Journal of
open psychology data

Bigelow, E and Piantadosi, S T 2016 A Large Dataset of Generalization Patterns in the Number Game. *Journal of Open Psychology Data,* 4: e4, DOI: http://dx.doi.org/10.5334/jopd.19

DATA PAPER

A Large Dataset of Generalization Patterns in the Number Game

Eric Bigelow[1] and Steven T. Piantadosi[2]

[1] Primary Conductor of Study; University of Rochester, Brain & Cognitive Sciences Dept., US
ebigelow@u.rochester.edu

[2] Principal Investigator; University of Rochester, Brain & Cognitive Sciences Dept., US

Corresponding author: Eric Bigelow

We present a dataset with 272,700 two-alternative forced choice responses in a simple numerical task modeled after Tenenbaum's "number game" experiment [6]. Subjects were shown a set (e.g. {16, 12}) and asked what other numbers were likely to belong to that set (e.g. 1, 5, 2, 98). Their generalization patterns reflect both rule-like (e.g. 'even numbers,' 'powers of two') and distance-based (e.g. 'numbers near 50') generalization. This dataset is available for further analysis of these simple and intuitive inferences, developing of hands-on modeling instruction, and attempts to understand how probability and rules interact in human cognition.

Keywords: Generalization; Bayesian inference; Structured cognitive model; Numerical cognition; Concept learning

Funding statement: We are grateful to the University of Rochester for funding this research.

(1) Overview

Collection Date(s)

March–April 2015.

Background

Numbers and related mathematical ideas form a complex set of interrelated concepts that can be used to study the origin and use of structured mental representations. To examine learning and generalization in this domain, we present an extension of the "number game" task originally developed by Tenenbaum [6]. In the number game, a subject is given a list of numbers sampled from an unknown rule. Subjects are asked to generalize from the samples and predict what other numbers ("targets") are likely to obey the rule. For example, a subject could be told that an unknown program generated the numbers {4, 16, 8}, and then is asked to rate whether 12 might be generated as well. In this example, subjects might rate 12 as relatively unlikely since the observed data suggests a rule like 'powers of two'. However, if the shown set were instead {4, 16, 8, 10}, the concept of 'even numbers' now seems like a better explanation than 'powers of two,' and subjects should generalize accordingly. The simplicity of this setup provides an simple toy domain for studying rule-like generalization: the set of hypotheses is likely to be very simple and concrete (e.g. basic arithmetic concepts), the input sets provided to subjects are small, and the findings are intuitive.

Tenenbaum [6, 7, 8] showed that subject generalizations in this task followed statistically sensible inferences combining both rule-like generalizations (e.g. 'even numbers,' 'powers of two,' 'multiples of ten') and magnitude/similarity-based generalizations (e.g. 'numbers near 50'). For instance, subjects' patterns of generalization depend strongly on the amount of data provided, consistent with models that quantify the likelihood of sampling the observed set given a possible rule. For example, if the participant sees {80, 10, 30}, the likelihoods for the hypotheses 'multiples of 10' and 'multiples of 5' should be significantly higher than for 'multiples of 2' since the latter includes more numbers and thus assigns {80, 10, 30} a smaller likelihood of being sampled. With short input lists like {8, 32}, or even single-number lists like {8}, responses are much more revealing of a priori inductive biases. For instance, given {8}, we can measure whether a participant prefers to generalize to 10 (suggestive of 'even numbers'), or 16 (suggestive of 'powers of two'). A comparison of these generalization therefore may tell us what types of numerical concepts subjects possess before our experiment – that is, the mathematical concepts that are most likely (highest prior) before data is observed. The number game task has also been used to study information gathering behavior [4], similar to that of Wason's 2-4-6 task [9].

In Tenenbaum's original task [6], 8 subjects were presented with the same 8 sets of input numbers, each with a list of 30 hand-selected targets. Subjects rated each number on a scale of 1–7, according to how likely the number was to be accepted by the program; they were instructed to take as much time as needed. Our experiment aims to replicate this general framework, but with a much larger numbers of subjects, concepts, and targets. We constructed a large space of concepts by applying several simple and intuitive transformations (e.g. x → 2^x) across a variety of basic number patterns (e.g. 'even numbers,' 'prime numbers,' etc.). We then sampled 255 sets from these concepts, and tested approximately 10 subjects on a two alternative forced choice (in the set or not) for the targets 1, 2, . . ., 100 for each input set. This resulted in a total of 606 subject participants, providing 253,564 total responses after trimming. We used a two-alternative forced choice instead of Tenenbaum's Likert scale rating in order to simplify later data analysis and model comparisons. Our binary response data is easily incorporated into either mixed effect logistic regressions, or as a binomial likelihood in more general probabilistic models. In forthcoming work, we present analysis aimed at capturing people's priors in the context of a Bayesian data analysis and model comparison [1].

(2) Methods

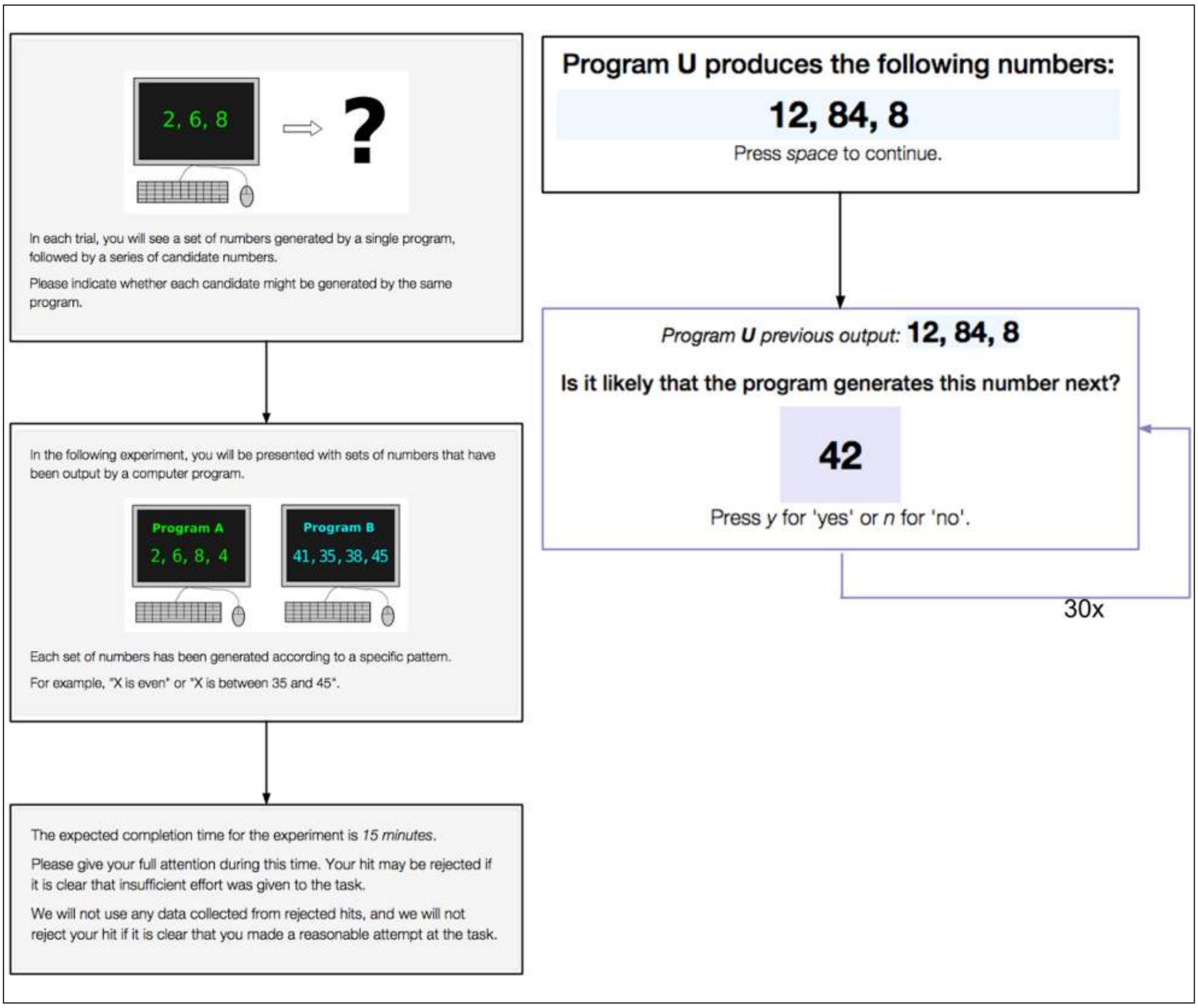
Sample

606 participants were recruited via Amazon Mechanical Turk (MTurk). We configured PsiTurk [3] to require participants to have an approval rating of at least 95% and be from the U.S. (note that the latter constraint is not absolute, as a small number of workers from outside the U.S. are able to circumvent this qualification). The first run included 510 subjects, of which 25 were rejected on qualitative criteria of not adequately attempting the task (see *Quality Control*). Rejected MTurk Human Intelligence Task data (HITs) were replaced in follow-up runs. After the first 485 subjects, an additional 126 were run and five were rejected. The experiment was run in March and April of 2015.

The modal education level participants reported was a "Bachelor's degree," accounting for 37.6% of subjects, followed by "some college" with 24.9%. Mean age of participants was 34.2, the median was 31.0, and the standard deviation 11.0. Reported gender was nearly even, with 51.1% of participants being male. The vast majority of subjects (97.4%) reported English as their first language. Age and education level for our subjects is typical of the MTurk population [2], though we find a higher than average proportion of male subjects.

Materials

Our version of the number game task was implemented in HTML and JavaScript, and distributed to Mechanical Turk workers via the PsiTurk interface [3]. **Figure 1** (left)



In each trial, you will see a set of numbers generated by a single program, followed by a series of candidate numbers.

Please indicate whether each candidate might be generated by the same program.

In the following experiment, you will be presented with sets of numbers that have been output by a computer program.

Each set of numbers has been generated according to a specific pattern.

For example, "X is even" or "X is between 35 and 45".

The expected completion time for the experiment is *15 minutes*.

Please give your full attention during this time. Your hit may be rejected if it is clear that insufficient effort was given to the task.

We will not use any data collected from rejected hits, and we will not reject your hit if it is clear that you made a reasonable attempt at the task.

Program U produces the following numbers:

12, 84, 8

Press *space* to continue.

Program U previous output: **12, 84, 8**

Is it likely that the program generates this number next?

42

Press *y* for 'yes' or *n* for 'no'.

Figure 1: Sequence of instructions shown before the experiment begins (left) and the display shown for main data collection (right).

shows the sequence of instruction pages shown before the beginning of the experiment. Once the experiment began, subjects rated a series of targets for each of 15 sets, shown in **Figure 1** (right). For each concept, the subject first saw a screen with the concept shown on it, and when they were ready they proceeded by pressing the spacebar. The subject then saw a target, and responded 'yes' or 'no' to the question, "Is it likely that the program generates this number next?" by pressing either the 'y' or 'n' key. Once they responded, another target was shown. This was repeated for 30 targets, until the first screen for the next set.

Stimuli

Sets

Numerical sets used as stimuli were constructed with the goal of spanning an interesting space of generalization stimuli. To generate sets, first we generated a collection of concepts. Beginning with six "primordial sets": all numbers, evens, odds, squares, cubes, and primes, the following functions were then mapped across each primordial set: f(n) = n, f(n) = n + 1, f(n) = n – 1, f(n) = n + 2, f(n) = n – 2, f(n) = 2 * n, f(n) = 3 * n, f(n) = 2 * n + 1, f(n) = 3 * n + 1, f(n) = 3 * n – 1, f(n) = 2^n, f(n) = 2^{n+1}, f(n) = 2^n + 1, f(n) = 2^n – 1. Numbers in our task were restricted to the domain of integers 1 through 100. We selected these primordial sets and functions in order to span concepts studied in previous work [6, 7, 8], and also to include some degree of pseudo-random number sets – for example, subjects are likely to perceive 2^{primes} – 1 (i.e. {3, 7, 31}) as being a set of random numbers, perhaps limited to some interval. Duplicates and extremely short (length < 3) full concepts were removed. We then added 21 additional full concepts: 4 * n, 5 * n, 6 * n, 7 * n, 8 * n, 9 * n, 10 * n, 5 * n + 1,2,3,4, and 10 * n + 1, . . . 9.

Given these 79 full concepts, we generated sets by the following procedure: if the length of the full concept was greater than 4, we chose 4 random numbers from the list without replacement; if the length was greater than 3, we chose 3 numbers; if the length was greater than 2, we chose 2 numbers. For each full concept, between 1 and 3 sets were created, for a total of 200 sets. Finally, 55 single-item sets were added to these 200, for a total of 255 sets. 16 of the single-item sets were hand selected, all integers 1 through 15 and 100, and the rest were chosen randomly from the range of 16 to 99. See **Table 1** for a full list of sets.

While we generated sets like {16, 8} from underlying concepts (e.g. 'powers of 2'), analysis of our data should primarily be interested in what generalizations the set {16, 8} leads subjects to, *not* whether subjects can recover the generating concept itself. There will often be too little data to infer the generating concept, particularly given the profusion of close alternative concepts (e.g. 'numbers between ten and twenty'). Examination of subjects' generalization from very little data like these small sets will be informative about their underlying inductive biases.

Targets

These 255 sets were divided into 17 groups of 15 sets each, where each participant assigned one of these groups. For each set presented to a participant, 30 targets (in 1. . .100) were shown, randomly selected without replacement, so that each participant made 450 decisions. All together, at least nine two-alternative forced-choice ratings were collected for each number from 1 to 100. Due to a small randomization error, targets for each set were slightly non-independent relative to one another, with no obvious effect on the experiment.

Procedures

After seeing the instructions, subjects provided 30 ratings for each of 15 different sets. At the conclusion of these forced-choice trials, the subject filled out a brief questionnaire. Basic demographics were collected: age, gender, first language, ZIP (postal) code, and highest level of education. Subjects were also asked to describe in English 5 sets randomly selected from the 15 they were shown during the experiment.

Quality Control

HITs were rejected based on qualitative criteria, under the determination that the task was not properly attempted. Many rejected HITs were exceedingly fast, including a number of HITs completed in less than 5 minutes. Other rejected HITs included significant repetitive answering patterns, such as many 'yes' responses followed by many 'no's, or alternating 'yes'/'no'. A small fraction of reaction times were corrupted during data recording; these were replaced with NA in the dataset.

The Dataset

The dataset contains a single row for each response collected, with columns for rating (1 for 'yes,' 0 for 'no'), set ("set"), target, subject id ("id"), trial number for subject ("trial"), reaction time ("rt"), subject demographics, number of HITs for this set and target pairing ("hits"), as well as probability of responding yes ("p"), entropy of responses ("H"), and a typicality measure. Probability was calculated as the number of 'yes' responses for a given set and target pairing, divided by the total number of responses for this pairing. Entropy was calculated as: – (p*log(p) + (1 – p)*log(1 – p)). Though not included, we have also used a response typicality metric to identify subjects whose responses strayed far from the average (e.g. due to low task effort) – calculated as log(p) for 'yes' ratings, and log(1–p) for 'no' ratings.

To illustrate our collected data, **Figure 2** shows ratings for targets under three related sets: {3, 63}, {33, 3}, and {93, 43, 83, 53}. Plots in the left and right columns of the figure are colored according to two distinct rules, 'numbers ending in 3' (left) and 'multiples of 3' (right). The data for {93, 43, 83, 53} strongly supports the first, but not the second, pattern. The data for {3, 63} is more uniform, and seems to correspond more closely to 'multiples of 3'. The distribution for {33,3} may be interpreted as a mixture of the two patterns – 'numbers ending in 3' are generally rated highest, with some probability mass assigned to 'multiples of 3'. The human ratings illustrate how the set may push generalizations one way or the other and reveal both categorical (all-or-nothing) and gradient generalizations across subjects.

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15
16	17	18	22	23
25	26	30	31	33
35	36	39	43	44
47	48	49	50	51
53	55	61	62	64
66	67	69	72	73
78	81	84	85	86
89	91	93	95	100
75, 4	81, 9	5, 51	13, 91	98, 83
16, 8	33, 9	3, 31	52, 24	25, 17
85, 7	71, 11	64, 4	5, 65	3, 63
55, 29	6, 74	3, 87	63, 67	94, 70
8, 92	2, 8	33, 3	7, 31	81, 25
26, 2	24, 35	83, 11	14, 47	50, 2
75, 27	19, 73	28, 13	2, 47	8, 64
28, 2	7, 63	10, 3	6, 25	16, 54
3, 81	55, 3	25, 82	23, 80	53, 2
60, 54	22, 96	59, 3	34, 26	15, 93
15, 11	94, 7	92, 56	8, 32	8, 16
5, 9	3, 7	83, 77	70, 15	6, 66
7, 67	73, 33	59, 14	10, 80	21, 71
42, 62	63, 43	64, 44	75, 95	76, 26
37, 57	68, 58	79, 59	84, 56	66, 78
28, 98	96, 48	90, 45	87, 8, 52	66, 93, 51
7, 51, 23	10, 55, 58	29, 62, 98	32, 4, 16	33, 3, 65
31, 3, 1	50, 76, 28	61, 9, 45	85, 19, 91	5, 77, 89
33, 5, 19	10, 74, 22	63, 27, 81	67, 99, 15	16, 82, 28
20, 32, 92	100, 1, 16	65, 26, 2	48, 99, 3	38, 18, 3
23, 62, 98	98, 18, 50	75, 48, 3	73, 33, 3	49, 76, 13
47, 2, 74	1, 8, 27	9, 65, 28	10, 66, 29	53, 47, 41
98, 54, 18	4, 16, 12	69, 19, 85	81, 87, 27	6, 34, 82
51, 39, 87	15, 39, 35	52, 22, 94	92, 68, 20	77, 17, 8
90, 100, 5	16, 26, 96	82, 67, 72	48, 63, 53	64, 94, 84
10, 50, 100	71, 31, 21	62, 32, 22	63, 23, 43	84, 94, 34
15, 85, 45	46, 76, 66	27, 37, 67	78, 48, 88	59, 89, 79
12, 84, 8	48, 30, 78	14, 98, 91	8, 80, 48	18, 45, 72
3, 33, 99	65, 47, 55, 13	39, 60, 12, 45	73, 35, 53, 59	31, 19, 49, 100
11, 41, 38, 92	16, 32, 2, 8	33, 17, 5, 9	31, 3, 1, 15	20, 8, 84, 100
29, 77, 37, 17	91, 25, 61, 31	17, 11, 65, 53	29, 9, 3, 31	34, 86, 30, 94
63, 27, 99, 33	99, 59, 31, 67	70, 10, 88, 40	68, 14, 8, 26	4, 64, 25, 81
17, 50, 5, 82	8, 24, 48, 63	6, 51, 66, 27	79, 47, 62, 98	18, 8, 72, 98
75, 48, 12, 3	9, 33, 73, 99	49, 28, 76, 13	11, 26, 74, 2	67, 31, 17, 61
98, 24, 30, 3	4, 66, 78, 96	43, 4, 91, 9	15, 17, 21, 5	22, 62, 26, 82
69, 9, 39, 21	23, 35, 95, 83	70, 88, 7, 22	92, 14, 20, 5	68, 20, 35, 92
55, 60, 25, 100	71, 61, 26, 81	92, 82, 27, 42	83, 38, 58, 68	59, 34, 14, 89
80, 70, 90, 100	81, 71, 21, 31	72, 92, 82, 42	93, 43, 83, 53	54, 14, 94, 34
55, 15, 75, 45	96, 86, 36, 66	97, 77, 87, 47	48, 78, 38, 98	59, 49, 99, 69
76, 44, 8, 48	96, 90, 6, 42	42, 35, 21, 84	64, 96, 24, 56	36, 45, 54, 81

Table 1: A complete list of sets used in experiment, sorted by length. Each cell shows a set of numbers that was presented to subjects, who then decide what other numbers in 1 . . . 100 are likely to be generated.

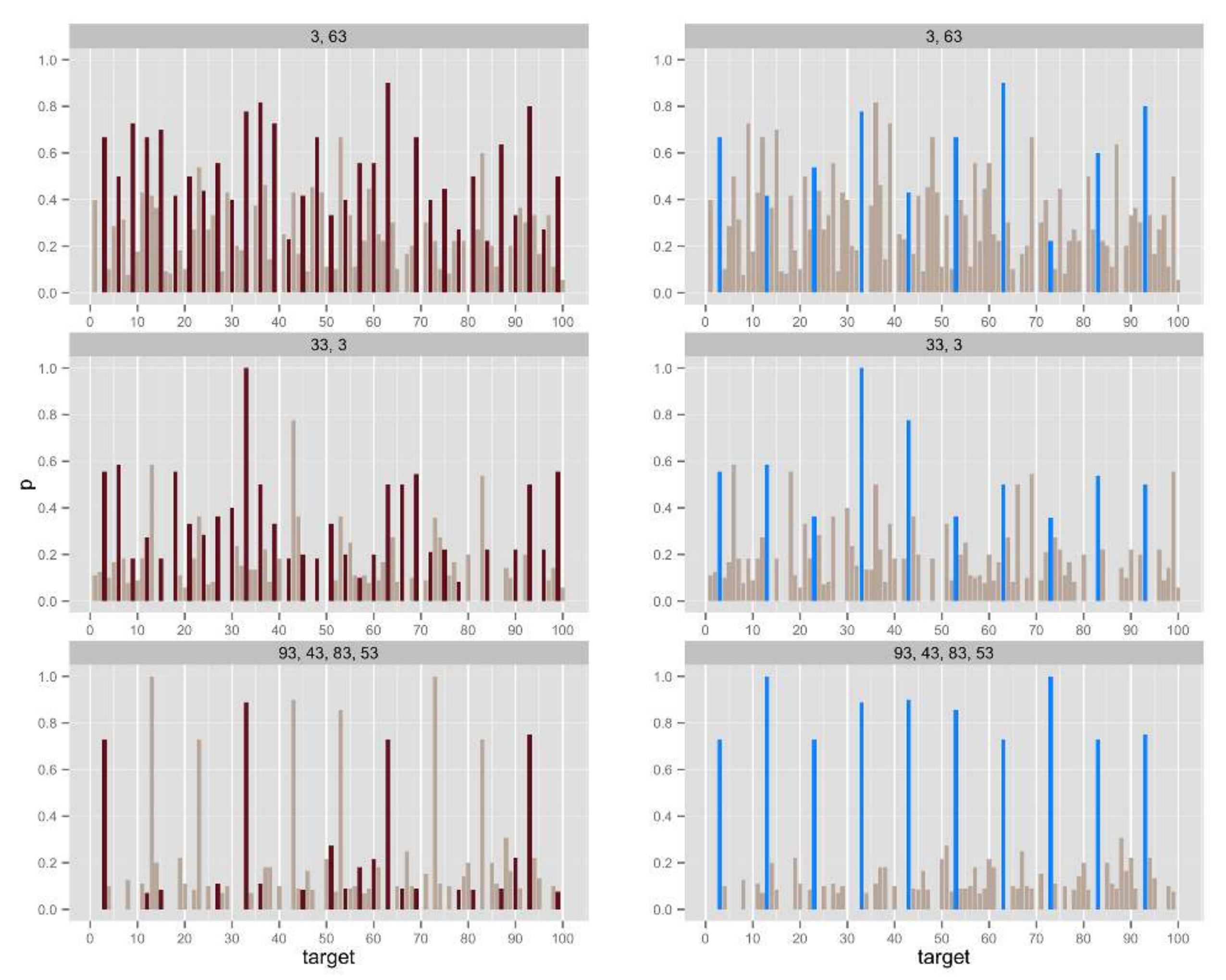


Figure 2: Human predictive distributions for 3 concepts (rows), selectively highlighting target numbers corresponding to 'multiples of 3' (left column) and 'ends in 3' (right column).

While we intended our experiment to present subjects with *sets* (unordered collections), it is possible that some subjects interpreted the goal as generalizing from *sequences* (ordered data). Further analysis may distinguish these possibilities in detail, although the sets were presented in random order and many subjects did not use sequential terminology in their descriptions of the concepts. Out of 3030 concept descriptions, only 21 included sequential terms such as "increasing order," "decreasing," or "ascending".

Ethical issues

All data presented collected in this experiment was anonymized prior to public release. This work was approved by the Research Subjects Review Board at the University of Rochester as part of a protocol for experimental data collection on Mechanical Turk.

(3) Dataset description

Object name

Our dataset consists of one primary file, *numbergame_data.csv* (described above), and 3 supplementary files including additional subject data:

- *instructions_rt.csv* : time spent looking at each instruction page
- *set_descriptions.csv* : qualitative set descriptions provided in questionnaire
- *show_set_rt.csv* : time spent looking at set presentation page, before responding to targets (see **Figure 1**, right)

To prevent ambiguity in file loading, commas have been replaced with underscores in the 'set' column of *numbergame_data.csv*, *set_descriptions.csv*, and *show_set_rt.csv*, and in the 'descr' column of *set_descriptions.csv* .

Data type

Raw data file.

Format names and versions

All data is in comma-delimited CSV format; scripts provided in R and python.

Data Collectors

Eric Bigelow.

Language
English.

License
Create Commons Attribution (CC-By).

Embargo
NA.

Repository location
DOI: http://dx.doi.org/10.7910/DVN/A8ZWLF
URL(alternate): https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FA8ZWLF&version=DRAFT

Publication date
Submitted for review – (5/20/15); edited and resubmitted – (7/28/15).

(4) Reuse potential

This data will be useful to basic research on human conceptual representation and generalization. Because of the simplicity of the task, we expect that it will provide a compelling teaching example, showing ways in which structured concepts influence generalization in a domain that is both simple and intuitive. Work using this data to infer the priors on human concepts is ongoing as part of LOTlib, a Python library for Language of Thought models [5]. Information about this project is available here: https://github.com/piantado/LOTlib.

Competing Interests
The author declares that they have no competing interests.

References

1. **Bigelow, E** and **Piantadosi, S T** (*under review*) Inferring priors in compositional cognitive models.
2. **Ipeirotis, P G** 2010 Demographics of Mechanical Turk (Tech. Rep. No. CeDER-10-01). New York: New York University. Retrieved from http://hdl.handle.net/2451/29585.
3. **McDonnell, J V, Martin, J B, Markant, D B, Coenen, A, Rich, A S** and **Gureckis, T M** 2014 psiTurk (Version 2.1.1)[Software]. New York, NY: New York University.
4. **Nelson, J D, Tenenbaum, J B** and **Movellan, J R** 2001. Active inference in concept learning. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the 23rd Conference of the Cognitive Science Society*, Mahwah, NJ: Erlbaum, pp. 692–697. Available at http://www.hcrc.ed.ac.uk/cogsci2001/pdf-files/0692.pdf.
5. **Piantadosi, S T** 2014 LOTlib: Learning and Inference in the Language of Thought [software]. Accessible from: https://github.com/piantado/LOTlib.
6. **Tenenbaum, J B** 1999 *A Bayesian framework for concept learning* (Doctoral dissertation, Massachusetts Institute of Technology).
7. **Tenenbaum, J B** 2000 Rules and similarity in concept learning. *Advances in neural information processing systems*, 12: 59–65.
8. **Tenenbaum, J B** and **Griffiths, T L** 2001 Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4): 629–640. DOI: http://dx.doi.org/10.1017/S0140525X01000061
9. **Wason, P C** 1960 On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12: 129–140. DOI: http://dx.doi.org/10.1080/17470216008416717

Peer review comments: http://dx.doi.org/10.5334/jopd.19.pr

How to cite this article: Bigelow, E and Piantadosi, S T 2016 A Large Dataset of Generalization Patterns in the Number Game. *Journal of Open Psychology Data,* 4: e4, DOI: http://dx.doi.org/10.5334/jopd.19

Published: 18 March 2016