Balancing experimental lists without sacrificing voluntary participation

Andrew Watts & T. Florian Jaeger

HLP Lab
Department of Brain and Cognitive Sciences
University of Rochester

July 27, 2011

UNIVERSITY of ROCHESTER

Andrew Watts

UNIVERSITY of ROCHESTER

UNIVERSITY of ROCHESTER

Why crowdsourcing?

- Provides ready access to a large heterogeneous pool of participants
  - Participants we normally wouldn't have access to in a local lab
  - Participants with multiple language backgrounds
- Allows for rapid data collection

What it requires - Mirroring procedures in the lab

- Want to be able to use balanced factorial designs (Latin square design)
- Ethics: voluntary participation
  - Need to allow participants to voluntarily abort experiments
- This can cause problems
  - Creates Zipf-distributed data and unbalanced lists
  - Makes data complicated or impossible to analyze (loss of power, etc)

Desiderata

- A way to ensure that participants see all (or most) of the items in their assigned experimental condition
  - While still allowing voluntary withdrawal from experiment *at most times during the experiment*
- A way to ensure that all items within an experimental condition are seen by relatively equal numbers of participants

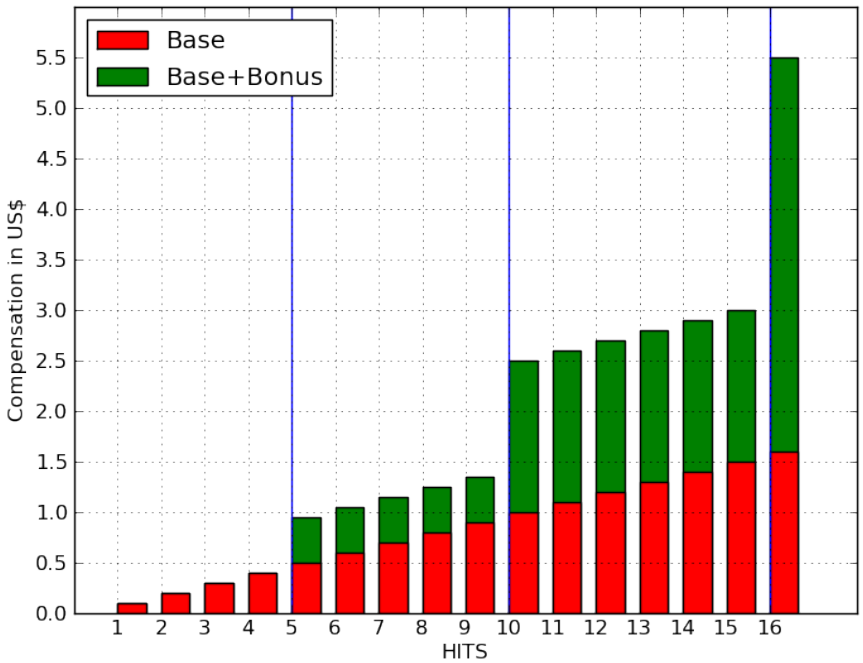
How do you get workers to do long experiments?

- Pay well and develop a reputation for doing so and for paying in a timely manner
- Take advantage of the Bonus system to reward good workers and weed out less interested and committed workers

Pay low, bonus high

Compensation bar graph shown to participants

- Example with $0.10 base and bonuses of:
  - $0.45 for 5 HITs
  - $1.50 for 10 HITs
  - $3.90 for all 16 HITs
- Vertical lines show levels where bonus increases

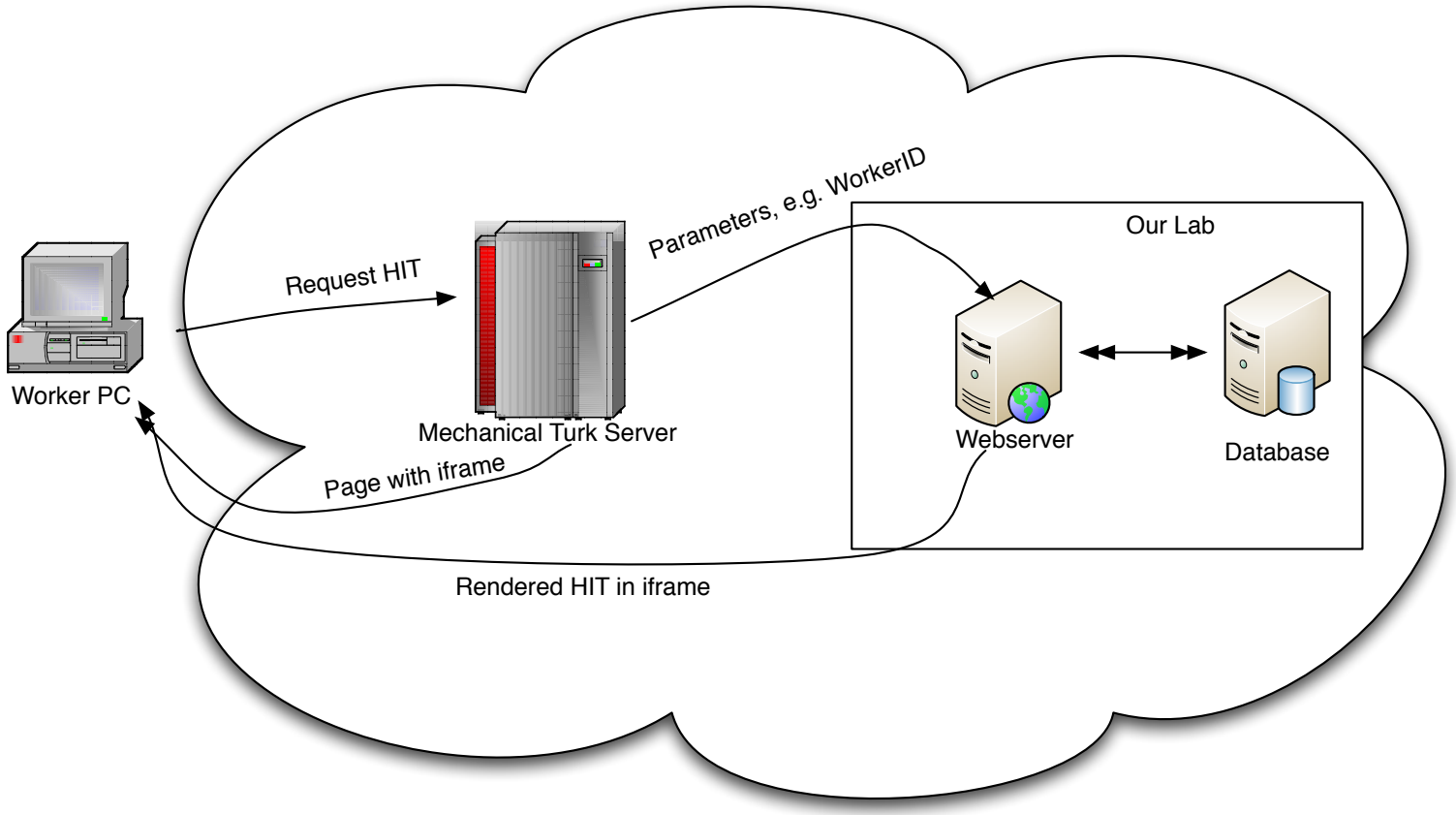


UNIVERSITY of ROCHESTER

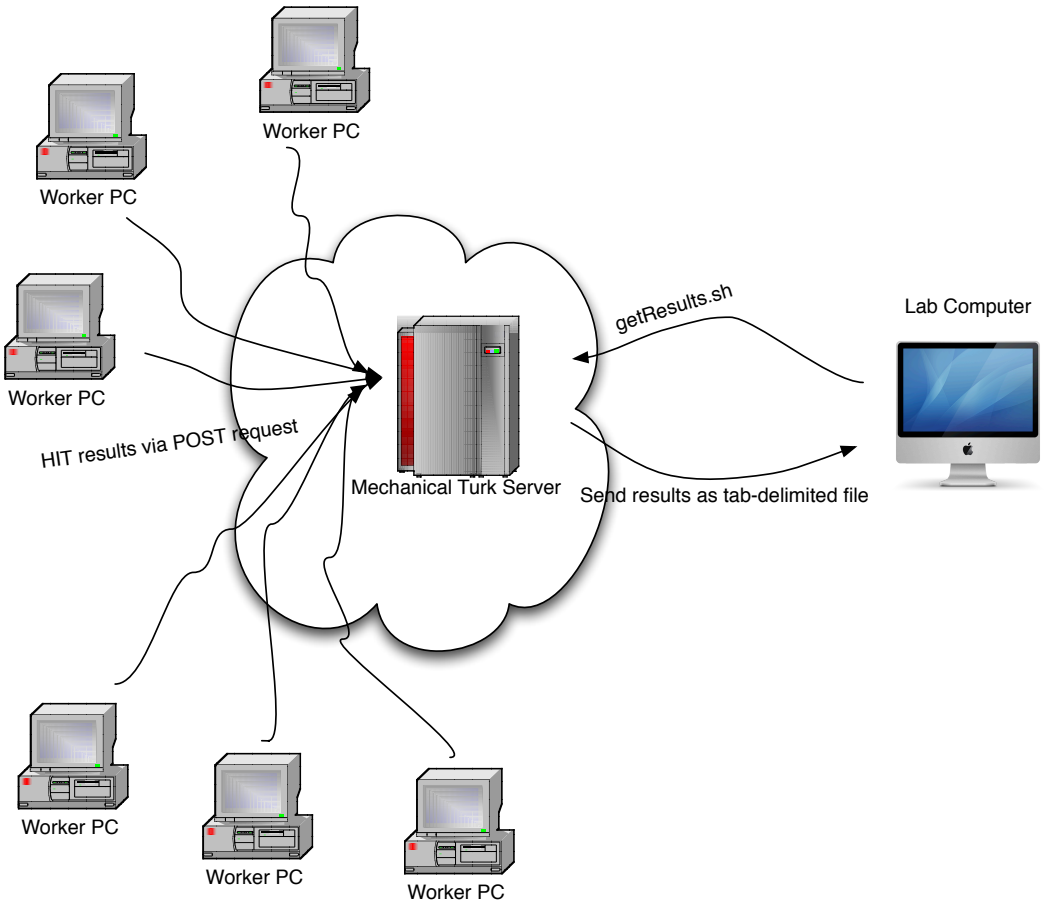
The problem of list balancing on MTurk

- We want to do factorial designs
- MTurk does not natively support the idea of multiple lists

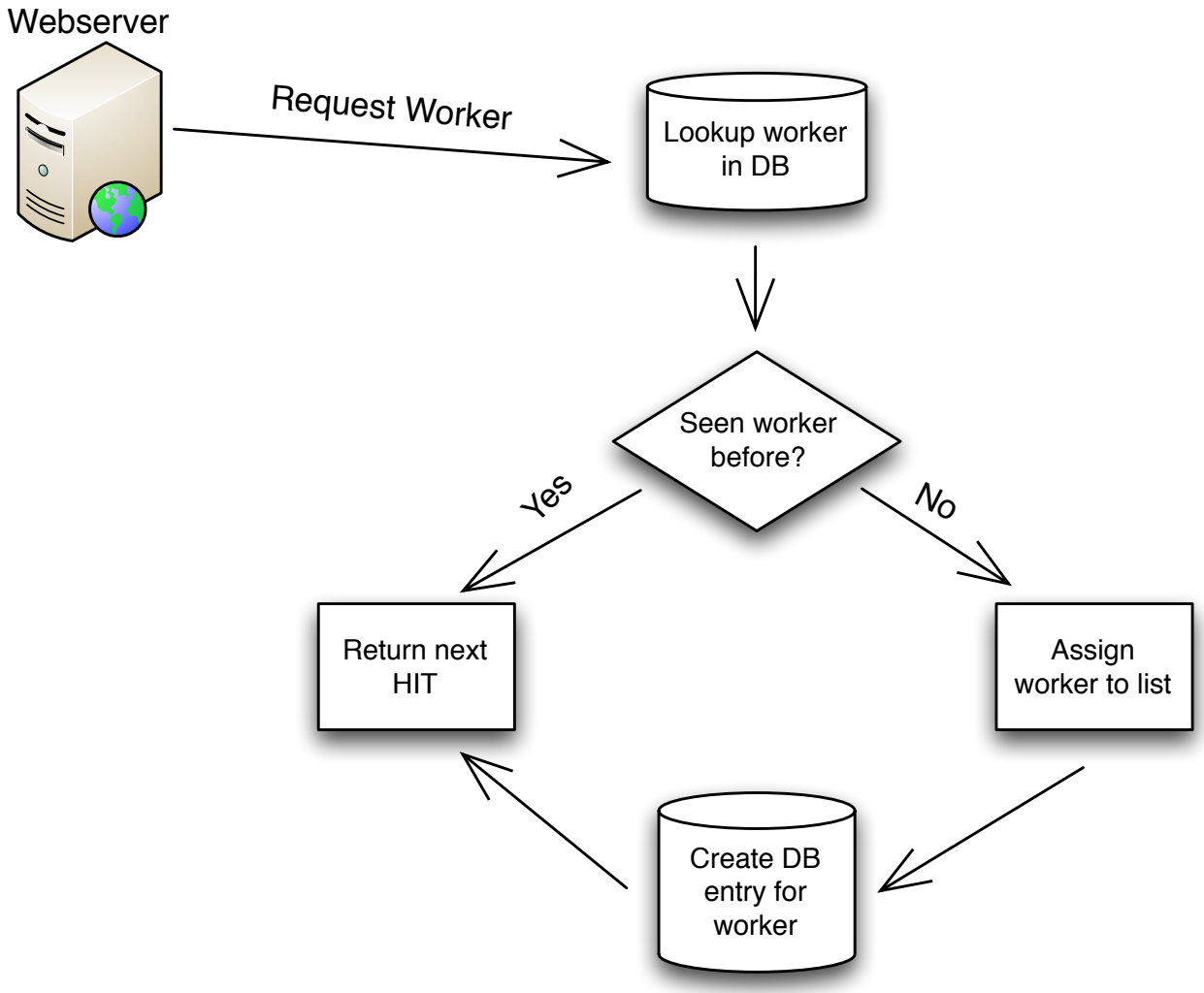
Our list balancing solution - mturker side



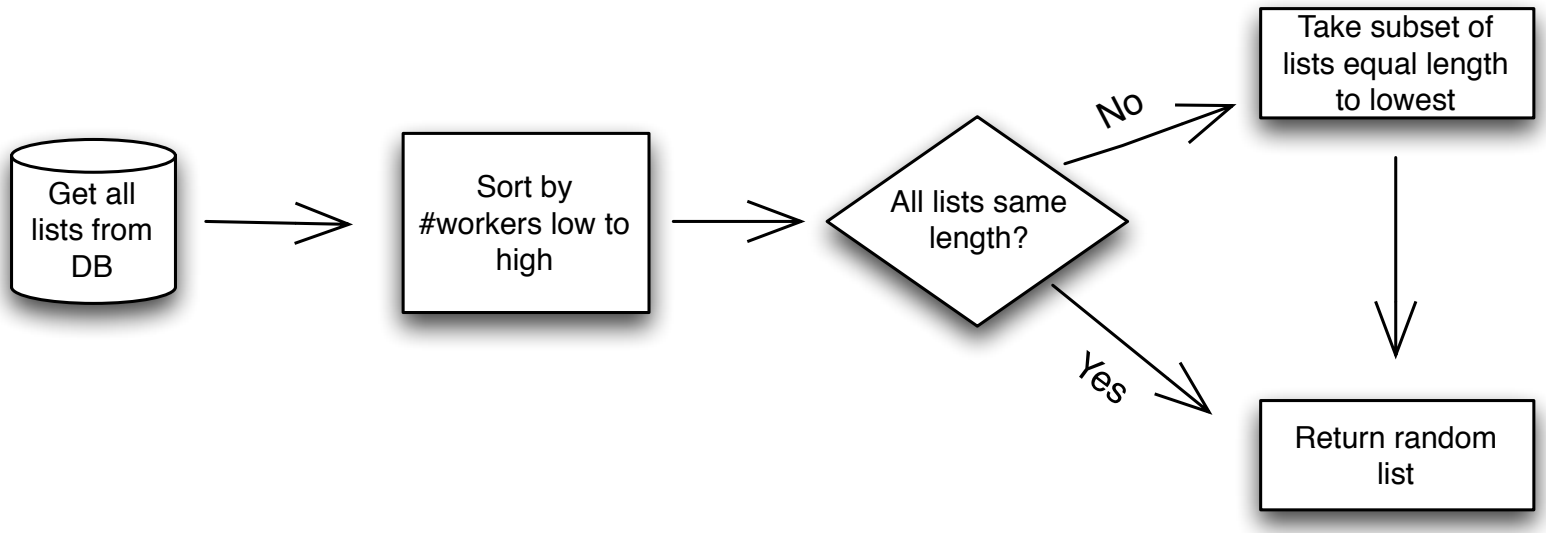
External Question workflow - experimenter side



On HIT request



List selection algorithm



Caveat: Balance remains imperfect

- Balanced list assignments in *our database*
- But MTurk provides no callbacks when HITs are submitted or returned (up to 25%)

→ Number of workers actually on each list can be unbalanced

  - For large number of participants, this evens out
  - Quick fix for all other cases: request half of desired HITs, examine balance, restrict lists to those that need to be filled, iterate sequence until lists are filled and close to balanced

Example studies

- Studies 1-3 (Jaeger, Levy, and Ferreira, 2010): using different progressive pay schemes
  - Syntactic Reduction of Object-extracted Relative Clauses
  - **Written Recall:** Encode sentence for 8 seconds, two simple math problems, recall cue shown for 2 seconds, type complete sentence.
- Studies 4-5 (Hansen-Karr, Ferris, and Jaeger, in prep): Using different ways to balance list
  - **Syntactic priming:** auditory comprehension to written production of ditransitives

UNIVERSITY of ROCHESTER

Studies 1-3: Data

- Study 1: 2304 critical items (31% data loss)
- Study 2: 452 critical items (34% data loss)
- Study 3: 2048 critical items (23% data loss)
- (comparable to lab-based experiments; e.g. Ferreira and Dell, 2000 had 30-60% data loss)

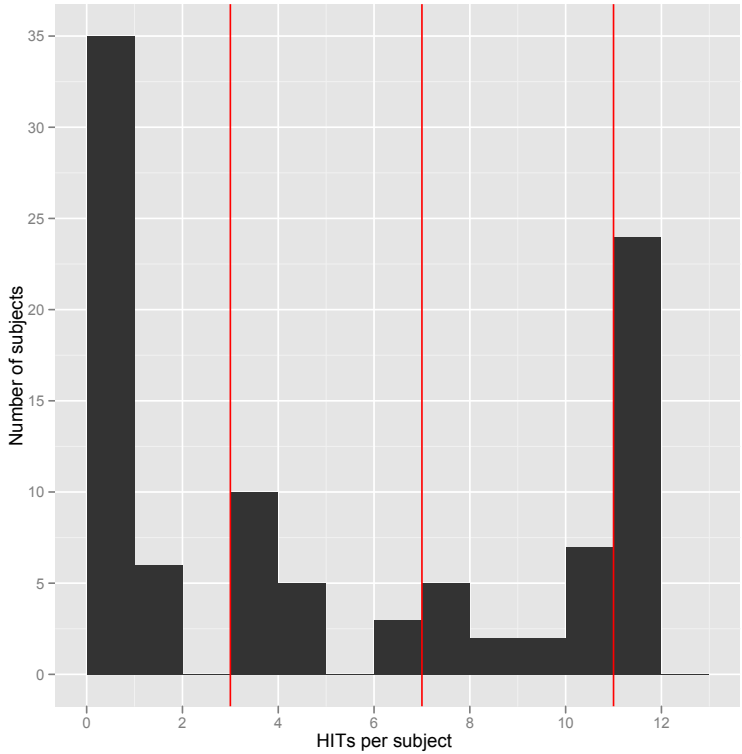
Study 1

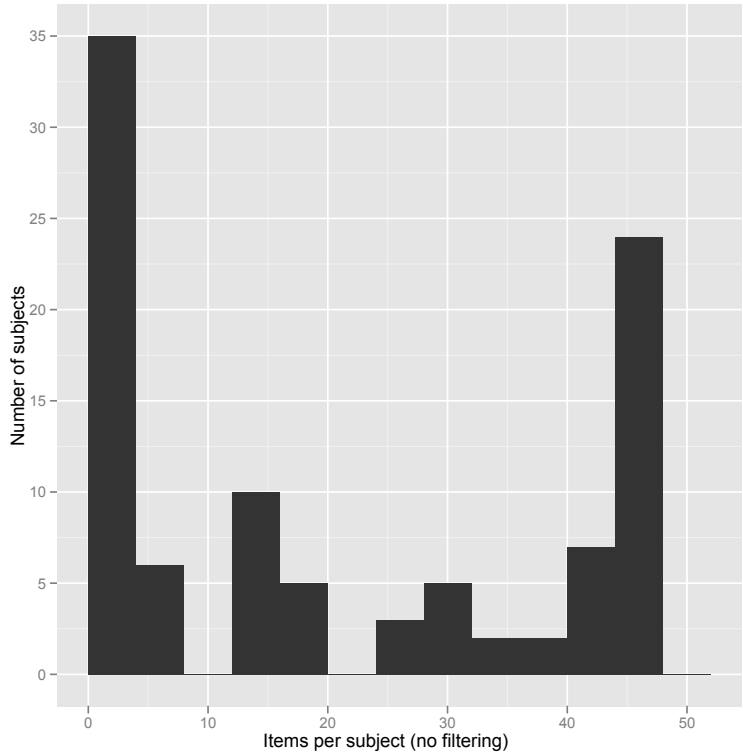
- Each HIT consisted of 8 trials (given the ordering constraints, these were likely to be 5-6 fillers and 2-3 targets)
- Each HIT paid $.10, plus
  - $0.40 for 4 HITs
  - $1.00 for 8 HITs
  - $2.00 for all 12 HITs

UNIVERSITY of ROCHESTER

Study 1: Result

max 12 hits/subj or 48 items/subj





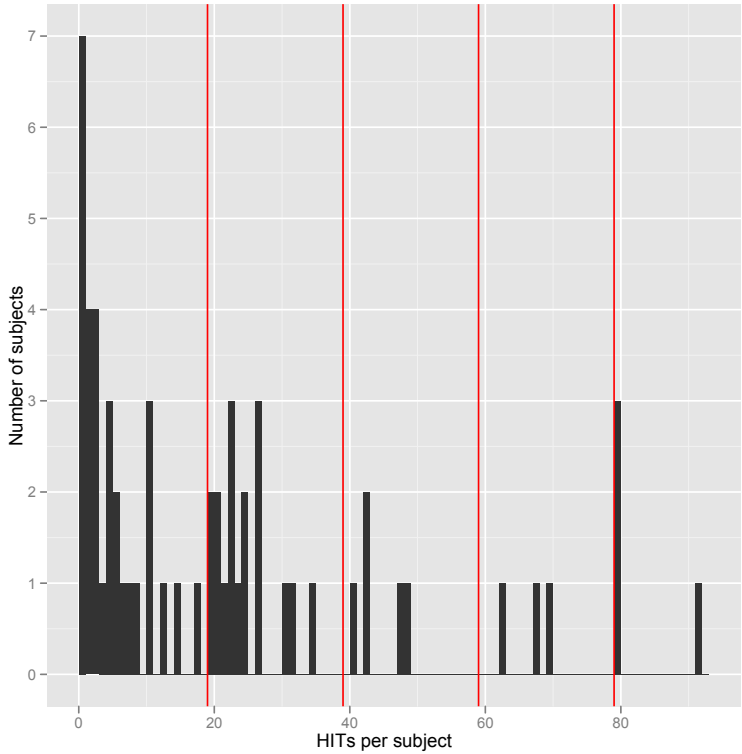
UNIVERSITY of ROCHESTER

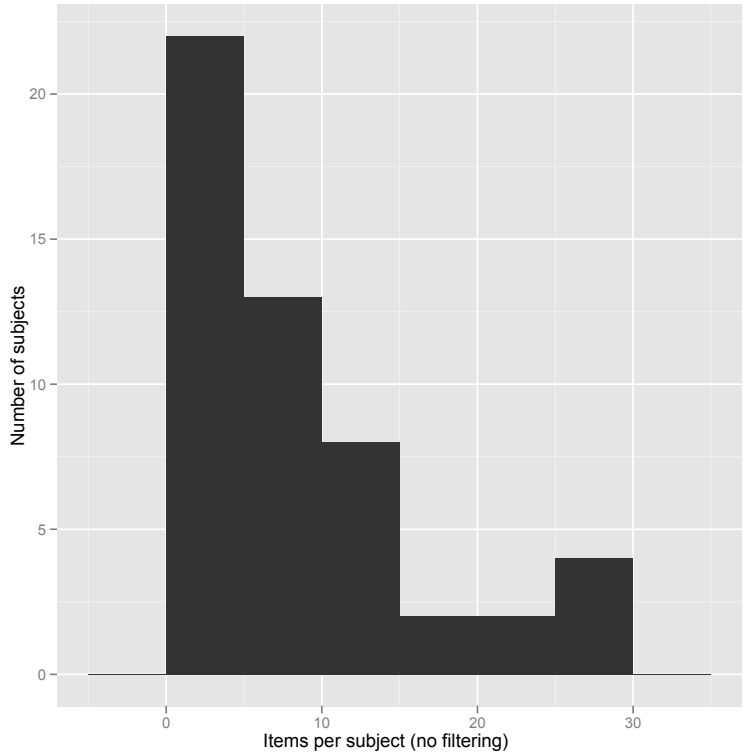
Study 2

- Each HIT consisted of one item **(subjects can stop after every trial)**
- Each HIT paid $.02, plus
  - $.20 for per every 20 HITs **(more regular increments)**

Study 2: Result

max 96 hits/subj or 32 items/subj





UNIVERSITY of ROCHESTER

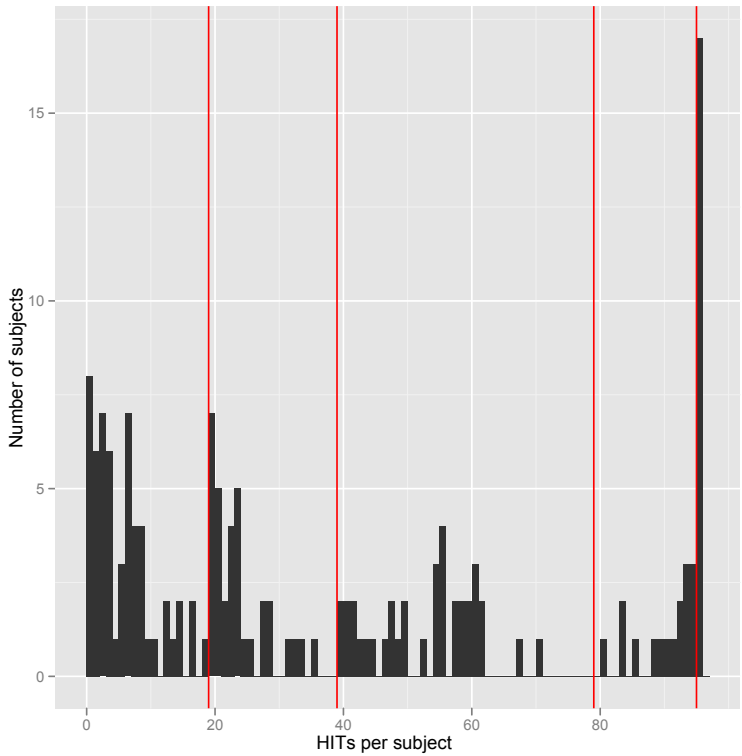
Study 3

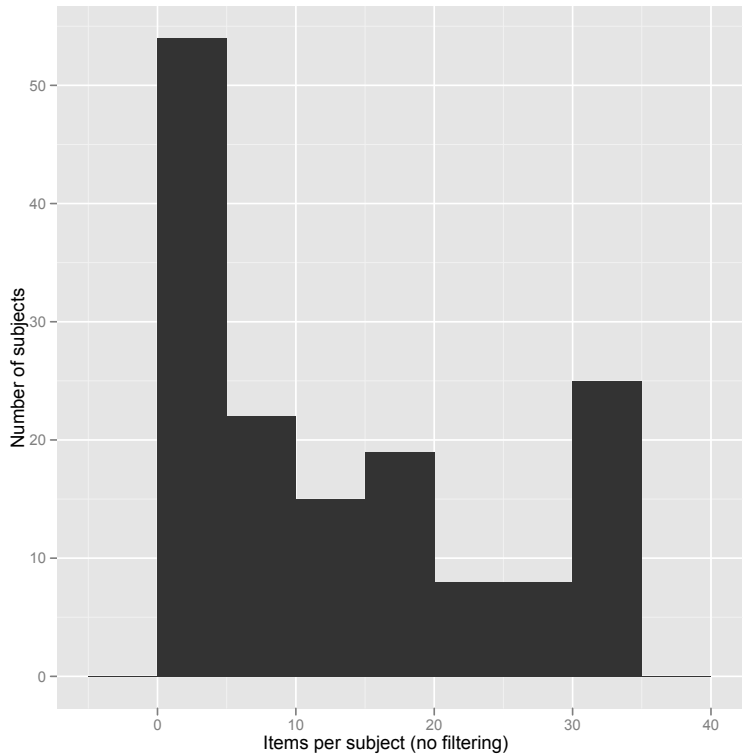
- Each HIT consisted of one trial
- Each HIT paid $.02, plus
  - $.20 for 20 HITs
  - $0.50 for 40 HITs
  - $1.25 for 80 HITs
  - $1.50 for all 96 HITs **(shifted last increment to end of list)**

UNIVERSITY of ROCHESTER

Study 3 - Results

max 96 HITs/subj or 32 items/subj



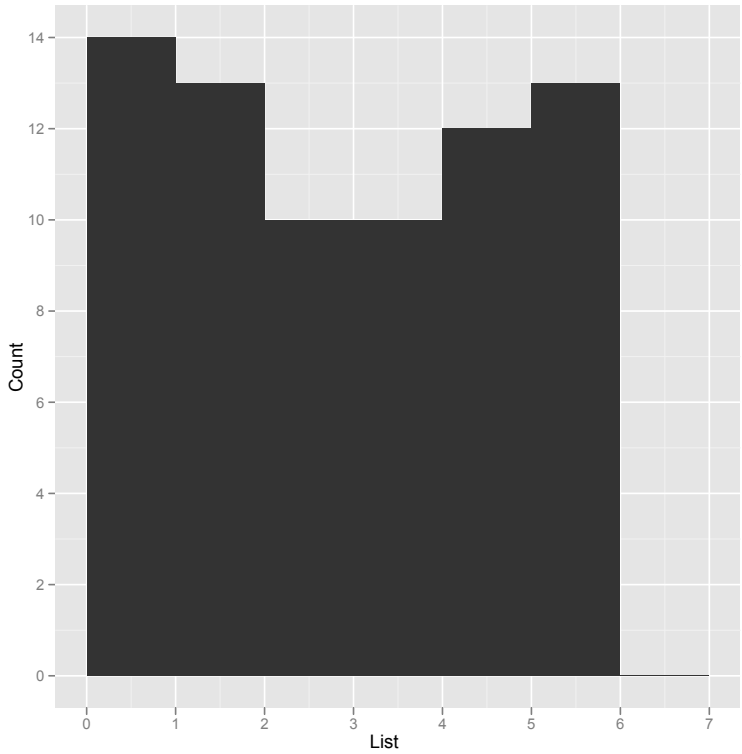


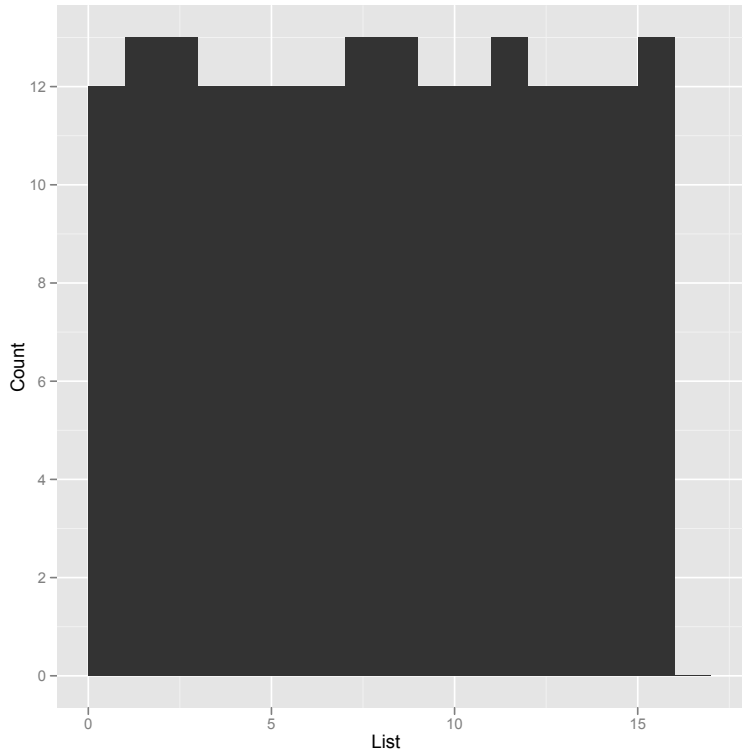
UNIVERSITY of ROCHESTER

Studies 4 and 5

- Short experiments (about 10 minutes) run with many participants:
  - Only 25 listen trials (10 primes, 15 fillers), followed by 10 production trials (4 targets, 6 fillers)
  - 80 (Study 4) to 192 participants (Study 5) in 1-3 days, each paid $1.
- Wanted 12 workers on each list for both experiments
  - Study 4: 6 lists
  - Study 5: 16 lists

List balance





UNIVERSITY of ROCHESTER

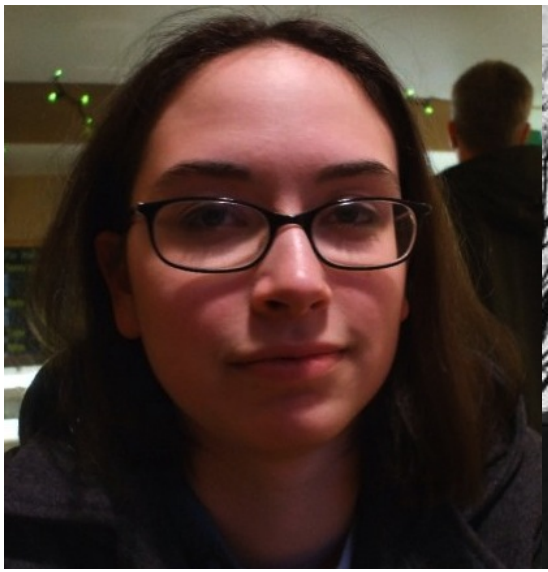
Conclusions

- Crowdsourcing can provide quality data
- Via progressive pay we can come close to achieving conditions similar to those in the lab without sacrificing the benefits of crowdsourcing
- We can achieve list balance on Mechanical Turk
- Changes how we do experiments (shorter experiments with more subjects; can avoid learning effects)

UNIVERSITY of ROCHESTER

Acknowledgements

- Thanks to Camber Hansen-Karr and Jeremy Ferris for the use of their experiment data
- Thanks to Whitney Gegg-Harrison for feedback on the presentation and general support



UNIVERSITY of ROCHESTER

Appendix

Additional material

UNIVERSITY of ROCHESTER

URLs

- External Question API docs
- Jinja2 template docs
- SQL Alchemy docs

Our List-Balancing Solution: Tech details

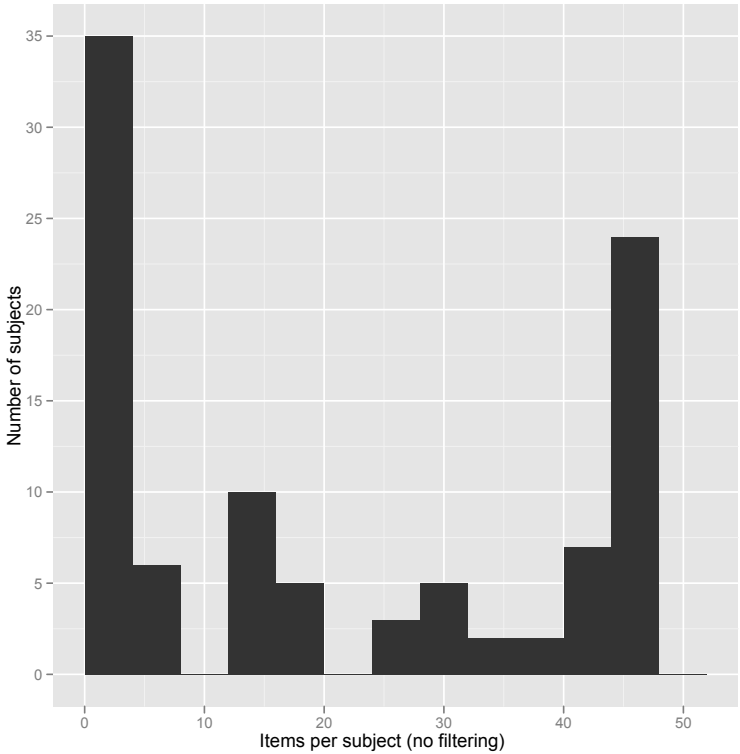
- Use the *External Question* API
- Create a database and CGI on our webserver
- When a worker accepts a HIT, assign them to a list using balancing algorithm
  - Original system used Perl w/ CGI, Template::Alloy::Velocity, and DBD::MySQL
  - Current system uses Python WSGI, Jinja2 templates, and SQL Alchemy

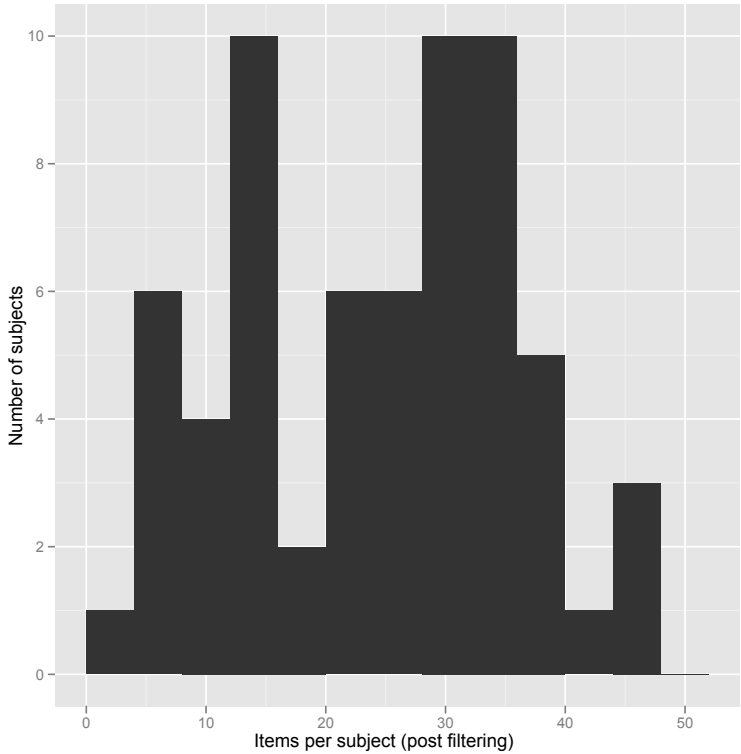
Data - Study 1

- Starting point: 3456 HITs ((12 HITs * 8 items) * 36) from 99 workers, including 2304 critical items
- Items excluded if participant didn't answer both math problems correctly. (145 cases, 6.3%)
- Items also excluded if their length (in words) didn't correspond to original sentence length or if response was like "I don't remember" or "I forgot". This excluded 505 items (22% of the data).
- Excluded all workers which contributed fewer than 5 data points
- After all exclusions, left with 1585 items, or 68.8% of original data (from 64 workers)

Items/subject - Study 1

max 48





UNIVERSITY of ROCHESTER

Experiment Design - Study 2

- 32 critical items and 64 fillers
- Each HIT consisted of one trial

Compensation - Study 2

- Each HIT paid $.02, plus
  - $.20 for per every 20 HITs

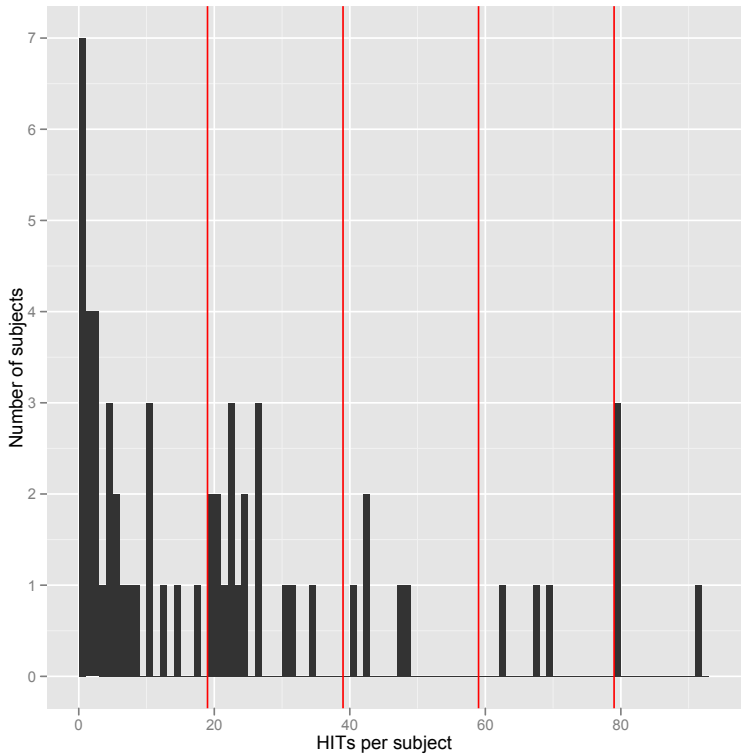
Data - Study 2

- Starting point: 1344 HITs (96 HITs * 14) from 59 workers, including 452 critical items (from 51 workers; 8 did only fillers)
- One item in list was corrupted and was excluded (16 cases) resulting in 3.6% data loss
- Items excluded if participant didn't answer both math problems correctly. (16 cases, 3.6%)
- 1 unacceptable answers and 24 different answers (excluded, 5.5%)
- 74 RConset mistakes (excluded, 16.4%)
  - 67 of which were just substitutions of pronouns for the repeated NP from the first sentence. Results don't differ either way.
- Left with 64.4% of original data (from 28 workers)

UNIVERSITY of ROCHESTER

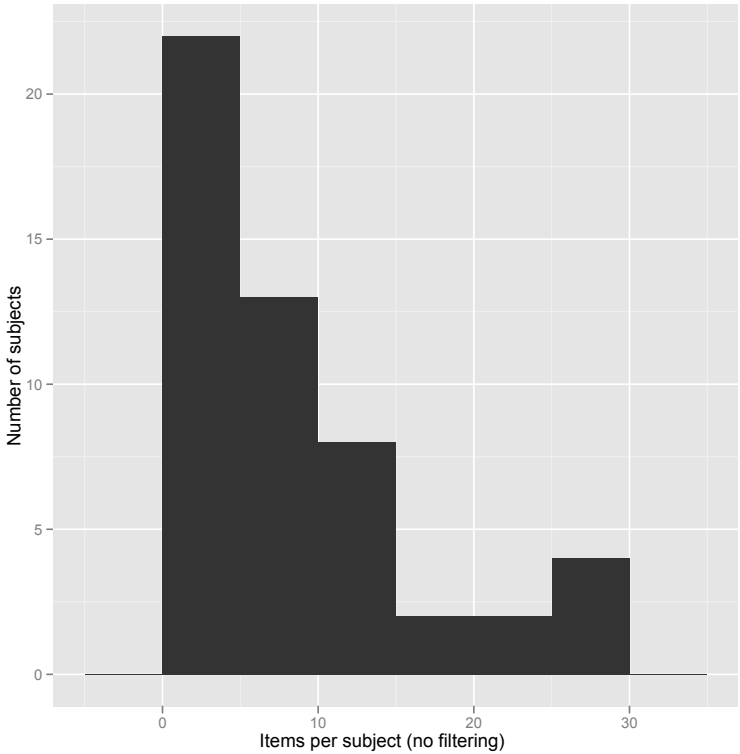
HITs per subject - Study 2

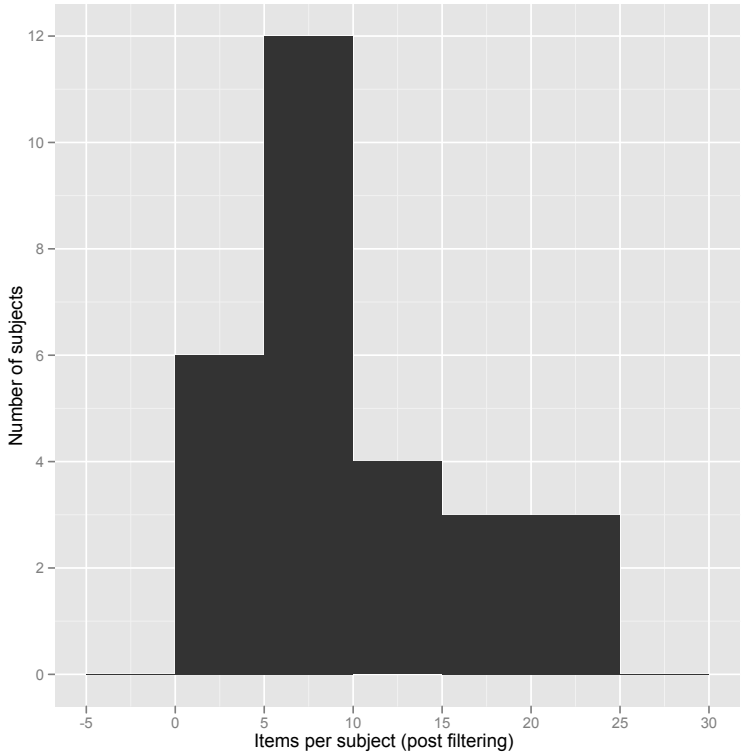
max 96



Items per subject - Study 2

max 32





Experiment Design - Study 3

- 32 critical items and 64 fillers
- Each HIT consisted of one trial

Compensation - Study 3

- Each HIT paid $.02, plus
  - $.20 for 20 HITs
  - $0.50 for 40 HITs
  - $1.25 for 80 HITs
  - $1.50 for all 96 HITs

UNIVERSITY of ROCHESTER

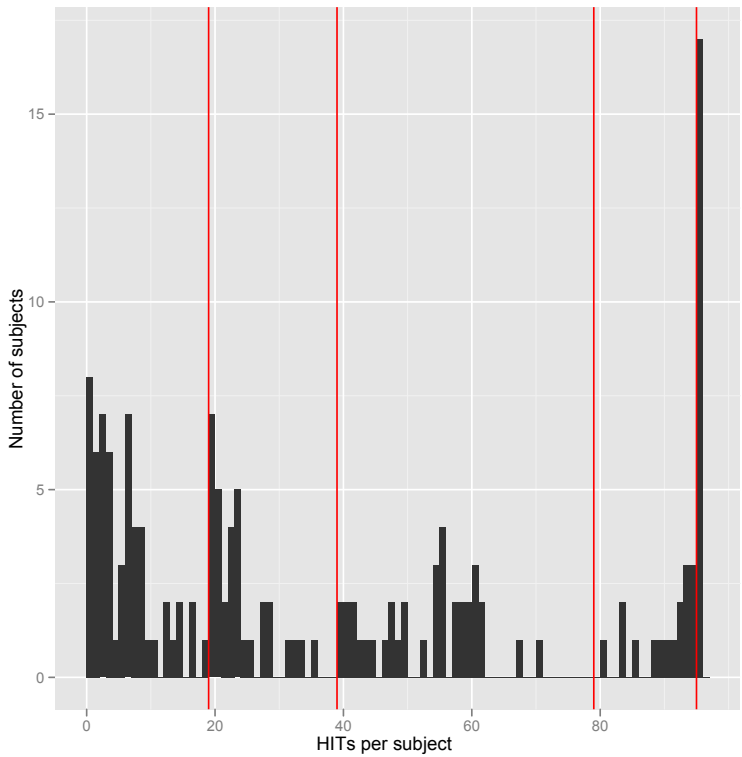
Data - Study 3

- Starting point: 6144 HITs (96 HITs * 64) from 158 workers, including 2048 critical items (from 151 workers; 7 did only fillers)
- Items excluded if participant didn't answer both math problems correctly. (192 cases)
- Items also excluded if their length (in words) didn't correspond to original sentence length. This excluded 3.3% of the data.
- Excluded all workers which contributed fewer than 5 data points
- After all exclusions, left with 87.3% of original data (from 97 workers)

UNIVERSITY of ROCHESTER

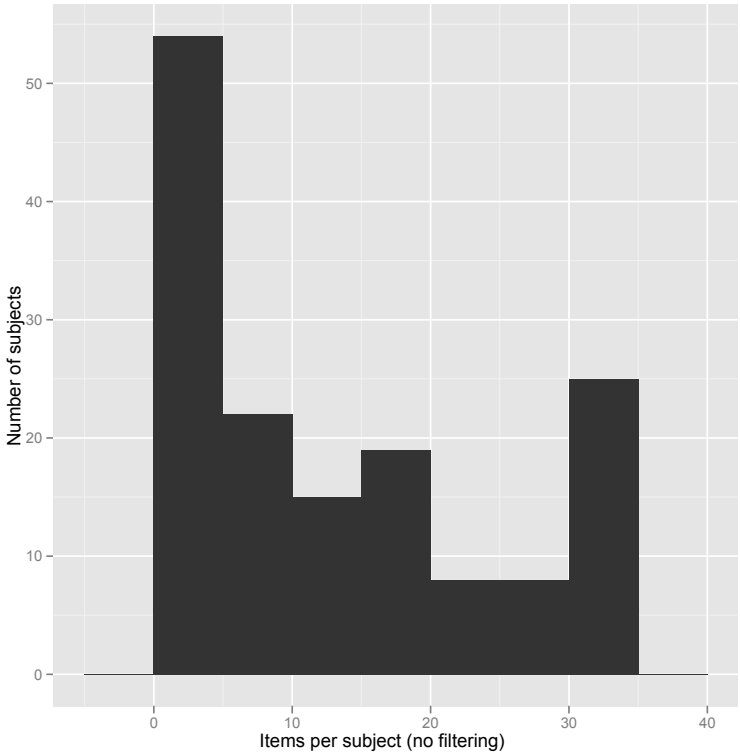
HITs per subject - Study 3

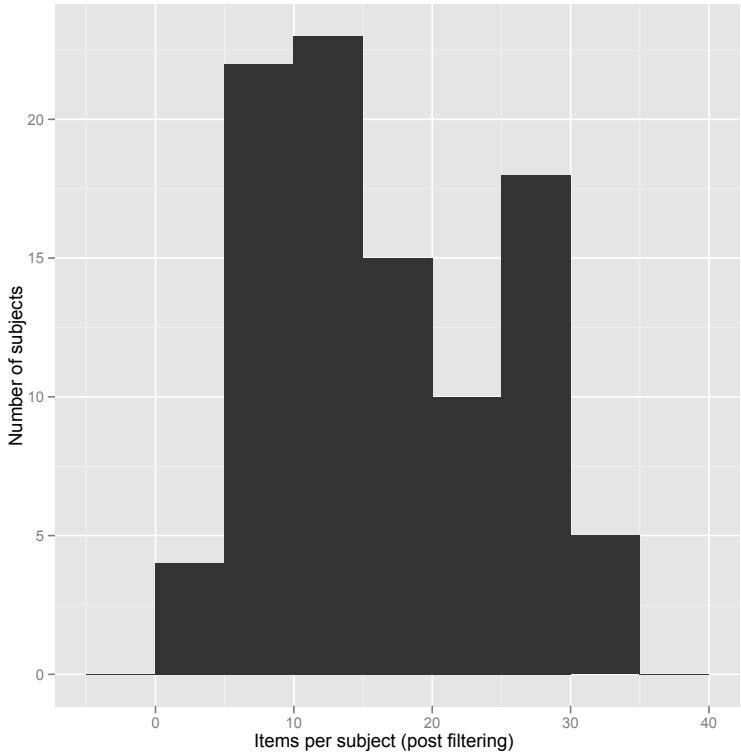
max 96



Items per subject - Study 3

max 32





UNIVERSITY of ROCHESTER