

Analysis: Multiplayer Concept Learning Pilot #1

Sahil Chopra

March 28, 2018

Introduction

Last time, we tried replicating the results of the Piantadosi 2016 paper. While the results seemed to indicate that the participants may be learning, it was somewhat inconclusive because of errors in the experimental procedure on our behalf. Namely, there was no held out test set, and critters were sampled with replacement during training, as there were only 27 unique critters but 50 trials.

Here we correct for these errors by:

- 1) Increasing the unique critter stimuli set from 27 to 81
- 2) Creating 2 lists: 1 Train Set (50 critters) & 1 Test Set (31 critters)

Additionally, we expand the single player game into multiplayer game, as detailed below.

Experimental Setup

Stimuli Generation

For this experiment, we generate stimuli by permuting four properties, each with three different possible values. The four properties are as follows: primary color, secondary color, size, and critter species. The primary color can either be blue, green, orange. The secondary color can be either red, yellow, or purple. The size is logarithmic in the relative differences between small, medium, and large. The possible critter species are fish, bug, and bird.

We enumerate all 81 possible critters and randomly split them into two groups – train (50) and test (31).

Selected Concepts

For this pilot experiment, we only gathered data about one difficult concept:

(Primary Color = Blue) XOR (Critter Species = Fish)

Running the Experiment

We ran the experiment on MTurk with two participants per game. The participants were split into teachers and listeners, with one teacher and one listener per game.

The teacher would begin before a listener joined the game itself. The teacher would have 50 training rounds, where they had to label the given critter as “wudsy” or “not wudsy”. At first teachers had to guess, but after each round we added the previous stimulus to a table, boxing the “wudsy” ones. If the teacher incorrectly labeled a critter they had to wait 5 seconds before the next round.

Once the teacher completed their 50 rounds of training, the student met them in an online chatroom, where they forced to talk for at least 30 seconds. Here, the teacher tried to explain the definition of a “wudsy” creature to the listener. After discussing the concept sufficiently, the listener would hit “continue” – progressing both participants to the test portion of the game, where both the teacher and the listener were given the same 31 held out critters and evaluated in terms of their performance on the task.

Results

We logged performance of the teachers during training, the conversations during the chat room, and the scores of both the teacher and listener during the test phase. Here we present the results of this pilot experiment. In total, we ran 8 games.

Places of Error

There far fewer places of error in this pilot experiment. The greatest error was failure to properly log `game_ids` to Mturk. This made it a bit more of a hassle to properly stitch together participants from the same game.

Additionally, I did not set a properly long duration for the hit itself. In the previous experiment, I set the hit duration to 30 minutes and forgot to extend that here – given that this is a multiplayer game. There was an instance or two of games that stalled out for this reason – but I properly compensated these individuals for their time, with separate compensation hits.

I have fixed both of these issues for subsequent versions of this experiment.

Accuracy Across Learning

```
library(readr)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##   filter, lag

## The following objects are masked from 'package:base':
##
##   intersect, setdiff, setequal, union

library(tidyr)
library(ggplot2)
library(rjson)
```