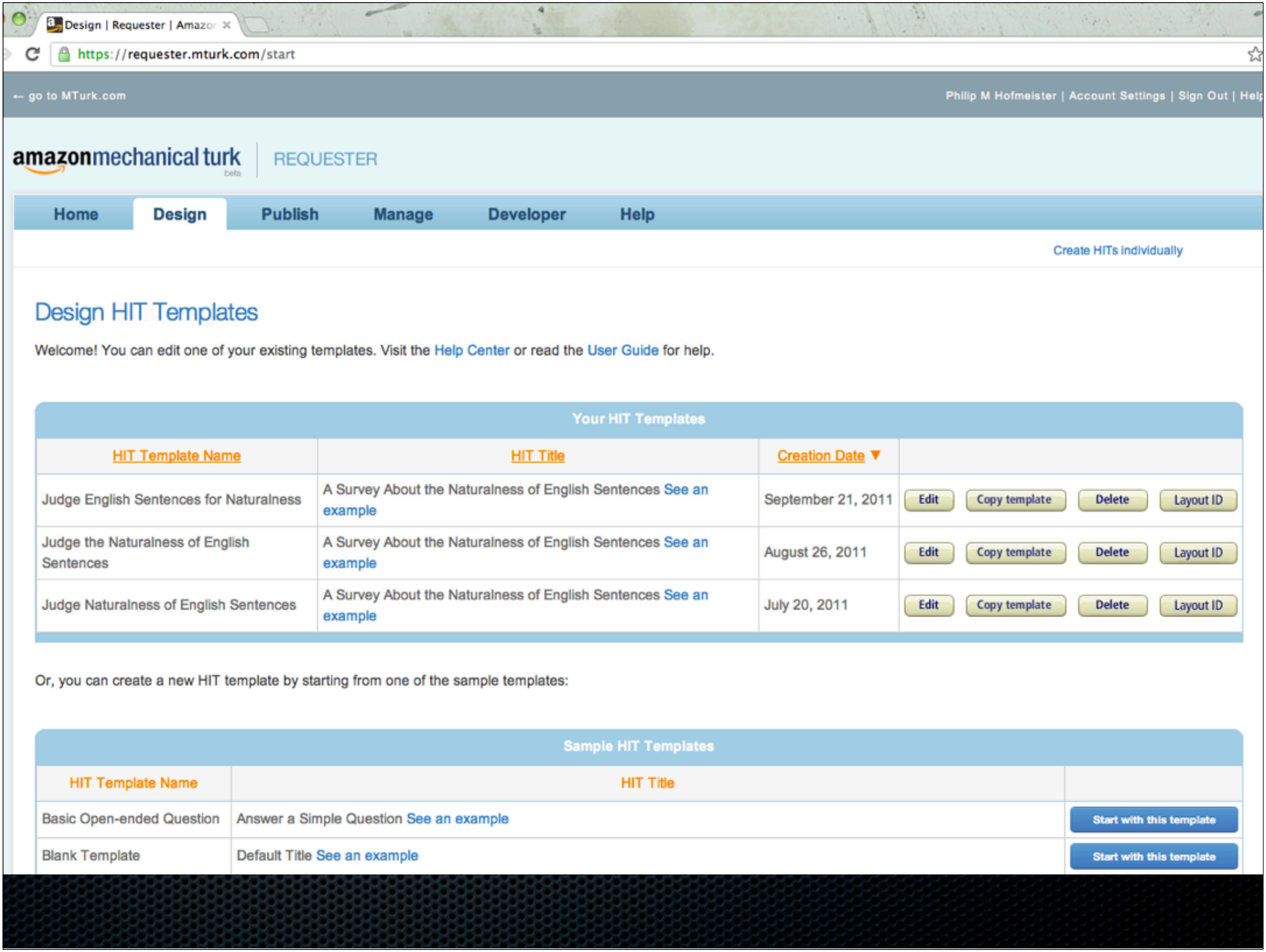

Terminal — bash — 159×53
- 36 trials
- number of questions: 1
- conditions:
['MC', 'MM', 'RC']
Experiment: cat-sluc
- 6 items
- 1 conditions
- 6 trials
- number of questions: 1
- conditions:
['x']
Experiment: mult-sluc
- 3 items
- 1 conditions
- 3 trials
- number of questions: 1
- conditions:
['x']
Experiment: PE
- 5 items
- 1 conditions
- 5 trials
- number of questions: 1
- conditions:
['x']
Experiment: w-o-sluc
- 4 items
- 1 conditions
- 4 trials
- number of questions: 1
- conditions:
['x']

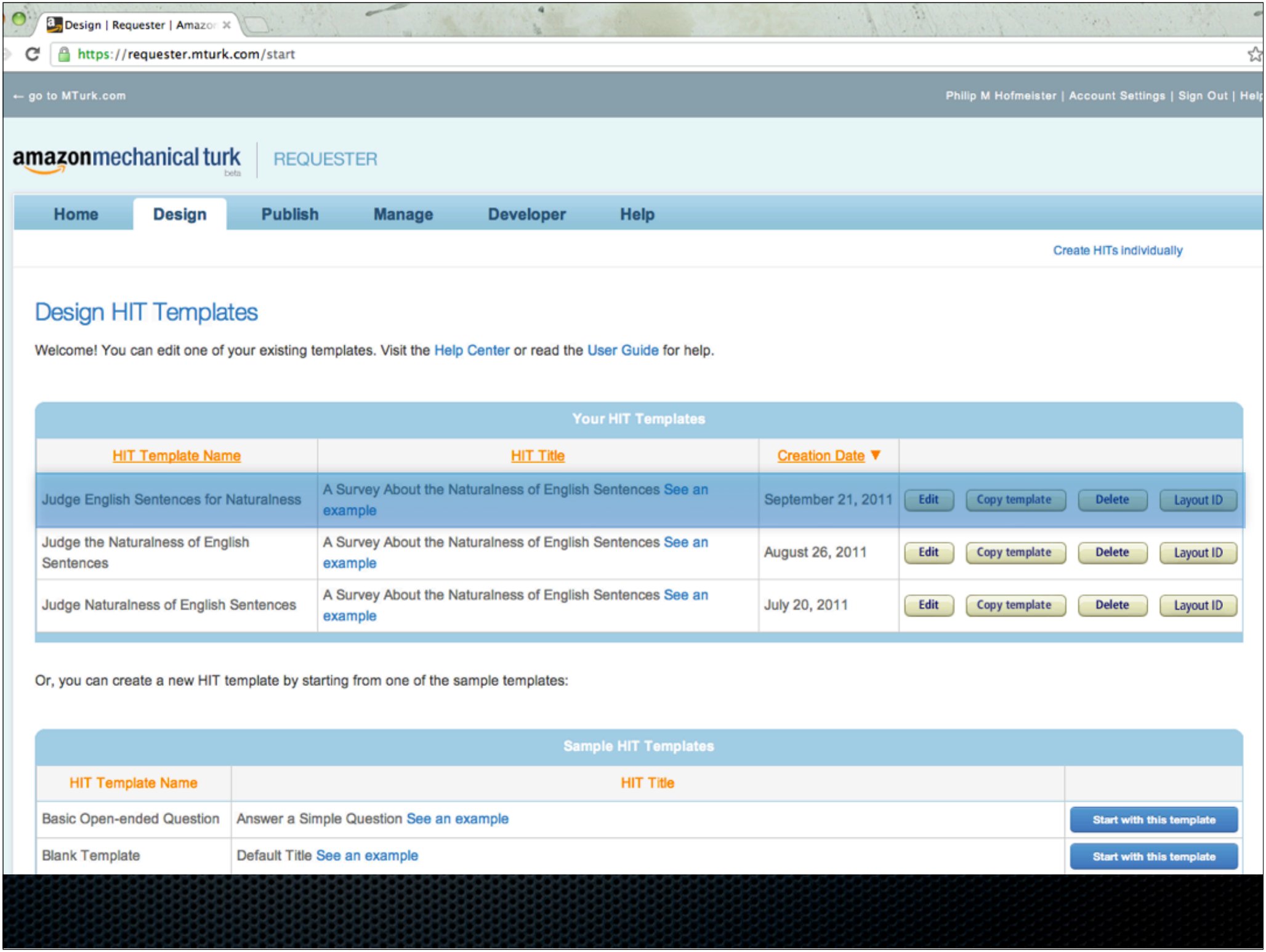
Performing a check of the parameters...
Creating a latin square...
Creating LCM (3) lists...
Creating 72 lists...
Randomizing each list...
Creating two csv files...
------- DONE! -------
unknown-c8-bc-c8-ea-b7-55:EN meaningwhat$

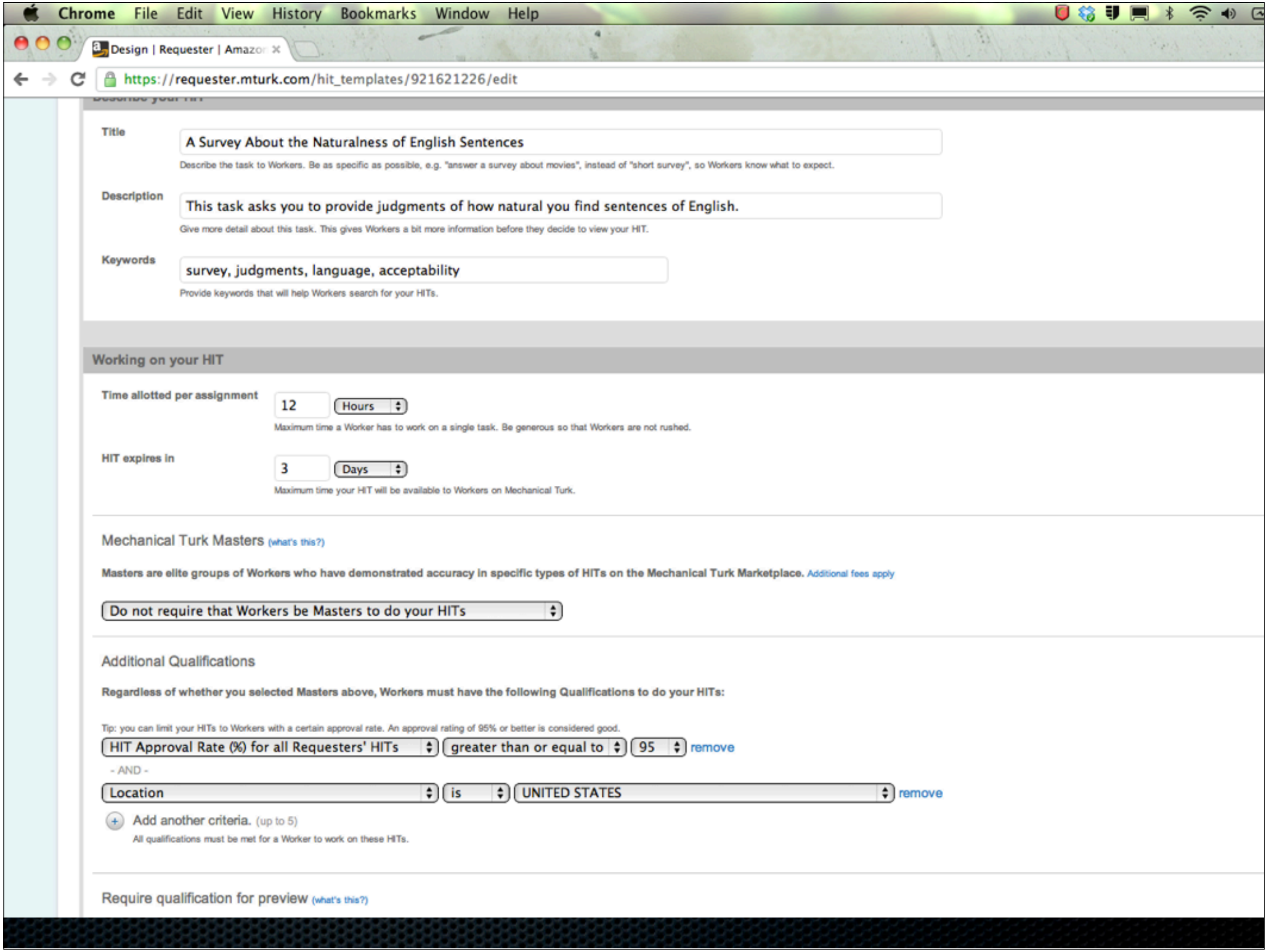

Terminal — bash — 159×53
unknown-c8-bc-c8-ea-b7-55:EN meaningwhat$ ls
Batch_587390_batch_results.csv
EN.R
EN_analyze
EN_sluicing.correct.csv
EN_sluicing.decode.csv
EN_sluicing.turk.csv
EN_sluicing.txt
deepres-spr.ps
sluice-plural.png
sluicesg.png
unknown-c8-bc-c8-ea-b7-55:EN meaningwhat$

MT STEPS

- 1) Select a template
- 2) Upload input data to template
- 3) Preview and publish
- 4) Download data


Design | Requester | Amazon
https://requester.mturk.com/start
← go to MTurk.com
Philip M Hofmeister | Account Settings | Sign Out | Help
amazonmechanical turk beta
REQUESTER
Home
Design
Publish
Manage
Developer
Help
Create HITs individually
Design HIT Templates
Welcome! You can edit one of your existing templates. Visit the Help Center or read the User Guide for help.
Your HIT Templates
HIT Template Name
HIT Title
Creation Date ▼
Judge English Sentences for Naturalness
A Survey About the Naturalness of English Sentences See an example
September 21, 2011
Edit
Copy template
Delete
Layout ID
Judge the Naturalness of English Sentences
A Survey About the Naturalness of English Sentences See an example
August 26, 2011
Edit
Copy template
Delete
Layout ID
Judge Naturalness of English Sentences
A Survey About the Naturalness of English Sentences See an example
July 20, 2011
Edit
Copy template
Delete
Layout ID
Or, you can create a new HIT template by starting from one of the sample templates:
Sample HIT Templates
HIT Template Name
HIT Title
Basic Open-ended Question
Answer a Simple Question See an example
Start with this template
Blank Template
Default Title See an example
Start with this template


Design | Requester | Amazon
https://requester.mturk.com/start
← go to MTurk.com
Philip M Hofmeister | Account Settings | Sign Out | Help
amazonmechanical turk beta
REQUESTER
Home
Design
Publish
Manage
Developer
Help
Create HITs individually
Design HIT Templates
Welcome! You can edit one of your existing templates. Visit the Help Center or read the User Guide for help.
Your HIT Templates
HIT Template Name
HIT Title
Creation Date ▼
Judge English Sentences for Naturalness
A Survey About the Naturalness of English Sentences See an example
September 21, 2011
Edit
Copy template
Delete
Layout ID
Judge the Naturalness of English Sentences
A Survey About the Naturalness of English Sentences See an example
August 26, 2011
Edit
Copy template
Delete
Layout ID
Judge Naturalness of English Sentences
A Survey About the Naturalness of English Sentences See an example
July 20, 2011
Edit
Copy template
Delete
Layout ID
Or, you can create a new HIT template by starting from one of the sample templates:
Sample HIT Templates
HIT Template Name
HIT Title
Basic Open-ended Question
Answer a Simple Question See an example
Start with this template
Blank Template
Default Title See an example
Start with this template


Chrome File Edit View History Bookmarks Window Help
Design | Requester | Amazo
https://requester.mturk.com/hit_templates/921621226/edit
Title
A Survey About the Naturalness of English Sentences
Describe the task to Workers. Be as specific as possible, e.g. "answer a survey about movies", instead of "short survey", so Workers know what to expect.
Description
This task asks you to provide judgments of how natural you find sentences of English.
Give more detail about this task. This gives Workers a bit more information before they decide to view your HIT.
Keywords
survey, judgments, language, acceptability
Provide keywords that will help Workers search for your HITs.
Working on your HIT
Time allotted per assignment
12
Hours
Maximum time a Worker has to work on a single task. Be generous so that Workers are not rushed.
HIT expires in
3
Days
Maximum time your HIT will be available to Workers on Mechanical Turk.
Mechanical Turk Masters (what's this?)
Masters are elite groups of Workers who have demonstrated accuracy in specific types of HITs on the Mechanical Turk Marketplace. Additional fees apply
Do not require that Workers be Masters to do your HITs
Additional Qualifications
Regardless of whether you selected Masters above, Workers must have the following Qualifications to do your HITs:
Tip: you can limit your HITs to Workers with a certain approval rate. An approval rating of 95% or better is considered good.
HIT Approval Rate (%) for all Requesters' HITs
greater than or equal to
95
remove
- AND -
Location
is
UNITED STATES
remove
Add another criteria. (up to 5)
All qualifications must be met for a Worker to work on these HITs.
Require qualification for preview (what's this?)

TEMPLATES

- An acceptability template is already available @ http://tedlab.mit.edu/software/tedlab-turk-survey1.html
- Simply copy and past the template in to a blank template, modifying the number of items as you need (template has 150 items)

Edit HIT Template

Use the HTML editor below to design the layout of your HIT. This layout is common for all of the HITs created with this template. You can define variables for data that will vary from HIT to HIT (Learn more).

1 Enter Properties | 2 Design Layout | 3 Preview and Finish

Template Name: Judge English Sentences for Nat This name is not displayed to Workers.

Frame Height 400 Height in pixels of the frame your HIT will be displayed in to Workers. Adjust the height appropriately to minimize scrolling for Workers.

Format | Font | Edit HTML Source

You are participating in a study being performed by scientists at the University of Tuebingen, Germany. This study has the name "EXEX_Sep21". Please do not complete this HIT if you have already completed a HIT with the same study name information.

What country are you from? ○USA ○Canada ○UK ○Australia / New Zealand ○India ○Other

Is English your first language? ○ Yes ○No

Please read each sentence, then answer the question immediately following, and rate the sentence for naturalness on a scale of 1 to 7, 1 being extremely unnatural and 7 being extremely natural. Assign higher numbers to sentences you find more natural, and lower numbers to sentences you find less natural.

Please note that there is a correct answer for each question.

Because some Mechanical Turk users answer questions randomly, we will reject users with error rates of 25% or larger. Consequently, if you cannot answer 75% of the questions correctly, please do not fill out the survey.

Note: **Please read the sentence** before answering the question and giving the rating!

We are interested in how natural you think the structure of the sentences below sound, not how plausible the meanings are. For example, "The man bit the dog" describes something less likely to happen than "The dog bit the man", but both are natural English sentences---there's nothing identifiably wrong with either sentence. So you should give them the same rating. You should provide ratings that match up with what would sound natural to you in a conversation or in reading a text, but you should NOT rely on what grammar books may have said is the right way to talk. There are no right or wrong ratings . . . we are exclusively interested in what your opinion is.

Sentence: ${trial_1}

Question: ${question_1_1}

○Yes ○No

Sentence rating:

○ 1(Extremely unnatural) ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7(Extremely natural)

Save | Preview and Finish

Sentence: ${trial_1}

Question: ${question_1_1}

◯ Yes ◯ No

Sentence rating:

◯ 1(Extremely unnatural) ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7(Extremely natural)

Sentence: ${trial_2}

Question: ${question_1_2}

◯ Yes ◯ No

Sentence rating:

◯ 1(Extremely unnatural) ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7(Extremely natural)

Sentence: ${trial_3}

Question: ${question_1_3}

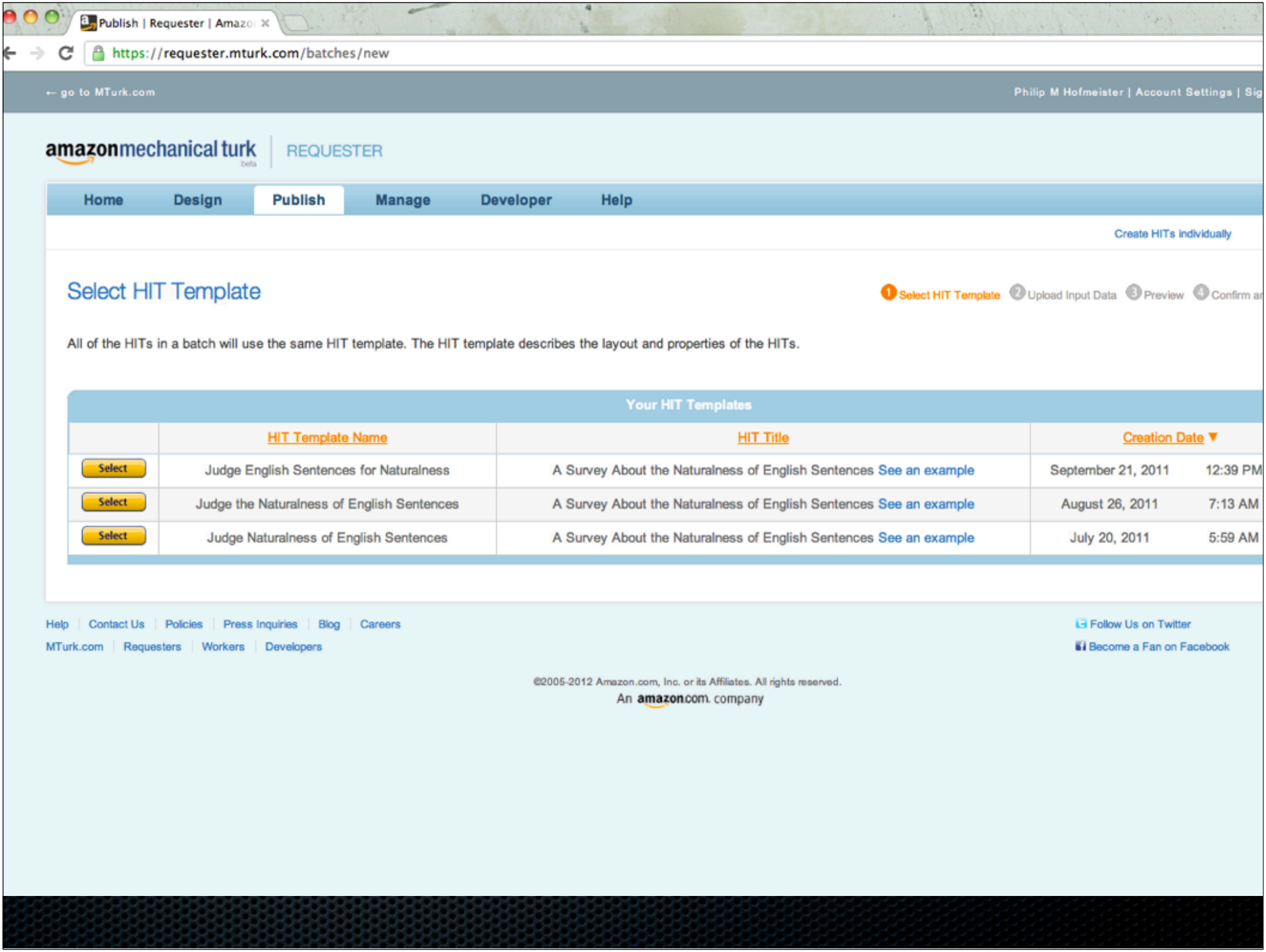
◯ Yes ◯ No

Sentence rating:

◯ 1(Extremely unnatural) ◯ 2 ◯ 3 ◯ 4 ◯ 5 ◯ 6 ◯ 7(Extremely natural)

PUBLISH

- You can make as many templates as you want for different experiment types
- Once you've got your template ready, it's time to merge the template with your items


Publish | Requester | Amazo
https://requester.mturk.com/batches/new
← go to MTurk.com
Philip M Hofmeister | Account Settings | Sig
amazonmechanical turk beta
REQUESTER
Home
Design
Publish
Manage
Developer
Help
Create HITs individually
Select HIT Template
1 Select HIT Template
2 Upload Input Data
3 Preview
4 Confirm a
All of the HITs in a batch will use the same HIT template. The HIT template describes the layout and properties of the HITs.
Your HIT Templates
HIT Template Name
HIT Title
Creation Date
Select
Judge English Sentences for Naturalness
A Survey About the Naturalness of English Sentences See an example
September 21, 2011 12:39 PM
Select
Judge the Naturalness of English Sentences
A Survey About the Naturalness of English Sentences See an example
August 26, 2011 7:13 AM
Select
Judge Naturalness of English Sentences
A Survey About the Naturalness of English Sentences See an example
July 20, 2011 5:59 AM
Help | Contact Us | Policies | Press Inquiries | Blog | Careers
MTurk.com | Requesters | Workers | Developers
Follow Us on Twitter
Become a Fan on Facebook
©2005-2012 Amazon.com, Inc. or its Affiliates. All rights reserved.
An amazon.com company

Upload Input Data

1 Select HIT Template 2 Upload Input Data 3 Preview 4 Confirm and Publish

Choose a .cvs file with the variables you specified in your HIT Template ("Judge English Sentences for Naturalness") (learn more).

Your file's column headers need to include: `trial_1, question_1_1, trial_2, question_1_2, trial_3, question_1_3, trial_4, question_1_4, trial_5, question_1_5, trial_6, question_1_6, trial_7, question_1_7, trial_8, question_1_8, trial_9, question_1_9, trial_10, question_1_10, trial_11, question_1_11, trial_12, question_1_12, trial_13, question_1_13, trial_14, question_1_14, trial_15, question_1_15, trial_16, question_1_16, trial_17, question_1_17, trial_18, question_1_18, trial_19, question_1_19, trial_20, question_1_20, trial_21, question_1_21, trial_22, question_1_22, trial_23, question_1_23, trial_24, question_1_24, trial_25, question_1_25, trial_26, question_1_26, trial_27, question_1_27, trial_28, question_1_28, trial_29, question_1_29, trial_30, question_1_30, trial_31, question_1_31, trial_32, question_1_32, trial_33, question_1_33, trial_34, question_1_34, trial_35, question_1_35, trial_36, question_1_36, trial_37, question_1_37, trial_38, question_1_38, trial_39, question_1_39, trial_40, question_1_40, trial_41, question_1_41, trial_42, question_1_42, trial_43, question_1_43, trial_44, question_1_44, trial_45, question_1_45, trial_46, question_1_46, trial_47, question_1_47, trial_48, question_1_48, trial_49, question_1_49, trial_50, question_1_50, trial_51, question_1_51, trial_52, question_1_52, trial_53, question_1_53, trial_54, question_1_54, trial_55, question_1_55, trial_56, question_1_56, trial_57, question_1_57, trial_58, question_1_58, trial_59, question_1_59, trial_60, question_1_60, trial_61, question_1_61, trial_62, question_1_62, trial_63, question_1_63, trial_64, question_1_64, trial_65, question_1_65, trial_66, question_1_66, trial_67, question_1_67, trial_68, question_1_68, trial_69, question_1_69, trial_70, question_1_70, trial_71, question_1_71, trial_72, question_1_72, trial_73, question_1_73, trial_74, question_1_74, trial_75, question_1_75, trial_76, question_1_76, trial_77, question_1_77, trial_78, question_1_78`

Judge English Sentences for Naturalness

Locate a New File

Choose File No file chosen Upload Download a sample of the input file for this HIT template

Or Choose an Existing File

Your Existing Files

	File Name	Input Lines	Creation Date ▼		
Select	extract-extrapose.turk.csv	72	September 21, 2011	12:48 PM PDT	Delete
Select	caps-context.turk.csv	64	August 26, 2011	7:14 AM PDT	Delete
Select	in-situ-items.turk.csv	72	August 17, 2011	4:49 AM PDT	Delete
Select	extractionPP-items.turk.csv	24	August 15, 2011	8:35 AM PDT	Delete
Select	ppdist_resump.turk.csv	60	July 20, 2011	6:29 AM PDT	Delete

Back

Sentence: We read which universities the evidence shows that has the best paid graduates.

Question: Did the evidence specify admission rates for the best universities?

○Yes ○No

Sentence rating:

○ 1(Extremely unnatural) ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7(Extremely natural)

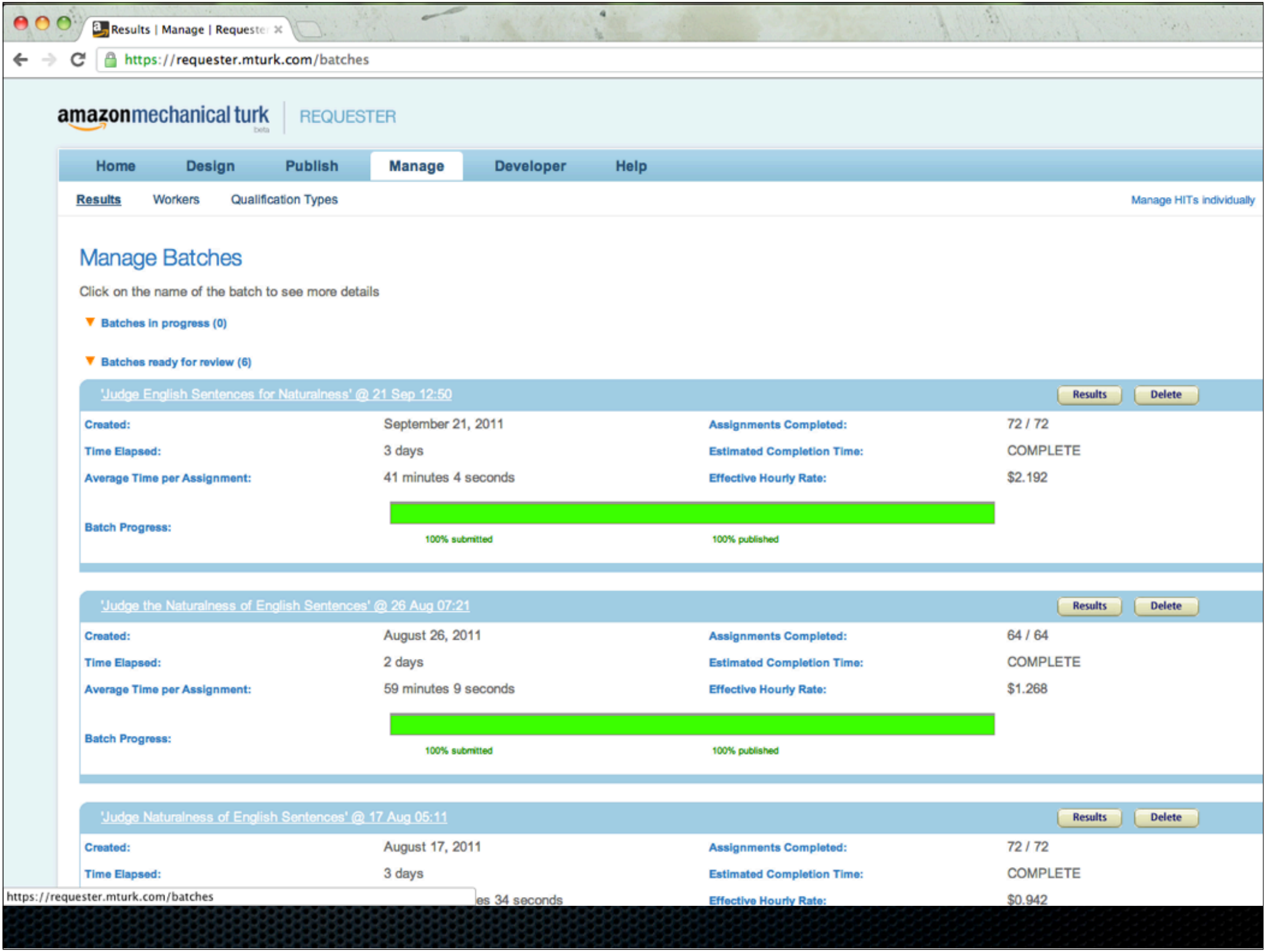
Sentence: Hundreds of people turned up at the house of a state minister demanding a solution to the problem.

Question: Were citizens hounding government officials at state buildings?

○Yes ○No

Sentence rating:

○ 1(Extremely unnatural) ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7(Extremely natural)


Results | Manage | Requester
https://requester.mturk.com/batches
amazonmechanical turk beta
REQUESTER
Home
Design
Publish
Manage
Developer
Help
Results
Workers
Qualification Types
Manage HITs individually
Manage Batches
Click on the name of the batch to see more details
Batches in progress (0)
Batches ready for review (6)
'Judge English Sentences for Naturalness' @ 21 Sep 12:50
Results
Delete
Created: September 21, 2011
Assignments Completed: 72 / 72
Time Elapsed: 3 days
Estimated Completion Time: COMPLETE
Average Time per Assignment: 41 minutes 4 seconds
Effective Hourly Rate: $2.192
Batch Progress:
100% submitted
100% published
'Judge the Naturalness of English Sentences' @ 26 Aug 07:21
Results
Delete
Created: August 26, 2011
Assignments Completed: 64 / 64
Time Elapsed: 2 days
Estimated Completion Time: COMPLETE
Average Time per Assignment: 59 minutes 9 seconds
Effective Hourly Rate: $1.268
Batch Progress:
100% submitted
100% published
'Judge Naturalness of English Sentences' @ 17 Aug 05:11
Results
Delete
Created: August 17, 2011
Assignments Completed: 72 / 72
Time Elapsed: 3 days
Estimated Completion Time: COMPLETE
es 34 seconds
Effective Hourly Rate: $0.942
https://requester.mturk.com/batches

MANAGING RESULTS

- MT delivers the results in a rather unwieldy format
  - Each row contains the values from each worker/HIT
  - For statistical analysis, you want each judgment on its row
- Results can be re-organized in SPSS/R/Matlab
- tedlab-turk-survey1-format.R does this for you

MANAGING WORKERS

- You must approve your workers' performance for them to get paid
  - Don't decline to pay unless the participant really ignored the instructions or cheated
  - MT workers track who pays and who doesn't (they even have a union!)

BEYOND JUDGMENTS

- Of course, MT can be used for more than just eliciting acceptability judgments
- It can be used for
  - Sociolinguistic surveys
  - Forced-choice tasks
  - Elicitation/completion tasks
  - Self-paced reading

CONCLUSION

- Gathering linguistic data is now easier than ever given technological and crowd-sourcing advances
- Costs are now considerably less, and you need increasingly fewer technical skills to implement experiments
- Experimental data allows for quantitative and statistical analyses that elude traditional methods

. . . UNFORTUNATELY

- Because MT is in beta mode, you or a colleague unfortunately need a US credit card at the moment to be a requestor

RESOURCES

- http://tedlab.mit.edu/tedlab_website/researchpapers/Gibson_et_al_2011_LLCO_mturk.pdf

For Christ's sake, stop!