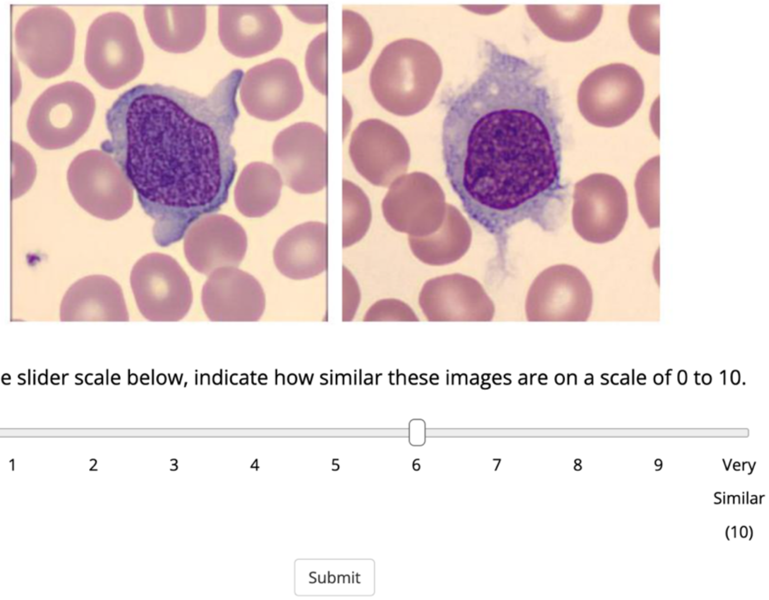
Outline

Procedure:

- Welcome/Study Information Sheet
- 40 Familiarization Trials - Image Without Showing Labels. [Counterbalanced]
- 3 Practice Trials Similarity
- **270 Similarity Judgment trials**
- Demographics/Attention Check



Using the slider scale below, indicate how similar these images are on a scale of 0 to 10.

Figure 1: Example of a similarity trial. Individuals indicate similarity using a slider.

Representations

- **General Representation:** GoogleNet trained on ImageNet only
- **Specialized Representation:** Fine-Tuned Representation (Holmes et al., 2020)

Materials

We use images from Trueblood et al. (2018) for our experiment. These are Wright-Stained White Blood Cell images. The full details of the image collection are available in the original paper. Below, we describe the procedure used to generate the trials in the experiment. The same set of trials is used for all participants.

Procedure

Familiarization Trials

In the familiarization trials, individuals are shown four images at a time without being told their labels.

We use the typicality ratings made by the experts in Trueblood et al. (2018) to create sets of 'easy' and 'hard' images. The easy images are in the top quartile of the 'typical' images (most typical), and the 'hard' images are in the bottom quartile (least typical). To create sets for the familiarization we use 40 images counterbalanced across the blast-easy, blast-hard, nonblast-easy, and nonblast-hard images. The reason for counterbalancing is that individuals could see various cell types of differing difficulty. We ensure that these sets are non-overlapping with the images used in the similarity trials.

Similarity Trials

For similarity trials, individuals are shown pairs of images and asked to report the similarity between these two images using a slider on a 0-10 scale, as shown in Figure 1. To avoid anchoring effects, the trial starts with no default slider position.

There are three different types of similarity trials - random, same, and expected similarity.

Random: For the random trials, both images are randomly chosen. [30 trials]

Same: The left image is randomly chosen. The right image is rotated by a randomly chosen amount - 90,180 or 270 degrees. [30 trials]

Expected: We first choose a representation R for which we are to generate a trial. The goal is to generate a pair of images that are expected to be similar (or dissimilar) according to R. For the expected similar trials, we randomly chose an image and one of its closest neighbors (top 5) based on the representation R to create a pair that is expected to be similar based on R (expected-similar-R trial). For the expected dissimilar trials, we randomly chose an image and calculated the similarity to all other images. We randomly chose an image in the bottom decile when images are ordered by distance to create a pair that is expected to be dissimilar based on R (expected-dissimilar-R trial). [2 (Similar, Dissimilar) x 2 (General, Specialized) x 30 trials=120 trials]

Disagreement: We choose two representations R and S based on which we generate a trial. The goal is to find a pair of images (s1,s2) such that the pair is similar according to R but dissimilar according to S. Specifically, we look for pairs such that s2 is in the top 5 neighbors of s1 according to representation R and s2 is in the bottom quartile when images are ordered based on distance based on S (disagreement-s1-nots2). [2 (disagreement-general-not-specialized, disagreement-specialized-not-general) x30 trials=60 trials]

Repeat: From the generated image pairs, we randomly chose trials to estimate the intra-rater reliability. [30 trials]

Practice Trials: Pairs are randomly selected to generate practice trials. [3 trials]

Sample Size:

100 (total) participants.

Planned Analysis

Exclusions: We will remove individuals who fail the attention checks before the analyses. We will also run the analyses with and without individuals with a Pearson intra-rater correlation of 0.3.

We will fit mixed-effects regression models with stimuli and participants as random effects to predict similarity judgments based on Trial Type.

Trial Type

- random
- same
- expected-similar-general
- expected-similar-specialized
- expected-dissimilar-general
- expected-dissimilar-specialized
- disagreement-general-not-specialized
- disagreement-specialized-not-general

Question 1:

(i) Can humans identify self-similarity for novel objects with which they have probably had no prior experience?

Prediction:

Similarity judgments are larger for the same trials than for random trials.

Analysis:

We predict we will see a significant trial-type variable, indicating a higher similarity rating for the same trials.

Question 2:

Do general and specialized representations predict human similarity judgments?

Prediction:

First, similarity judgments will be higher for the expected similar trials than for random trials. Second, similarity judgments will be lower for the dissimilar trials than for random trials. This will be true for both representations – specialized and general.

Analysis:

We predict a significant positive coefficient for trial type between the expected-similar-R for both representations. We predict a significant negative coefficient for trial type between the expected-dissimilar-R for both representations.

Question 3:

Which representation does a better job of predicting similarity judgments?

Prediction(s):

The general representation might better predict similarity trials since its experience is similar to that of individuals not exposed to white blood cells. However, the specialized representation might perform better at predicting similarity since it picks up on task relevant features. We compare the coefficients on the expected similarity / dissimilarity trials and disagreement trials to see which representation can better predict human similarity.

Analysis

We will conduct this analysis in three parts (i) We will compare the expected similarity coefficients for the general and specialized representation. A higher coefficient indicates that the representation can better predict human similarity (ii) We will compare the expected dissimilarity coefficients for the general and specialized representation. A smaller coefficient will indicate that the representation can better predict the dissimilar trials (iii) We will compare the coefficients on the disagreement trials. If the coefficient for the disagreement-specialized-not-general is larger than the coefficient for disagreement-general-not-specialized then it indicates that the specialized representation can better predict similarity than the general representation.