A Model for Categorization: STEM Stereotypes

Huiwen (Dorothy) Duan, Tianying (Teanna) Feng, Tiffany Kanamaru, Juliette Le Saint

Background

Categorization allows individuals to identify and understand objects by attributing them to and placing them in a group (a memory representation) whose existing members embody the object's properties better than other groups. Two prominent theories have emerged to explain the processes that might underlie category judgments. The **prototype theory** states that categories are represented in memory by one average and abstract, "most typical" member of that category, called a prototype. A category is chosen for a new stimulus such that the stimulus is most similar to the prototype for a category in question, minimizing the conceptual distance from the stimulus to the prototype (Rosch, 1978). On the other hand, the **exemplar theory** states that to make a category judgment, a new stimulus is compared to all remembered prior instances of category members, called exemplars, with the most "typical" exemplars being remembered first. The general consensus today is that both theories trade off under different circumstances, on the basis of experimental results showing that different processes underlie categorization at different stages of learning, concerning categories of different sizes and structures (Smith, 2014).

Smith (2014) demonstrates a current understanding of the cases under which either prototype or exemplar processes dominate category judgment. His results show a prototype process categorizes transfer items most efficiently in cases involving family resemblance, where category members have "variable and probabilistic similarity relationships" such that members will share some but not all features. Similarly, a prototype process is more efficient in cases of prototype-exception categorization, (categories include logically disjoint subclusters where category members are not linearly separable) but only when items are typical of the category. Ultimately, this shows the exemplar process is advantaged when dealing with exception items, though the prototype process is best overall for categories with fewer exception members. Smith clarifies that still, an interaction between both processes is likely in any category judgment.

Noteable reference. Our team consulted Medin and Schaffer's (1978) second experiment for classification learning in order to guide our project. Medin and Schaffer's models considered conceptual distances and weights and were not probabilistic, unlike the model we implemented in this experiment, which was probabilistic. Furthermore, Medin and Schaffer presented novel stimuli to participants, including geometric shapes and schematic faces. In particular, the stimuli had four feature dimensions; the geometric shape stimuli were distinguished by color, form, position and size. His denotation of each feature by a '0' or '1' allowed him to interpret each stimuli using the equations delegated to each model, with an additive (averaging) equation corresponding to his prototype model and a multiplicative equation capturing the interaction between individual features in his exemplar model. To clarify, Medin and Schaffer referred to their models models as the "independent cues" and "context" models — these are analogous to the prototype and exemplar models and so were consulted in the current experiment.

Our study attempted to better understand category judgment by comparing human judgment (participants' categorical judgments of imaginary students as either STEM (i.e., science, technology, engineering, and math) or non-STEM majors) to the results of prototype and exemplar models for categorization. Stereotypes such as those presented in the features we assigned STEM and non-STEM students are used in social categorization to place individuals into certain social groups (Jhangiani & Tarry, 2019). When participants are asked to determine whether an imagined student is a STEM or non-STEM student, do they imagine two abstract prototypes of a STEM and non-STEM students and compare the stimulus to those representations to make their decision, or do they compare the single stimulus to numerous previously encountered instances of STEM and non-STEM students?

Research Questions: *Our project attempts to answer the following questions:*

1. In contrast to human category judgments regarding novel stimuli, which theory better anticipates humans category judgments concerning familiar stimuli for which they have an priori judgment or understanding (a stereotype): the prototype or exemplar theory?
2. Which of the following features of stereotypical STEM versus non-STEM majors do participants weigh more heavily when making category judgments: gender, appearance, personality or favorite TV show?

We hypothesized that the prototype theory would better capture how people judge whether a person is a STEM major, as the distinction between STEM and non-STEM majors revolves around family resemblance, which states a pair of members will share some but not all features. Intuitively, it would also make sense that a prototypical, abstract STEM student embodying all of the stereotypical features of a STEM student might exist in participants' minds and is consulted when performing social categorization, as stereotyping involves generalization. Recalling all previously-met STEM and non-STEM students, who may each embody only some features of a stimulus, seems far more unlikely. This notion is complicated, however, by participants' potential effort not to make discriminatory judgments based on stereotypes. We hypothesized gender and personality would be weighed more heavily than favorite TV shows and appearance, as these features are more salient and variable across individuals, respectively.

Method

Selection of Features

Existing literature was consulted to determine four stereotypes directed at individuals belonging to STEM or non-STEM fields: gender, personality, favorite TV show, and appearance (Banchefsky et al., 2016; Cheryan et al., 2011; Cundiff et al., 2013; Nassar-McIllian et al., 2011).

Gender. Cundiff et al. (2013) tested gender stereotypes on 1,700 undergraduate students and found that women were deemed more likely to be non-STEM majors while men were deemed more likely to be STEM majors. (0 = female, non-STEM ; 1 = male, STEM)

Personality. STEM majors are more strongly associated with introverted behavior and non-STEM students with sociable behavior. Nassar-McIllian et al. (2011) documented these stereotypes, listing several for individuals involved in STEM (i.e., being uninteresting, unsociable, and dull). Thus, those who are not involved in STEM are stereotypically interesting, social, and fun. (0 = sociable, non-STEM ; 1 = unsociable, STEM)

TV show. Cheryan et al. (2011)'s studied the effect of male and female role models on women's anticipated success in STEM. The study identified stereotypical and non stereotypical items indicated by participants as being either stereotypical or non stereotypical of STEM majors. Results found that the television show "Mystery Science Theater 3000" was significantly more stereotypical of a computer science (STEM) major than "The Office." "Mystery Science Theater 3000" is identified as a fantasy show and "The Office" as a sitcom show, meaning STEM majors are stereotyped as being more likely to watch fantasy shows and non-STEM students are more likely to watch sitcoms. (0 = sitcom, non-STEM ; 1 = fantasy, STEM)

Appearance. Banchefksy's study on the effect of appearance on the perception of an individual being a scientist was used to define the last feature (Banchefsky et al., 2016). Participants rated photos of males and females based on their femininity, the likelihood of being a scientist, and the likelihood of being an early childhood educator. The more feminine females were rated, the less likely they were rated to be a scientist and the more likely they were rated to be an early childhood educator (non-STEM). In our study, femininity was considered linked to individuals' concern with their appearance**,** based on previous literature indicating feminity is stereotypically linked to such behavior (Jackson, 1992). Students who quickly comb their hair were associated with STEM and those who carefully styled their hair were associated with non-STEM. (0 = concerned w/ appearance, non-STEM ; 1 = not concerned w/ appearance, STEM)

Materials

We created 16 stimuli belonging to the categories 'STEM' (8) or 'non-STEM' (8) modelled after the stimuli in Medin and Schaffer's (1978) second experiment. Each stimulus, a short paragraph describing part of an imaginary student's day, contained four binary feature dimensions, where '1' indicated a STEM stereotype and '0' indicated a non-STEM stereotype (Table 1). Each stimulus was identified by a unique combination of these four features. A feature set of '0000' indicated a prototypical non-STEM student (with all non-STEM stereotypical features). This is analogous for a feature set of '1111' in relation to a prototypical STEM student. Furthermore, the features were highlighted for participants as underlined phrases of interest in the stories such that they would be noticed. Notably, the gender feature was the only inconsistent feature across stimuli as a control for redundancy. It was represented by typical male and female names and was implicated by the gender pronouns used in the story. Each story presented a different name; In total, there were 8 female names and 8 male names.

These 16 stimuli were separated into training and testing groups as per Medin's method, with nine stimuli pertaining to the training phase and seven to the testing phase. The training stimuli were further labeled and separated into STEM and non-STEM categories where the majority of dimensions in the STEM category had more 1's (STEM), and the majority of dimensions in the non-STEM category had more 0's (non-STEM) (Table 2 and Table 3).

Three different story styles were written for each stimulus to control for redundancy. The first story centered around an individual getting ready for a date, the second story centered around an individual getting ready for bed, and the third story centered around an individual getting ready for work. An example of each type of story is presented in Appendix A. Therefore, there were 48 versions in total, three for each story.

24 undergraduate students were recruited for this experiment. Each participant was presented with a 24-slide presentation containing 16 stories (stimuli) such that categorical judgment data was collected for every possible combination of features. The order in which the stimuli were presented was randomized in both the training and testing groups. Each story style was also randomly chosen from the story pool for that stimulus.

Model Implementation

Context and independent cues models. The context and independent cues models implemented by Medin and Schaffer's (1978) corresponded to exemplar and prototype models, respectively. Both models computed the hamming distance between a stimulus and the training examples or the prototype. For example, a hamming distance of 1 indicates a stimulus differs from the prototype stimulus by one feature feature dimension (e.g., stimulus: 1110 and STEM prototype: 1111). Given this difference measure, Medin and Schaffer weighted each feature dimension that differed. The weights were not probabilistic and had to be defined. For their exemplar model, they used weights of .16, .16, .18, and .14. For their prototype model, they used weights of .38, .10, .40, and .20.

Probabilistic prototype and exemplar models. Our model implementations were similar to Medin and Schaffer's (1978) models, but instead of using the weighted hamming distance as a difference measure, we assigned a probability of supporting the categorization of STEM to each feature dimension. The formula for assigning such probabilities is shown below. 1 - ε indicates the probability of a stimulus having the same k^{th} feature as the k^{th} feature of a training example or the prototype. The probability ε indicates the probability of a stimulus having its k^{th} feature differing from the k^{th} feature of a training example or the prototype.

$$P(x_k \mid \text{STEM}) = 1 - \varepsilon \text{ if } x_k = \mu_{\text{STEM},k}$$

$$P(x_k \mid \text{STEM}) = \varepsilon \text{ if } x_k \neq \mu_{\text{STEM},k}$$

Using this formula, we maximized the difference between 1 - ε and ε, assigned a probability to each of the four feature dimensions, and computed the product of the four probabilities. We calculated the ratio between the probabilities of a stimulus falling into the STEM category versus the non-STEM category as the outcome measure. Then, we rank-ordered these predicted outcomes for the exemplar and prototype models, respectively.

Procedure

Nine training stimuli were presented. Participants were instructed to read each story and indicate on the answer sheet whether the individual in the story was a STEM or non-STEM major. After marking their answer, participants were told in a following slide whether the individual in the story was a STEM or non-STEM major. Then, participants were given a distraction task showing ten consonant-vowel- consonants (e.g., cat) and prompting participants to rate how meaningful each word was using a 7-point Likert scale. After the distraction task, the seven testing stimuli were presented. Participants read each story and were asked to indicate whether the individual in the story was a STEM or non-STEM major. The correct answers were not given, unlike the training phase. The participant answer sheet can be seen in Appendix B.

Results

Human responses were averaged for each stimulus to compute the observed categorization probabilities, thus generating 16 observed categorization probabilities for each stimulus. Because our human responses, Medin and Schaffer's (1978) human responses, and model predictions each used different scales, we decided to use the rank orders.

Human Responses

Our human responses were not significantly correlated with Medin and Schaffer's human ratings for the nine training stimuli, $r(24) = 0.17$, $p > .05$, but were significantly correlated with Medin and Schaffer's human ratings for the seven testing stimuli, $r(24) = 0.89$, $p < .05$.

Fits of the Models

According to Medin and Schaffer (1978), "the main prediction of interest concerns Stimuli 4 and 7. Since the modal prototype for STEM is 1111, stimulus 4 (1110) must be at least as close as 7 (1010) is to the prototype, no matter how the dimensions are weighted" (p. 221). Our study also used 1111 as our STEM prototype, and because stimulus 4 shares more features with the prototype than stimulus 7, the prototype model should give stimulus 4 as high a probability of belonging to the STEM category as the exemplar model would predict. Thus, our first step to validating both models was to look at their predictions for stimuli 4 and 7. Our prototype model favored stimulus 4, as it gave a higher probability than stimulus 7. Contrastingly, our exemplar model assigns to stimulus 7 a higher probability of indicating a STEM major. This pattern is consistent with the feature combinations of stimuli 4 and 7.

We also compared both models against observed categorization probabilities over the 16 stimuli. The correlation between predicted and observed rank orders was significant, $r(24) = 0.53$, $p < .05$ for the prototype model and insignificant for the exemplar model, $r(24) = 0.17$, $p > .05$. Therefore, the prototype model gave a better fit. Figure 1 shows a bar graph comparing the rank order of human responses vs. exemplar model predictions, and Figure 2 compares the rank order of human responses vs. prototype model predictions. Compared to the exemplar model, the prototype model's predictions were closer to our participant's responses on all stimuli except for stimulus 6. Therefore, we can conclude that in our experimental settings, the prototype model did a better job in predicting human categorization about STEM versus non-STEM major students.

Weighted Features

The first set of epsilon ($1 - \varepsilon$) values were set [0.90, 0.80, 0.70, 0.92] based on our hypothesis that gender and personality would serve as the most influential features in categorization. To test our model's robustness to different epsilon values for different feature weights, we changed the values to [0.60, 0.60, 0.60, 0.60] and the prototype model still performed better than the exemplar model. Next, to test which feature was weighted most heavily by the participants, we assigned a weight of 0.90 to each of the four feature dimensions while keeping the other three set to 0.55. The specific weights used and the correlation coefficients of our participant's responses versus model predictions are shown in Table 4. Results indicated that participants heavily used personality to categorize whether an individual was a STEM major or not , $r(24) = 0.53$, $p < .05$. Other features were non significant, and thus, did not have a role in categorization.

Discussion

Our observed data showed participants' categorization method was more consistent with the prototype model than the exemplar model, thus confirming our first hypothesis. It should be noted, however, that it is generally accepted that both exemplar and prototype processes are implicated when people categorize social events (Smith and Zarate, 1990). Here, because the exemplar process was found to be consistently insignificant, our results do not fall completely in line with this previous finding, which indicates both processes should be implicated. Our results indicate that when people are deciding whether an individual is a STEM or non-STEM student, they use an abstract prototype of a STEM and non-STEM student to make their decision.

We found that participants weighed personality the most when making category judgements based on the effect of varying feature weights. This partially follows our hypothesis since gender was not found to have a significant influence on judgment. This could imply that participants, aiming to avoid discriminating between individuals using gender, chose to rely on a more neutral and less politically-charged feature, personality, to judge social group membership.

In future, we will further study the weight (influence) of each feature, select more precise epsilon values, and do more to ensure no confounding interaction between the features, as we suspect gender might have interacted with appearance. It is also worth exploring how other stereotypes influence human reliance on different processes while making category judgments.

References

Banchefsky, S., Westfall, J., Park, B., & Judd. C. M. (2016). But you don't look like a scientist!: Women scientists with feminine appearance are deemed less likely to be scientists. *Sex Roles, 75,* 95-109: doi: 10.1007/s11199-016-0586-1

Cheryan, S., Siy, J. O., Vichayapai, M., Drury, B. J., & Kim, S. (2011). Do female and male role models who embody STEM stereotypes hinder women's anticipated success in STEM? *Social Psychological and Personality Science, 2*, 656-664. doi: 10.1177/1948550611405218

Cundiff, J. L., Vescio, T. K., Loken, E., & Lo, L. (2013). Do gender-science stereotypes predict science identification and science career aspirations among undergraduate science majors? *Social Psychology of Education, 16,* 541-554. doi: 10.1007/s11218-013-9232-8

Jackson, L. A. (1992). Physical appearance and gender: Sociobiological and sociocultural perspectives. Albany: State University of New York Press.

Jhangiani, R. & Tarry, H. (2019). Social categorization and stereotyping [Website]. Retrieved from https://opentextbc.ca/socialpsychology/chapter/social-categorization-and-stereotyping

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238. doi: 10.1037/0033-295X.85.3.207

Nassar-McMillan, S. C., Wyer, M., Oliver-Hoyo, M., & Schneider, J. (2011). New tools for examining undergraduate students' STEM stereotypes: Implications for women and other underrepresented groups. *New Directions for Institutional Research, 2011*(152)*,* 87-98. doi: 10.1002/ir.411

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum

Smith, E. R., & Zdrate, M. A. (1990). Exemplar and prototype use in social categorization. *Social Cognition, 8*, 243-262.

Smith, J. D. (2014). Prototypes, exemplars, and the natural history of categorization. *Psychonomic Bulletin & Review, 21*, 312-331.

Stangor, C., & Crandall, C. (2013). Stereotyping and prejudice. New York: Psychology Press.

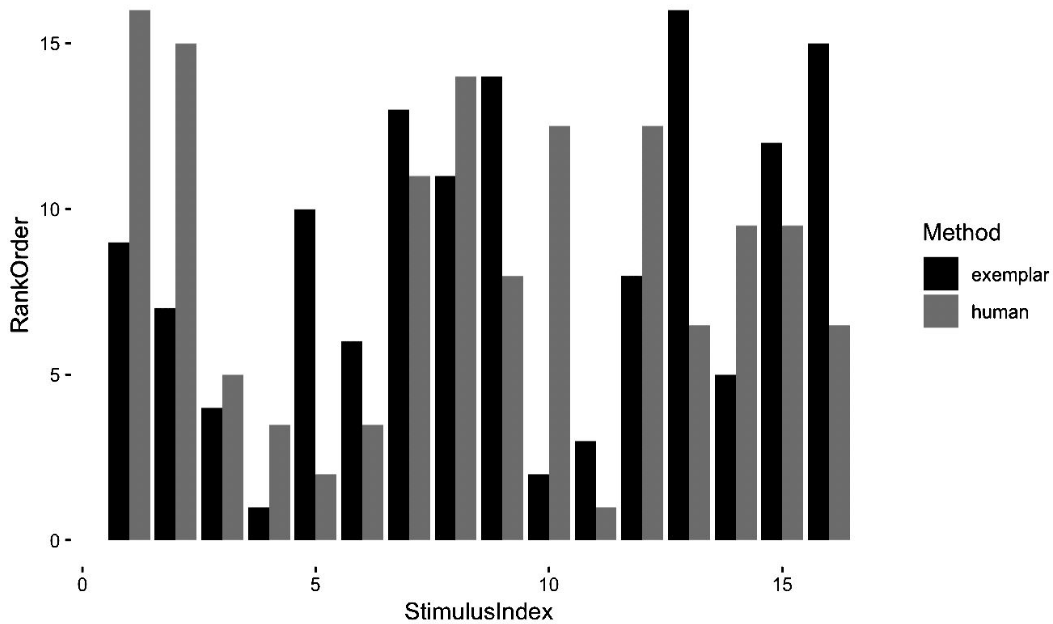

Figure 1. Comparison of rank orders between human ratings and predictions by the exemplar model

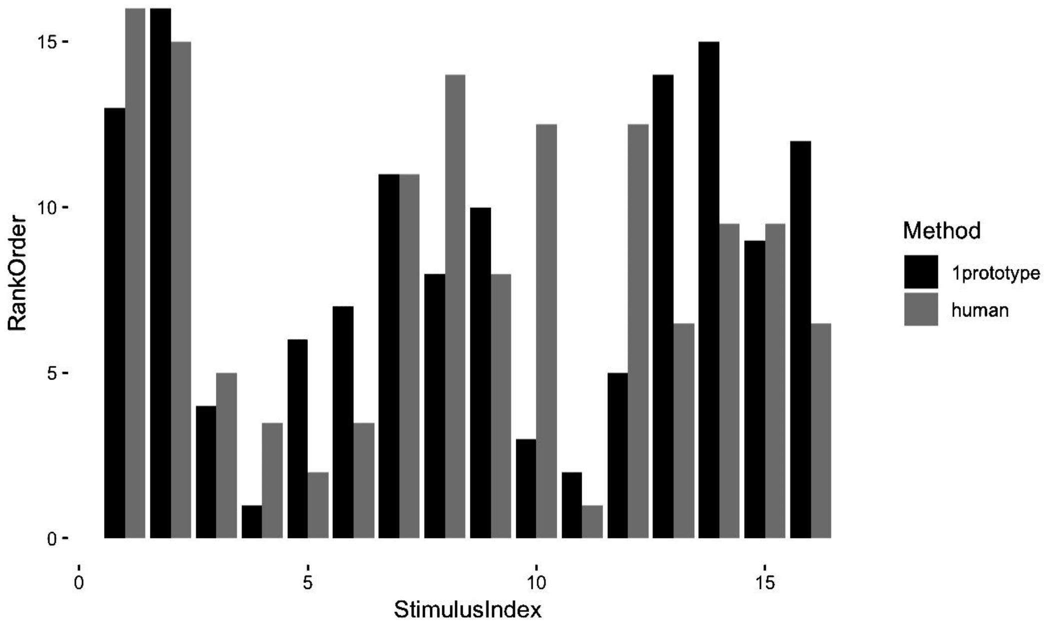

Figure 2. Comparison of rank orders between human ratings and predictions by the prototype model

Table 1

Feature Selection

	1	0
Gender	Male	Female
Hair	Quickly combs hair	Carefully styles hair
Favorite show	Fantasy	Sitcom
Personality	Not social	Social

Table 2

Training stimuli based on Medin and Schaffer's (1978) *study*

STEM category	Gender	Hair	Favorite Show	Personality
STEM major	1	1	1	0
	1	0	1	0
	1	0	1	1
	1	1	0	1
	0	1	1	1
Non-STEM major	1	1	0	0
	0	1	1	0
	0	0	0	1
	0	0	0	0

Table 3

Testing stimuli based on Medin and Schaffer's (1978) *study*

STEM category	Gender	Hair	Favorite Show	Personality
STEM	1	0	0	1
Non-STEM	1	0	0	0
STEM	1	1	1	1
Non-STEM	0	0	1	0
STEM	0	1	0	1
STEM	0	0	1	1
Non-STEM	0	1	0	0

Table 4

Correlations between human ratings and model predictions with different epsilon values

Feature	Gender		Hobby (Favorite Show)		Appearance (Hair Styling)		Personality (Socialness)	
Weights (1-ε)	[0.90 0.55 0.55 0.55]		[0.55 0.90 0.55 0.55]		[0.55 0.50 0.90 0.55]		[0.55 0.50 0.55 0.90]	
Model	Exemplar	Prototype	Exemplar	Prototype	Exemplar	Prototype	Exemplar	Prototype
Pearson's Correlation	-0.0297	-0.0700	0.3105	0.2246	-0.0087	-0.0368	0.1871	0.5316

Appendix A

A Set of Stories for the Stimulus 1 1 1 0

Ryan is going on a date tonight. He turns off his favorite fantasy show and quickly combs his hair. He loves meeting and hanging out with new people so he is excited about this date.

Ryan gets out of bed and realizes it's almost time to go to class. He heads over to his bathroom and quickly combs his hair. Walking back to his room, he makes plans to watch his favorite fantasy show with his friends when he gets home from school.

It's 9am, and so it's time for Ryan to head to work. He turns off his favorite fantasy show and goes over to his bathroom to quickly brush his hair. His phone suddenly buzzes. His friend is giving him a call, and is inviting him to grab lunch with their friend group on Sunday. Ryan happily agrees, because he enjoys hanging out in large groups.

Appendix B

Participants' Answer Sheet

(each part was presented on a separate piece of paper in the actual study)

PART 1

For each question, please type in 0 if you think the person is not a stem major, or type in 1 if you think that person is a stem major.

1	2	3	4	5	6	7	8	9

PART 2

Please rate the meaningfulness of each consonant-vowel-consonant item you just saw using the following scale. Write down your rating for each item.

How meaningful is this consonant-vowel-consonant?

1	2	3	4	5	6	7
Extremely not meaningful	Very not meaningful	Somewhat not meaningful	Neutral	Somewhat meaning	Very meaningful	Extremely meaningful

1	2	3	4	5	6	7	8	9	10

PART 3

For each question, please type in 0 if you think the person is not a stem major, or type in 1 if you think that person is a stem major.

1	2	3	4	5	6	7