Replication of "Seeking confirmation is rational for deterministic hypotheses"
by Joseph L. Austerweil & Thomas L. Griffiths (2011, *Cognitive Science*)

Natalia Vélez
nvelez@stanford.edu

Introduction

Confirmation bias, the tendency to select information that confirms one's current hypothesis, is a ubiquitous and well-studied bias in human decision making. The present study argues that confirmation bias is the optimal strategy for testing deterministic hypotheses. In the original study, participants were trained by viewing number sequences that were predominantly generated by a single rule. In a test phase, they viewed a new set of sequences, predicted which number came next, and provided a hypothesis of the rule that generated the sequence. The results of the original study demonstrate that a simple Bayesian model can capture people's predictions about which numbers come next. We replicated Experiment 4 of their study, which showed that manipulating the prior probabilities of hypotheses changed the hypotheses that participants chose to test.

Methods

Power Analysis

The key result of the original study is that participants select tests that confirm the hypothesis that is most probable during the testing phase. Half of the participants were trained on 100 number sequences where, in 89 sequences, the next number of the sequence was generated by summing the previous two (*sum last two*). Conversely, half of the participants were trained on 100 number sequences where, in 87 sequences, the next number was generated by multiplying the last number by a constant *C* and/or adding a constant *K* (*xC+K*).

In the original study, participants who were trained on the sum last two sequences tested the sum last two rule more often than those in the xC+K condition. The original effect size for this comparison is $\chi^2(2) = 8.14,\ p < 0.05$, with n = 29. Using R's power analysis library (pwr), we determined the original effect size was $w = 0.58$. Achieving 80%, 90%, and 95% power with this effect size and $p < 0.05$would require 29, 38, and 46 subjects, respectively. Because of the ease of acquiring large samples on Amazon Mechanical Turk, we will attempt replication at 96.5% power with 50 subjects.

Planned Sample

46 US participants from Amazon Mechanical Turk will participate in our study. To qualify for the study, participants will need to have completed at least 100 Human Intelligence Tests (HITs) and received a HIT approval rating of 95% or greater.

Materials

In this study, participants will be trained on 100 sequences of numbers and test their hypotheses on 21

new number sequences. The original study describes the number sequences as follows:

> In order to establish the priors in different sequence prediction environments, participants in the xC + K and *sum last two* conditions were trained on 100 sequences of numbers. The training sequences in the xC + K condition had a high prevalence (87%) of sequences generated by rules of the form ‘‘xC + K’’ and no sequences generated by summing the last two numbers, and vice versa in the sum last two condition (with 89% of sequences conforming to the ‘‘sum of the last two numbers’’ rule). No training was provided in the control condition. Test selection was probed with 21 sequences consistent with both the ‘‘sum of the last two numbers’’ and the ‘‘xC + K’’ rules, shown to participants in all three conditions. For example, one of the 21 test sequences, (3,6,9), can be interpreted as ·1 + 3 or the sum of the last two numbers (3 + 6 = 9).

Procedure

The procedure follows the original experiment. All participants will give informed consent in accordance with the requirements of the internal review board at Stanford University. The original study describes the procedure as follows:

> The experiment had two phases: a feedback phase and a test phase. In the feedback phase, participants were given the task of discovering the underlying rule of a sequence for 100 sequences. They were told that to achieve the goal, they got to choose a number and they would receive feedback as to whether or not it is the next number in the sequence. They were reminded that more than one rule could fit the sequence they saw, so that they should pick the number that helps them figure out the underlying rule as best as possible. For the test phase, participants were told that it was identical to the first part, except that they would not be given feedback, but that they should still make their choices as if it were given. They were asked to write down both what they thought the rule was and their number choice. The experiment was administered on a computer with instructions read by an experimenter. The participants were also provided a calculator to ensure that arithmetic ability was not a factor in people’s responses.

Our replication will follow this procedure exactly, except that the calculator will be provided as an applet on the experiment website.

Analysis Plan

A coder who is blind to participants’ assignment to conditions during the feedback phase will code their responses in the test phase according to the hypothesis being tested: *sum last two*, xC+K, or other. We will use chi-squared tests to compare how often each type of hypothesis is invoked during the test phase in different training conditions.

Differences from Original Study

Our subjects will be online respondents on Amazon Mechanical Turk, not undergraduates at UC Berkeley. This difference motivated one deliberate change to the methodology: instead of being provided with physical calculators, participants will use a calculator applet provided on the experiment page. The context and setting in which the experiment is carried out will also inevitably vary between subjects. We

do not foresee that these changes will affect our ability to replicate the study.

(Post Data Collection) Methods Addendum

Actual Sample

Our actual sample was 50 US workers from Amazon Mechanical Turk (26 male), ages 18-69 (mean = 33.06, SD = 10.71). 22 participants were randomly assigned to the "xC + K" condition, and 28 were assigned to the "sum last two" condition. Most participants reported having received "some college" education, and the distribution of education levels was matched across the two conditions ($\chi^2(8) = 10$, p = 0.27). We did not exclude any participants.

Differences from pre-data collection methods plan

Due to the constraints of online-based studies, we halved the length of the training phase. Thus, participants saw 50 training sequences, rather than 100, in which the two classes of rules appeared in similar proportions as the original study: 44 sequences were generated by the predominant rule, and 6 were generated by the other rule. In the original study, all participants saw the same training sequences. However, because our replication was carried out online, it would have been simple for participants to look up all of the answers to the training phase by looking at the source code. Thus, training sequences were randomly generated in order to discourage cheating, while test sequences were preassigned because they needed to be ambiguous. We do not expect that these changes will affect our ability to carry out the replication.

Results[1]

Data preparation

Data preparation followed the analysis plan. Subjects' responses were recorded as JSON files, which were then parsed using R's jsonlite library. In order to code participants' responses, participants were assigned a random ID not tied to their worker ID, and information about their condition assignment was removed. Condition data were not added back to the test phase data until all responses had been coded.

Responses were classified into three categories: "xK+C," "sum last two," and "other." Responses that described adding or multiplying by some number (e.g., "add 20," "multiples of 2," "add 7 to the previous number") were marked as "xK+C" responses. Responses that explicitly referenced summing the last two numbers (e.g., "add the two numbers before," "sum the last two numbers") were coded as "sum last two" responses. Finally, responses that did not fall neatly into each category (e.g., "add 20 to the sum of the last two numbers") or where the participant did not provide a rule (e.g., "no idea," "don't know") were coded as "other".

Confirmatory analysis

In the original study, the critical test of the authors' claim—that changing the prior probability of different rules changes which hypotheses people choose to test—was a chi-square test comparing the mean number

[1] The raw data and documented analysis scripts for this replication can be found here.

of response types across the two training groups (Figure 1). In our results, however, it seems that the proportion of responses of each type did not vary between the two training groups ($\chi^2(2) = 2.54$, $p = 0.28$). One salient difference between our replication and the original study is the incidence of "sum last two" responses in the "xK+C" group. In the original study, the mean number of "sum last two" responses was 0; in our replication, that number rose to 2.78.

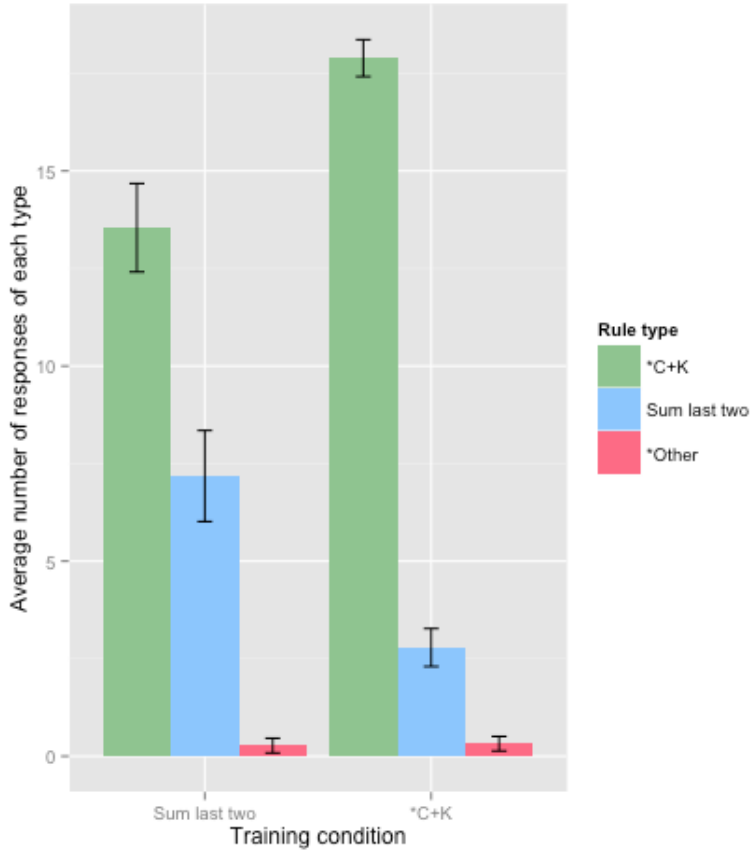

Figure 1: Mean number of responses of each type in the test phase. The divisions on the x-axis correspond to the predominant rule that participants saw during the training phase; e.g., participants in the "sum last two" group saw 44 number sequences of the form "sum last two" and 6 sequences of the form "xK+C" during the training phase. Responses were coded using the conventions described above.

Exploratory analyses

Excluding other responses

One possible issue with the critical analyses is that the chi-squared test may be inaccurate when there are too few responses in one cell—in particular, in our data, there were very low numbers of "other" responses. Therefore, we excluded these from our analyses and compared only the average proportion of sum last two and xK+C responses in the test phase. Here, we again did not find that the dominant rule during the training phase influenced which rules participants chose to test ($\chi^2(1) = 2.54$, $p = 0.11$).

Logistic regression

The previous analyses suggest that there is not a statistically significant relationship between training

condition and the frequency with which participants cited different classes of rules during the test phase. However, the chi-squared test is a relatively coarse test of this relationship; for example, it does not account for intersubject variability or item-specific effects.

Therefore, we used a mixed-effects logistic regression to compare the probability of participants citing an xK+C rule (which were coded as a 'success') during the test phase. Responses classified as 'other' were excluded from the analysis in order to minimize small-sample bias in the model. Number sequence and subject number were both treated as random effects. Our model suggests that training condition indeed predicts response type: participants in the xK+C training condition were more likely to provide xK+C responses than participants in the sum last two condition ($\beta = 2.82$, $z = 3.92$, $p < 0.001$).

Accuracy & reaction time

As noted above, participants were not evenly split across conditions: more participants completed the sum last two condition than the xK+C condition. One possible reason for this discrepancy is that participants were simply unevenly assigned to conditions. Another possibility is that the rates of attrition differed for each condition; for example, participants who were assigned to the xK+C condition could have returned the HIT more often than participants who were assigned to the sum last two condition. While we did not record completion rates for each condition, we did collect two measures during the test phase—response time and accuracy—which may provide a clue as to whether one of the conditions was more difficult than the other.

We first used a mixed-effects logistic regression, with subject number as a random effect, to analyze the effects of training condition and sequence type on accuracy (Figure 3). Subject number was treated as a random effect; because training sequences were randomly generated for each subject, we did not include number sequence as a random effect. Overall, participants in the xC+K condition were less accurate ($\beta = -1.88$, $z = -2.28$, $p < 0.05$). There was also an interaction between rule type and training condition on accuracy ($\beta = 4.53$, $z = 8.11$, $p < 0.001$): while there were no differences in performance by question type in the sum last two group, participants in the xC+K group were particularly inaccurate when predicting sum last two sequences.

We next used a mixed-effects linear model to examine the effects of training condition and sequence type on reaction time (Figure 2). Our results suggest that there was an effect of training condition on response time: on average, participants in the xC+K condition were 56.43 seconds slower to complete training number sequences than participants in the sum last two condition ($\beta = 56.43$, $t = 2.34$). Participants seem to have been particularly slow to complete sum last two sequences in the xK+C condition. However, response times were highly variable; thus, there was neither an effect of sequence type ($t = -0.16$) nor an interaction between training condition and sequence time ($t = -1.30$) on response time.

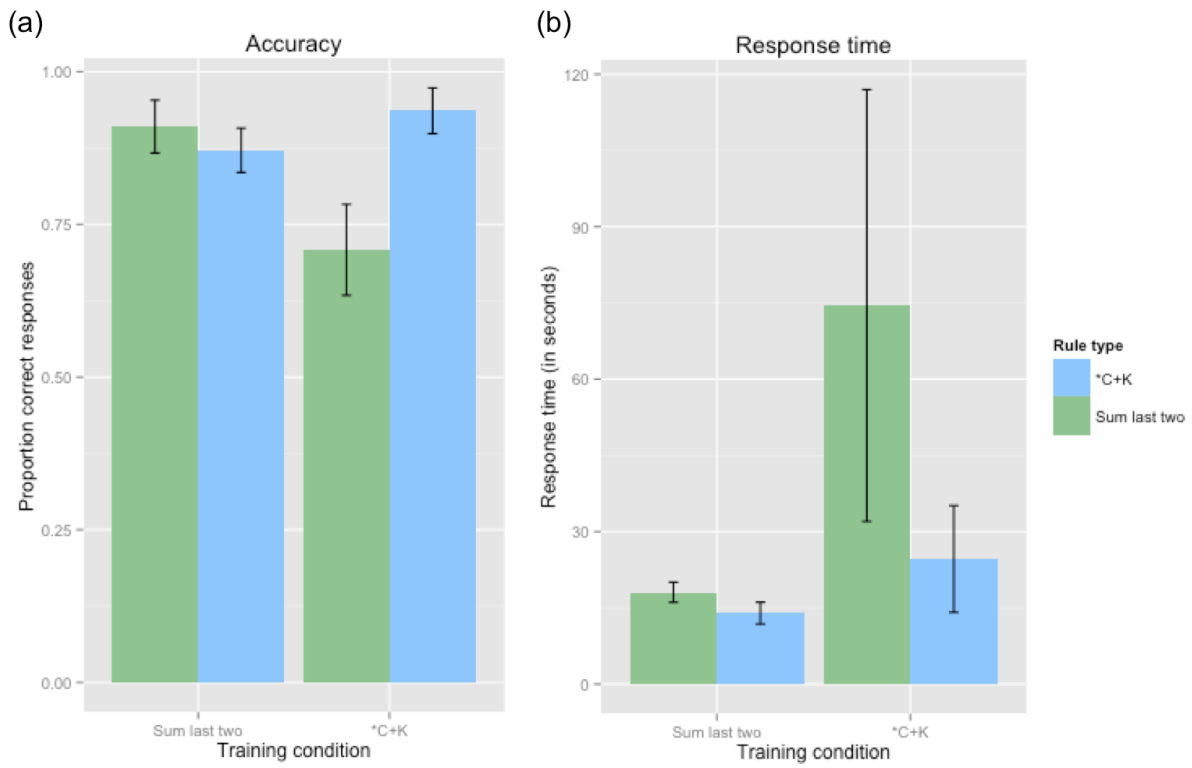

Figure 2: (a) Accuracy and (b) response time during the training phase by sequence type and training condition.

Discussion

Summary of Replication Attempt

The present study was not a straightforward replication of the original study. We failed to replicate Austerweil & Griffith's (2011) key result—that changing the prior probability of underlying rules changed how likely people were to test certain hypotheses—using a chi-squared test. It is worth noting that the results of our chi-squared test were likely inaccurate because several cells had a very small number of responses. In order to circumvent this bias, we instead analyzed the data using a mixed-effect logistic regression. Using this analysis, we indeed found that training condition predicted how often participants tested each class of rule.

Commentary

It is unclear how to reconcile the results of the logistic regression above with the critical test of our replication. As noted previously, our chi-squared test was likely biased by small samples, and it considered only the average number of responses of each type while glossing over inter-subject and inter-item variability. In this respect, the mixed logistic regression likely a more complete and robust test of whether training condition predicts which rules participants choose to test. However, this analysis was neither present in the original study nor planned for the current replication.

Because we changed some aspects of the original procedure, it is unclear whether these affected our

results. The exploratory analyses and methodology above highlight three differences between the present replication and the original study that may have impacted the results. First, our analyses of participants' accuracy and reaction time also suggest that participants were both slower and less accurate in the xK+C condition than in the sum last two condition. Second, rather than using the same training sequences in the training phase, the present replication used randomly generated sequences—while this change was intended to discourage cheating, the sequences may have been less informative, more difficult, or more redundant than the original sequences. Finally, our participants spanned a broad range of ages and education levels, and they participated in the study online. These changes were motivated by the limitations of online testing, and they do not seem to conflict with the predictions of the original authors. Nevertheless, they may have affected participants' engagement in or ability to complete the task.