# Replication of Optimal Predictions in Everyday Cognition by Griffiths & Tenenbaum

Calvin Wang
thecw@stanford.edu

## Introduction

Human perception and memory are often explained as optimal statistical inferences that are informed by accurate prior probabilities. In contrast, cognitive judgments are usually viewed as following error-prone heuristics that are insensitive to priors. Griffiths & Tenenbaum examined the optimality of human cognition in a more realistic context than typical laboratory studies, asking people to make predictions about the duration or extent of everyday phenomena such as human life spans and the box-office take of movies. Their results suggest that everyday cognitive judgments follow the same optimal statistical principles as perception and memory, and reveal a close correspondence between people's implicit probabilistic models and the statistics of the world. The goal of this project is to replicate this finding through experiments on Mechanical Turk.

## Methods

### Power Analysis

Effect size is not applicable to this study, but the replication ensured a sample size comparable to the original study, which consisted of 208 participants in the first condition and 142 participants in the second condition. The details of the conditions are described in the Procedure section below.

### Planned Sample

As mentioned in the Power Analysis section, this replication has a sample size of at least 208+142=350 participants. Since data collection happens on Mechanical Turk, the experiment automatically stopped collecting data when the goal was reached. The original experiment was done with undergraduates, which was simply the most convenient option for the authors, who were university professors. The paper is on human cognition in general and does not suggest any reason to restrict the participant population in any way except for the basic requirements that they can read and have real-world intuition. Therefore, the replication used all Mechanical Turk workers in the US.

**Materials**

Each participant made a prediction about one instance from each of the five different classes seen by his or her group. Each prediction was based on one of five possible values of t, varied randomly between subjects. These values were $1, $6, $10, $40, and $100 million for movie grosses; 2, 5, 12, 32, and 67 lines for poem lengths; 18, 39, 61, 83, and 96 years for life spans; 1, 3, 7, 11, and 23 years for reigns of pharaohs; 1, 3, 7, 11, and 23 years for lengths of marriages; 30, 60, 80, 95, and 110 min for movie run times; 1, 3, 7, 15, and 31 years for terms of U.S. representatives; 10, 20, 35, 50, and 70 min for baking times for cakes; and 1, 3, 7, 11, and 23 min for waiting times. In each case, participants read several sentences establishing context and then were asked to predict $t_{total}$ given t.

The questions were presented in survey format. Each survey began as follows:

"Each of the questions below asks you to predict something – either a duration or a quantity – based on a single piece of information. Please read each question and write your prediction on the line below it. We're interested in your intuitions, so please don't make complicated calculations - just tell us what you think!"

Each question was then introduced with a couple of sentences to provide a context. Following are sample questions:

- Insurance agencies employ actuaries to make predictions about people's lifespans – the age at which they will die – based upon demographic information. If you were assessing an insurance case for an 18 year old man, what would you predict for his lifespan?
- If you opened a book about the history of ancient Egypt to a page listing the reigns of the Pharoahs, and noticed that at 4000 BC a particular Pharoah had been ruling for 11 years, what would you predict for the total duration of his reign?
- Imagine you hear about a movie that has taken in 10 million dollars at the box office, but don't know how long it has been running. What would you predict for the total amount that of box office intake for that movie?
- If your friend read you her favourite line of poetry, and told you it was line 5 of a poem, what would you predict for the total length of the poem?
- A friend is telling you about an acquaintance whom you do not know. In passing, he happens to mention that this person has been married for 23 years. How long do you think this person's marriage will last?
- If you made a surprise visit to a friend, and found that they had been watching a movie for 30 minutes, what would you predict for the length of the movie?

- If you heard a member of the House of Representatives had served for 15 years, what would you predict his total term in the House would be?
- Imagine you are in somebody's kitchen and notice that a cake is in the oven. The timer shows that it has been baking for 35 minutes. What would you predict for the total amount of time the cake needs to bake?
- If you were calling a telephone box office to book tickets and had been on hold for 3 minutes, what would you predict for the total time you would be on hold?

### Procedure

Participants were tested in two groups, with each group making predictions about five different phenomena. One group (Condition 1) made predictions about movie grosses, poem lengths, life spans, reigns of pharaohs, and lengths of marriages. A second group (Condition 2) made predictions about movie run times, terms of U.S. representatives, baking times for cakes, waiting times, and lengths of marriages. The two groups were implemented as two different HIT groups on Mechanical Turk. The HITs were given in a survey format, each having a time limit of 5 minutes.

### Analysis Plan

Griffiths & Tenenbaum compared optimal Bayesian statistical inference based on real-world distributions that they extracted from various data sources with results from their study and found a much closer correspondence than suggested by previous research. If the Mechanical Turk data match those obtained by Griffiths & Tenenbaum, then we can conclude that the experiment replicated.

### Differences from Original Study

This is a fairly accurate replication of the original experiment, with only a few minor differences:

- The original sample was undergraduates, while our sample was Mechanical Turk workers.
- The original setting was a questionnaire in the lab, while our setting was an online survey.

Since Griffiths & Tenenbaum's context is everyday cognition of people in general, the above differences should not impact the outcome of the study.

## Methods Addendum

### Actual Sample

The sample size goal (350 participants) was reached in about one day. I approved all but one submission (due to negative inputs).

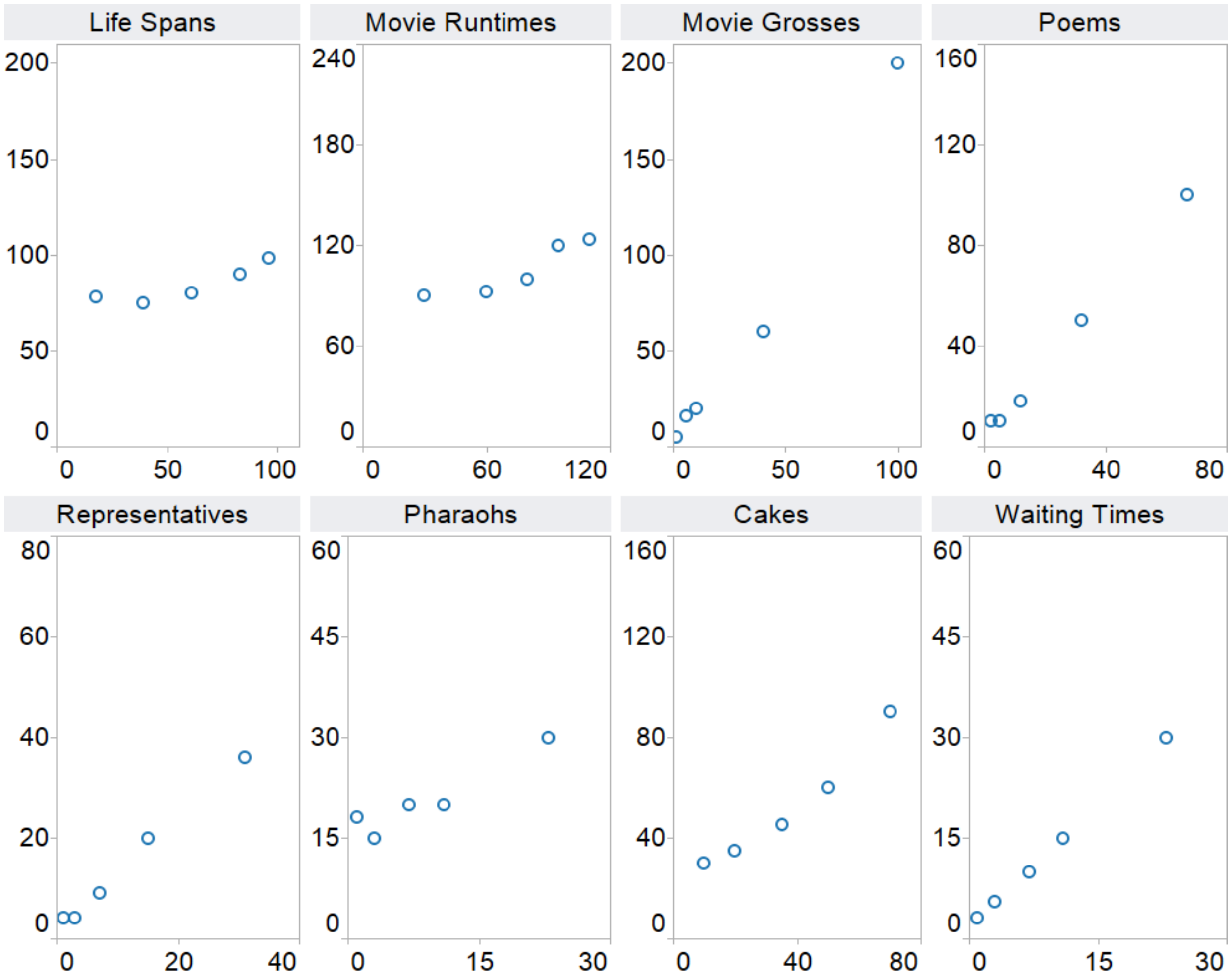
## Results

### Data preparation

Initial data cleanup was performed to ensure that all data are ready for software consumption. This included, for example, editing both “10,000,000” and “10 million” to “10”, and editing “20 years” into “20”, etc.

### Confirmatory analysis

I fed cleaned data into Tableau and made the following graph. Horizontal axes represent input, and vertical axes represent answers. Each class (life spans, poems, etc.) has 5 possible inputs (t), and the circles represent the median answers ($t_{total}$) by participants for the corresponding inputs. They are plotted on the same scales as the original paper for easy comparison.



Results from the original study are summarized in this graph from Griffiths & Tenenbaum’s paper:

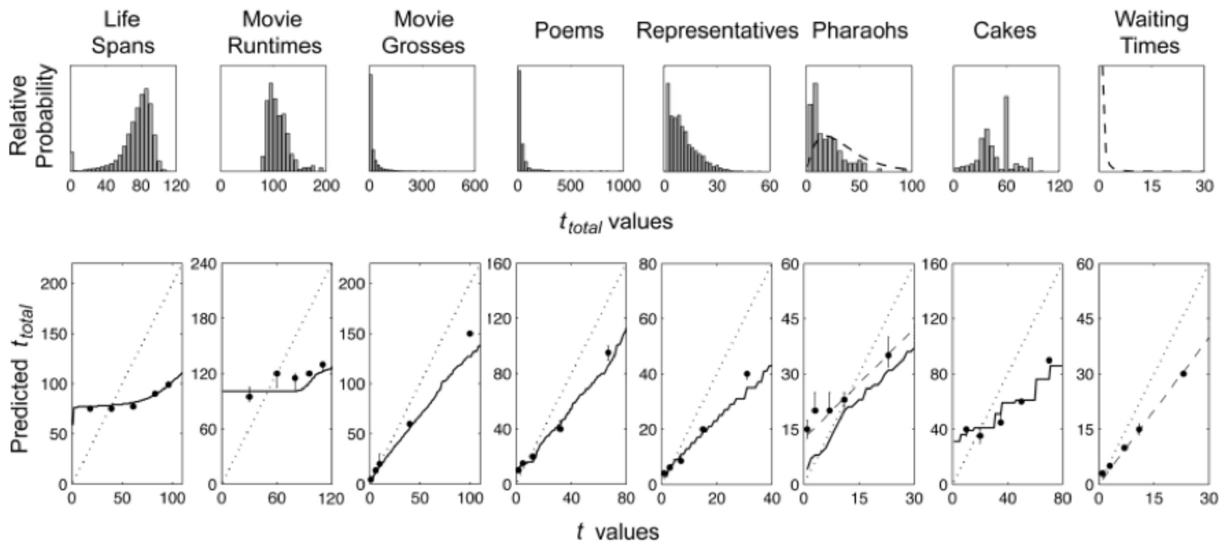


**Fig. 2.** People's predictions for various everyday phenomena. The top row of plots shows the empirical distributions of the total duration or extent, $t_{total}$, for each of these phenomena. The first two distributions are approximately Gaussian, the third and fourth are approximately power-law, and the fifth and sixth are approximately Erlang. The bottom row shows participants' predicted values of $t_{total}$ for a single observed sample $t$ of a duration or extent for each phenomenon. Black dots show the participants' median predictions of $t_{total}$. Error bars indicate 68% confidence intervals (estimated by a 1,000-sample bootstrap). Solid lines show the optimal Bayesian predictions based on the empirical prior distributions shown above. Dashed lines show predictions made by estimating a subjective prior, for the pharaohs and waiting-times stimuli, as explained in the main text. Dotted lines show predictions based on a fixed uninformative prior (Gott, 1993).

## Discussion

### Summary of Replication Attempt

Comparing the original graph with ours, we see that the data for most of the classes match almost perfectly. Those that don't match perfectly only deviate from each other by a reasonable amount. Therefore, the replication is successful.

After a closer look at the graphs, it seems that the best replication happened in the life spans class, and the worst in the pharaohs class. This is not surprising since most people have a very good picture of human life expectancy, but know almost nothing about how long pharaoh reigns typically are. Movie grosses is another interesting deviation: the first four median answers match almost exactly, but the last one is about 200 compared to about 150 in the original study. Times have changed. According to IMDB, 4 of the top 10 all-time box office hits were movies after 2006. This has had a priming effect on people and might have caused their expectations for movie grosses to rise accordingly, and it makes sense that this would only affect the fifth and last data point because box office hits always start strong.

### Commentary

The replication's success confirms Griffiths & Tenenbaum's hypothesis that everyday cognitive judgments follow implicit Bayesian statistical models. However, the degree to which this is true remains to be quantified, perhaps with a precise statistical analysis of the theoretical and actual distributions involved as a possibility of a further study.