Concept Learning: Single Player Pilot Analysis

Sahil Chopra

Wednesday, March 28, 2018

Introduction

Here, we attempt to try to replicate some of the results from Piantadosi 2016 (The Logical Primitives of Thought: Empirical Foundation for Compositional Cognitive Models). As a part of this replication, we use different stimuli - but mantain three axes of variability with three potential values for a total of 27 possible objects. To examine some examples of stimuli and play the concept learning game for yourself, please checkout the experiment page.

Experimental Setup

Selected Concepts

For this initial pilot, we attempted to replicate two concepts from Piantadosi 2016. We chose one concept that seemed relatively easy and one concept that seemed relatively difficult for participants in the original paper. We purposely avoided using any rules that required quantifiers, selecting two boolean concepts. This focus on boolean concepts, allowed us to leverage a more traditional concept learning paradigm – showing one object at a time, rather than a set of objects as in the Piantadosi paper.

Instead of using shapes, size, and color for the three axes of variability, we user different critters from the stimuli project. Specifically, we use three different types of critters: fish, bugs, and birds. Each critter could be one of three different sizes and one of three different colors.

The two selected concepts for ths initial pilot were:

1. Easy: ORANGE
2. Hard: FISH XOR BLUE

Stimuli Generation

For this experiment, we create the stimuli from randomly sampling the three properties described above with replacement. We then stored this set of randomly sampled objects in a list – so that each participant was exposed to each of the critters in the same order.

Running the Experiment

We ran the experiment on MTurk such that each participant had 50 rounds, where they had to a label the given critter as “wudsy” or “not wudsy”. At first participants had to guess, but after each round we added the previous stimulus to a table, boxing the “wudsy” ones. If the participant incorrectly labeled a critter they had to wait 5 seconds before the next round.

We ran the easy concept across 4 participants and the hard concept across 5 participants.

Results

Places of Error

If I were to repeat this pilot experiment, I would fix a bug in my game – where I did not properly log how the participants' description of the wudsy concept. Additionally, I would mantain a held out test set of critters so that we could evaluate the performance of the participants on unseen creatures. In order to facilitate the development of a held out test set, I would have to increase the size of my stimuli set. Rather than the 27 possible critters as we had in this pilot experiment, I would have 81 – adding another variable feature with three possible values. This would also enable me to construct a trainining phase of the experiment, where critters are sampled without replacement, i.e. the participant sees a creature only once.

```r
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
```

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

# Read Data
setwd("../../mturk/game-5/pilot")
easy_concept_data <- read_csv(
  file="./easy_concept/game-5-trials.csv",
  col_names=TRUE,
  col_types = "iilild"
)
hard_concept_data <- read_csv(
  file="./hard_concept/game-5-trials.csv",
  col_names=TRUE,
  col_types = "iilild"
)

# Get the accuracies for the individual participants
num_easy = 4
num_hard = 5
num_trials = 50
get_accuracy <- function(num_participants, num_trials, data) {
  acc <- vector("list", num_participants)
  i = 0
  while (i < num_participants) {
    single_worker_data <- data %>% filter(workerid == i)
    num_correct = length(which(single_worker_data$labels == single_worker_data$true_labels))
    acc[i+1] = num_correct / num_trials
    i = i + 1
  }
  acc
}
```

Accuracy First 25% Vs. Last 25%

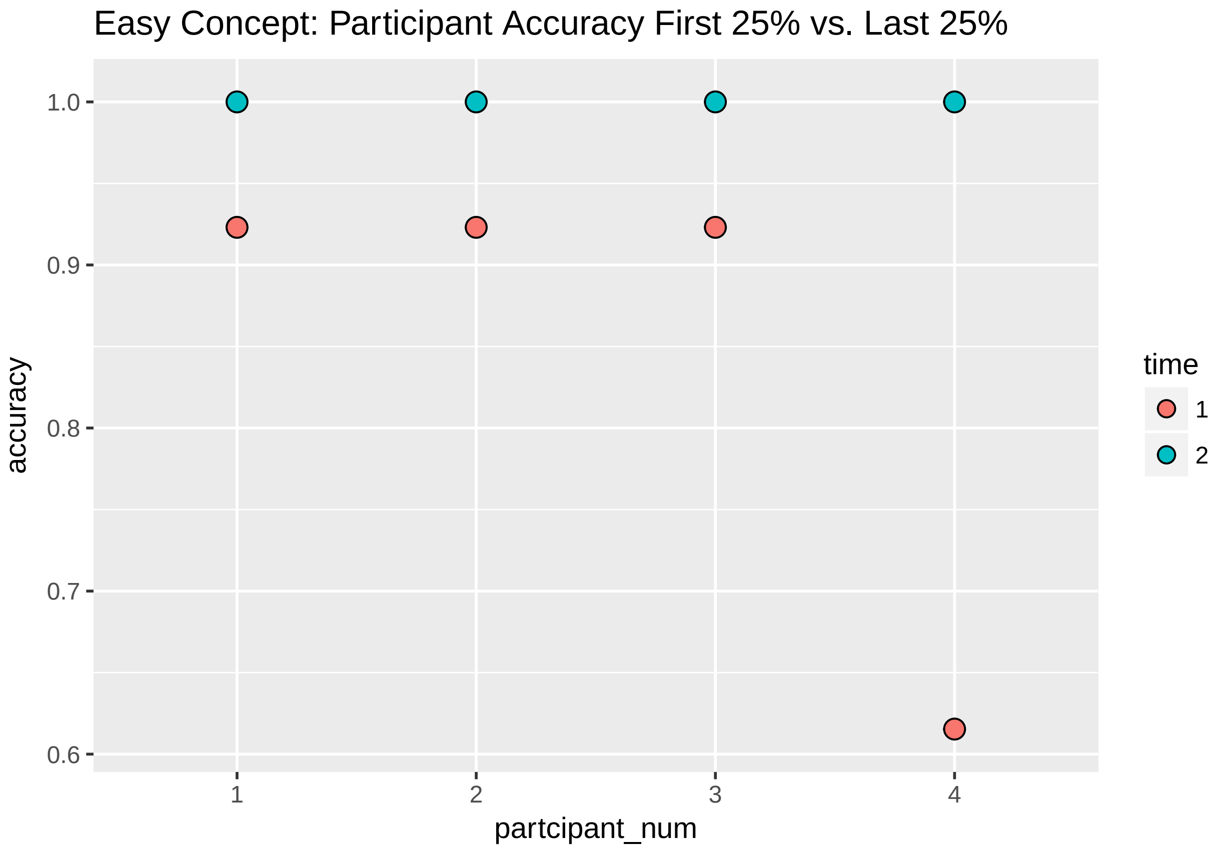
Since we didn't have a designated test set for this experiment, we examine the difference in performance between the first 25% and last 25% of the the training trials.

Easy Concept: Orange

```
# Accuracy throughout learning process (Easy Concept)
trials_25p = ceiling(num_trials * 0.25)
easy_acc_first_25p <- data.frame(
  partcipant_num=1:num_easy,
  accuracy=unlist(
    get_accuracy(
      num_easy,
      trials_25p,
      easy_concept_data %>% filter(trial_num <= num_trials * 0.25))
    ),
  concept=rep(1, num_easy),
  time=rep(1, num_easy)
)
easy_acc_final_25p <- data.frame(
  partcipant_num=1:num_easy,
  accuracy=unlist(
    get_accuracy(
      num_easy,
      trials_25p,
      easy_concept_data %>% filter(num_trials * 0.75 <= trial_num))),
  concept=rep(1, num_easy),
  time=rep(2, num_easy)
)
easy_acc_delta <- rbind(easy_acc_first_25p, easy_acc_final_25p)
easy_acc_delta$time <- as.factor(easy_acc_delta$time)
easy_acc_delta$partcipant_num <- as.factor(easy_acc_delta$partcipant_num)

# Plot
ggplot(data=easy_acc_delta, aes(x=partcipant_num, fill=time, y=accuracy)) +
  geom_dotplot(binaxis = "y", stackdir = "center") +
  labs(title = "Easy Concept: Participant Accuracy First 25% vs. Last 25%")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



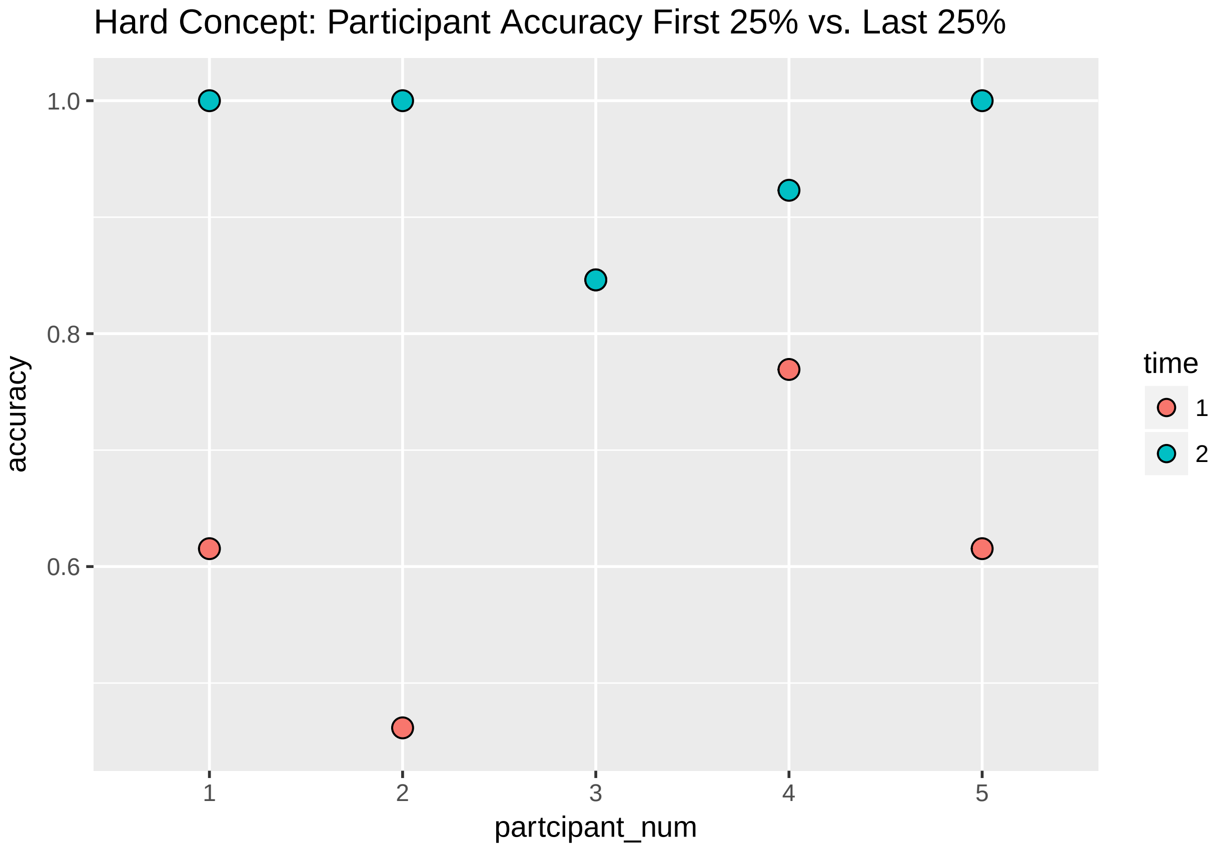
Here we see that the four participants in the “Easy Concept” group seem to have learned the concept over the course of the game, as they all started at sub 100% performance in the first 25% of trials and reached ceiling by the last 25% of trials.

Hard Concept: Blue XOR Fish

```r
# Accuracy throughout learning process (Hard Concept)
hard_acc_first_25p <- data.frame(
  partcipant_num=1:num_hard,
  accuracy=unlist(
    get_accuracy(
      num_hard,
      trials_25p,
      hard_concept_data %>% filter(trial_num <= num_trials * 0.25))
    ),
  concept=rep(1, num_hard),
  time=rep(1, num_hard))
hard_acc_final_25p <- data.frame(
  partcipant_num=1:num_hard,
  accuracy=unlist(
    get_accuracy(
      num_hard,
      trials_25p,
      hard_concept_data %>% filter(num_trials * 0.75 <= trial_num))
    ),
  concept=rep(1, num_hard),
  time=rep(2, num_hard)
)
hard_acc_delta <- rbind(hard_acc_first_25p, hard_acc_final_25p)
hard_acc_delta$time <- as.factor(hard_acc_delta$time)
hard_acc_delta$partcipant_num <- as.factor(hard_acc_delta$partcipant_num)

# Plot
ggplot(data=hard_acc_delta, aes(x=partcipant_num, fill=time, y=accuracy)) +
  geom_dotplot(binaxis = "y", stackdir = "center") +
  labs(title = "Hard Concept: Participant Accuracy First 25% vs. Last 25%")
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



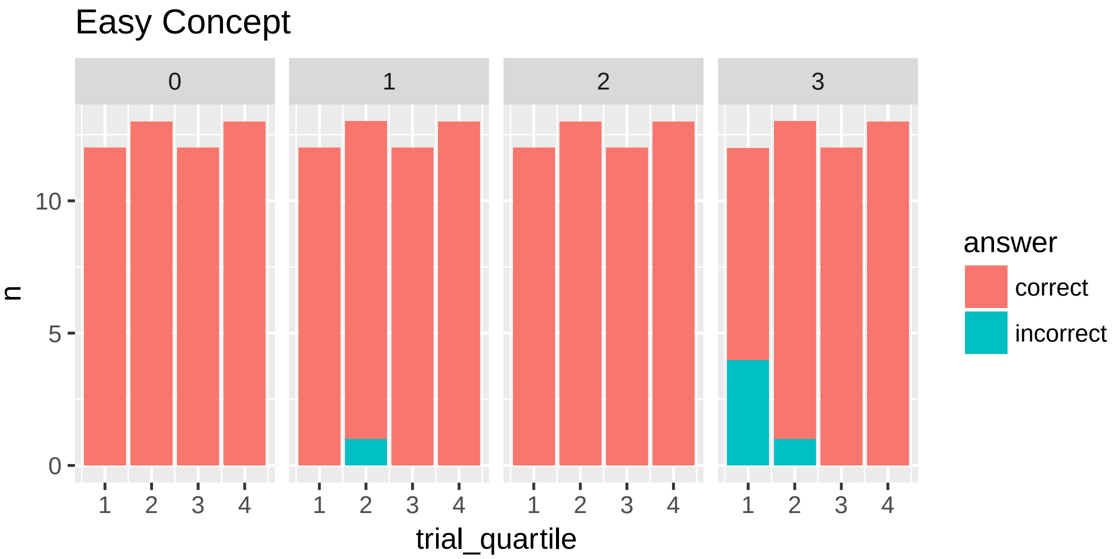
Here we see that the four participants in the "Hard Concept" started with much poorer performance during the first 25% of training trials, then those in the "Easy Concept". This is reasonable considering that the concept is much more difficult – so it may take a few trials to even be exposed to the exemplars necessary to understand the concept. 3 of the 5 participants seem to reach ceiling in the performance by the last 25% of the game, and the other 2 remaining participants seem to show marked improvement. It may seem suprising that 3 of the 5 participants reach ceiling at all – but considering that there are only 27 critters that can be possible generated during the game and that the participants have access to whether all previous creatures are "wudsy" or "not wudsy" – it may be the case that they simply examined the previous critters – when providing answers for later rounds of the game.

Response Analysis: Accuracy

```
easy_acc <- easy_concept_data %>%
  mutate(answer = ifelse(true_labels,
                          ifelse(labels, "correct", "incorrect"),
                          ifelse(labels, "incorrect", "correct")
                         )) %>%
  group_by(workerid) %>% # for each workerid, pick out their max
    mutate(max_n_trials = max(trial_num)) %>%
  ungroup() %>%
  mutate(trial_quartile = ceiling(4*(trial_num/max_n_trials) )) %>%
  group_by(workerid, trial_quartile, answer) %>%
  count()
```

Easy Concept: Orange

```
ggplot(easy_acc, aes(x = trial_quartile, fill = answer, y = n))+
  geom_col() +
  facet_wrap(~workerid, nrow = 1) +
  labs(title = "Easy Concept")
```

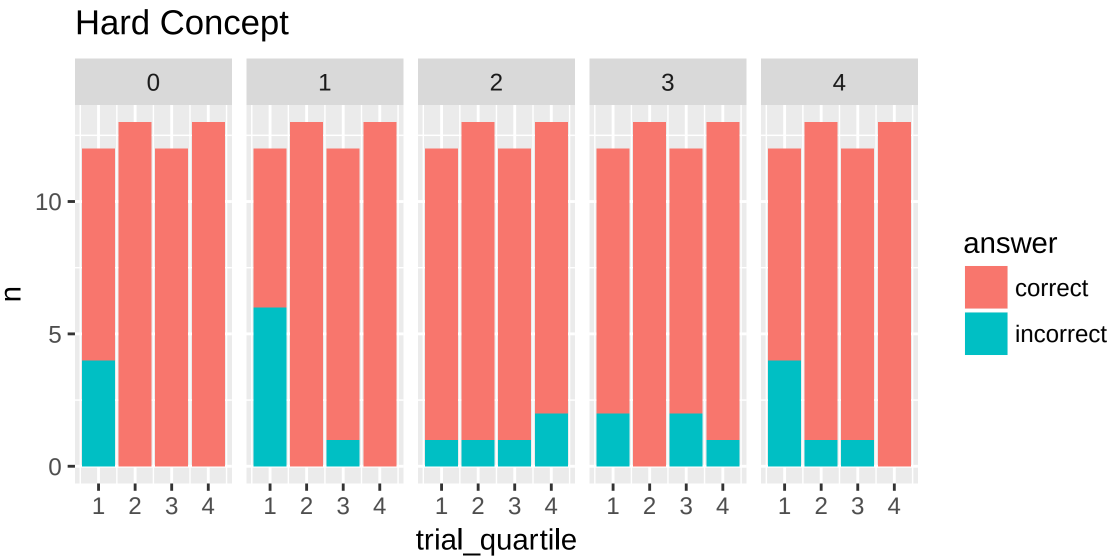


Here we examine the performance of each of the participants in the "Easy Concept" group – across the training period, split across four quarters. We see that participants 0 and 2 start at ceiling and remain at ceiling throughout the game, whereas participants 1 and 3 err a bit before hitting ceiling consistently in third and four quarters of the training period.

Hard Concept: Blue XOR Fish

```
hard_acc <- hard_concept_data %>%
  mutate(answer = ifelse(true_labels,
                          ifelse(labels, "correct", "incorrect"),
                          ifelse(labels, "incorrect", "correct")
                         )) %>%
  group_by(workerid) %>% # for each workerid, pick out their max
    mutate(max_n_trials = max(trial_num)) %>%
  ungroup() %>%
  mutate(trial_quartile = ceiling(4*(trial_num/max_n_trials) )) %>%
  group_by(workerid, trial_quartile, answer) %>%
  count()
```

```
ggplot(hard_acc, aes(x = trial_quartile, fill = answer, y = n))+
  geom_col() +
  facet_wrap(~workerid, nrow = 1) +
  labs(title = "Hard Concept")
```



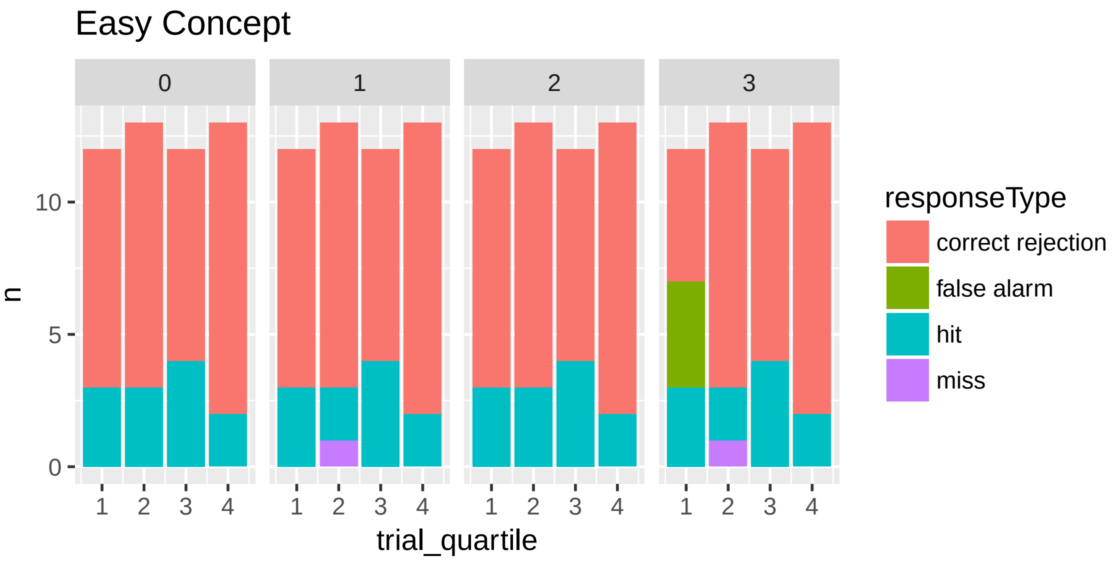
Here we examine the performance of each of the participants in the “Hard Concept” group – across the training period, split across four quarters. We see that all 4 participants start off the game with a large number of errors but they all seem to improve or hit near-ceiling by the end of the game. As mentioned earlier, this might be do the failure on our part to have a seperate test set, such that repeated examples are shown in the last 25% of the game, wherein the participant my lookup the answer to the posed “wudsy”/ “not wudsy” question if the critter was previously encountered, as this will be listed in the table of preceding examples, underneat the presented trial.

Response Analysis: Hits/Misses, False Alarms/Correct Rejections

```
easy_concept_data_quartileSummary <- easy_concept_data %>%
  mutate(responseType = ifelse(true_labels,
                               ifelse(labels, "hit", "miss"),
                               ifelse(labels, "false alarm", "correct rejection")
                               )) %>%
  group_by(workerid) %>% # for each workerid, pick out their max
    mutate(max_n_trials = max(trial_num)) %>%
  ungroup() %>%
  mutate(trial_quartile = ceiling(4*(trial_num/max_n_trials) )) %>%
  group_by(workerid, trial_quartile, responseType) %>%
  count()
```

Easy Concept: Orange

```
ggplot(easy_concept_data_quartileSummary, aes(x = trial_quartile, fill = responseType, y = n))+
  geom_col()+#position = position_dodge())+
  facet_wrap(~workerid, nrow = 1) +
  labs(title = "Easy Concept")
```

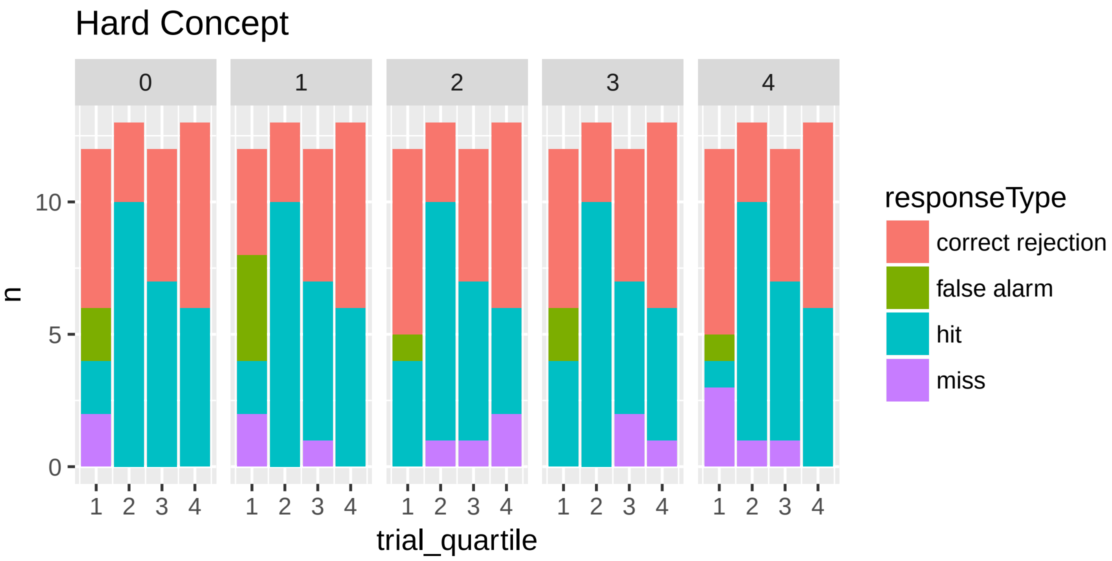


Here we see that that the members of the "Easy Concept" group, are correctly splitting their responses between correct rejections and hits, by the final 25% of the game.

```
hard_concept_data_quartileSummary <- hard_concept_data %>%
  mutate(responseType = ifelse(true_labels,
                             ifelse(labels, "hit", "miss"),
                             ifelse(labels, "false alarm", "correct rejection")
                             )) %>%
  group_by(workerid) %>% # for each workerid, pick out their max
    mutate(max_n_trials = max(trial_num)) %>%
  ungroup() %>%
  mutate(trial_quartile = ceiling(4*(trial_num/max_n_trials) )) %>%
  group_by(workerid, trial_quartile, responseType) %>%
  count()
```

Hard Concept: Blue XOR Fish

```
ggplot(hard_concept_data_quartileSummary, aes(x = trial_quartile, fill = responseType, y = n))+
  geom_col()+#position = position_dodge())+
  facet_wrap(~workerid, nrow = 1) +
  labs(title="Hard Concept")
```



Here we see that that the members of the “Hard Concept” group, are mostly correctly splitting their responses between correct rejections and hits, by the final 25% of the game. The fact that there are still some inaccuracies in the responses of participants 2 and 3 by the final 25% of the game, i.e. the misses, hopefully indicates that the particpants actually were applying some learned concept rather than simply examining exemplars from the provided table of previously seen training examples.