Comparing the Threshold and Prototype Model for Gradable Adjectives

Tamar Johnson (t.johnson@uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam

Alexandra Sarafoglou (A.S.G.Sarafoglou@uva.nl)
Psychological Methods, University of Amsterdam

Julia Haaf (julia.haaf@uni-potsdam.de)
Psychological Methods, Statistics and Evaluation, University of Potsdam

Ingmar Visser (I.Visser@uva.nl)
Developmental Psychology, University of Amsterdam

Jakub Szymanik (jakub.szymanik@unitn.it)
Center for Brain/Mind Sciences, University of Trento

Abstract

In logical theories of meaning, threshold and prototype models are two distinctive formal approaches. In cognitive science literature, however, where the two models are operationalized, there is support for the use of a threshold model in categorization (Schmidt, Goodman, Barner, & Tenenbaum, 2009; Ramotowska, Haaf, Van Maanen, & Szymanik, 2022) as well as support for the prototype model (Douven, 2016; Douven, Wenmackers, Jraissati, & Decock, 2017), and in many cases the two models are used interchangeably (Kruschke, 2008). We test for the case of relative gradable adjectives whether a) there is a difference between predicted degrees of membership from the two models when relying on explicit reports of threshold and prototype values, and b) which of the models better predicts behavioral data from categorization tasks. Results suggest that prototype and threshold models are highly predictive of behaviour in a categorization task and that the two models yield similar results with a slight advantage of the threshold model.

Keywords: gradable adjectives; threshold; prototypes; categorization; meaning representation

Introduction

Categorization is central in cognitive sciences. Here, we focus on relative gradable adjectives and explore their representation and the underlying cognitive process by which objects in the world are divided into relative gradable categories like *big* and *tall*. Relative gradable adjectives are contrasted with absolute adjectives that describe a maximum or a minimum value (such as *full*) and are characterized by two critical aspects. The first is their context-sensitive interpretation (Kennedy, 2007). The meaning of these adjectives is inherently linked to the noun they modify (comparison class). A temperature of 20 degrees Celsius can count as *warm* if it describes the weather in London, but as *cold* if it is the temperature of the water in a bathtub. Second, the concept of vagueness, or graded membership, is integral to the use of gradable adjectives. There are borderline cases, even after contextual factors are accounted for. Therefore, the underlying model for the categorization process with gradable adjectives should consider these aspects.

Two categorization models that are often linked in the literature to gradable adjectives are the prototype model and the threshold model. The prototype model, particularly in the framework of Conceptual Spaces (Gärdenfors & Williams, 2001), offers a foundation for modeling categorization processes with graded membership (Decock & Douven, 2014). According to this model, the categorization is made by comparing the distance of an item to the prototypical values of each category. The item is assigned to the category of the closest prototype. There is empirical support for this approach in various linguistic domains, including color adjectives (Douven et al., 2017), shape categories (Douven, 2016), and relative gradable adjectives (Verheyen & Égré, 2018).

The threshold model presents an alternative perspective. It posits that for a predicate like *tall* to be applicable, an object must surpass a certain threshold on a relevant dimension, such as height. It is used extensively in accounting for pragmatic aspects of categorization in gradable adjectives (e.g., Lassiter & Goodman, 2013; Qing & Franke, 2014; Pezzelle & Fernández, 2023), but not exclusively (for interpretation of quantifiers see e.g., Ramotowska et al., 2022).

Although the two models stem from theoretically different approaches, they are treated in many cases as sharing the same underlying process (e.g., Kruschke, 2008) and there is evidence that they perform similarly, under some circumstances (van Tiel, Franke, & Sauerland, 2021). We compare the performance of the two models and their ability to predict behavioral data in two categorization tasks with gradable adjectives. Due to the context-sensitivity of gradable adjectives, we test categorization in adjectives in a specific context. We implement prototype-based and threshold-based models following the method of Douven et al. (2016; 2017) and allow for multiple prototype and threshold values to account for graded membership (vagueness) in categorization. Our experiment and simulation results show high similarity in how they fit to behavioural data, supporting, to some extent, the view that sees them as two descriptions of the same underlying process (Kruschke, 2008).

Methods

We conducted an online behavioral study to test whether the prototype or threshold model better explains human categorization with gradable adjectives. Specifically, we test the

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

Table 1: Overview of materials.

Adjective pair	Comparison class	Unit	Categorization task	prototype selection task	Model X-instances
			Mean Value, sd	min, max, every	min, max, every
short-tall	adult male	ft	5.9, 0.8	3.6, 7.6, 0.1	3.6, 7.6, 0.1
young-old	person	years	45, 15	0, 120, 3	0, 120, 1
slow-fast	cycling commute in London	mph	15, 5	0, 40, 1	0, 40, 1
cold-warm	London summer's day	°C	20, 6	-20, 40, 1.5	-6, 60, 0.5

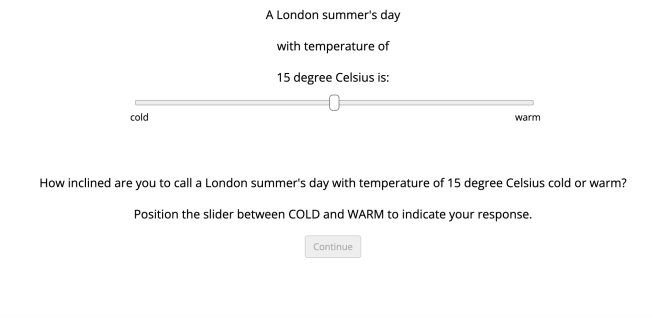

Figure 1: Example trial in the continuous categorization task for the adjective pair *cold-warm* for the temperature of *15°C*. The 'continue' button is clickable once the slider's is position is modified.

use of four pairs of gradable adjectives (*short-tall*, *slow-fast*, *cold-warm* and *young-old*) in categorization. Our behavioural task closely follows Verheyen & Égré's study (2018). In the first part of the behavioural task, participants complete two types of categorization tasks, continuous categorization and binary categorization. In the second part, participants were asked to report threshold and prototypical values for each of the 8 adjectives. In the continuous categorization task, participants are asked to indicate, using a slider, how inclined they are to describe different items using each of the appropriate pairs of adjectives. On each trial, they are presented with one of four objects (comparison class, see Table 1) with a value specifying their height/ speed/ temperature or age. Participants can then move the slider to any degree between the negative and the positive relevant adjectives to indicate the object's membership. For example, for the item 'A London summer's day', the slider can be moved between *cold* and *warm*. Figure 1 shows an example trial for this item.

In the binary categorization task, participants are presented on each trial with a statement describing an instance of one of the four objects with the respective positive adjective and are asked to indicate whether the statement is true or false. E.g., 'A person whose age is 66 years is old.' After indicating their judgement, participants are asked to indicate, using a slider, how certain they are of their response. The instances participants are asked to categorize in the continuous and binary categorization tasks are sampled normally around a mean value and standard deviation taken from previous work on gradable adjectives (Verheyen & Égré, 2018) and presented in Table 1. Both categorization tasks include 120 trials each, 30 sampled instances for each adjective pair, presented in randomized order across adjectives. The order of the continuous and binary categorization tasks is counterbalanced across participants.

In the second part of the behavioural task, in the threshold generation task, for producing the threshold value for *old*, for example, participants were asked to complete the text 'When is it true to say that a person is old? It is true to say that a person is old if their height is greater than or equal to [blank]'. For the prototypical values, participants were asked to generate a prototypical value for each adjective by filling in the text 'What age (in years) comes spontaneously to mind when you imagine an old person? [blank]' for the adjective old, for example. In a following set of trials, participants were also asked to select values they found prototypical for each adjective from a set of values presented to them (for the presented values see Table 1). The order of the threshold and prototypical values generation and selection tasks is counterbalanced across participants. Participants either complete the threshold generation task and then the prototype generation and selection tasks, or first the prototype generation and selection tasks and then the threshold generation task.

Participants 98 self-reported native English speakers participants were recruited via the Prolific crowd-sourcing platform. The mean duration of the task was 33 minutes and participants were compensated £4.5 for their time.

Models Simulation

Prototype Based Model We use participants' prototype reports for each of the eight adjectives to predict degrees of membership for all X- instances with the prototype-based model.[1] Following (Douven, 2016) approach, we predict degrees of membership to the negative or positive adjective for a set of possible X-instances. For each X-instance, we sample

[1] we use the term X-instances to refer to all values that can be assigned to an item. For example, 15 degree Celsius in the example trial in Figure 1, or 64 years for the age of a person.

a reported prototypical value of the negative adjective, one of the positive adjective, and calculate whether the instance is closer to the negative adjective's prototype or to the positive adjective one. Consider a participant who selected 8, 10, and 12 as prototypical values for temperatures of a cold London summer's day and 20, 22, and 24 as prototypes for a warm London summer's day. For each X instance in the set of temperatures from -6 to 60 degrees Celsius, we sample one prototype for *cold*, one for *warm* and compare the distances between the X-instance and the sampled prototypes. If the instance is closer to the positive prototype, we assign one as its membership, and zero otherwise. We repeat this process 10,000 times, and the predicted degree of membership for each X-instance is the average over all results from the 10,000 iterations. We combine participants' reports from the prototype generation and prototype selection tasks as the reported prototypical values. X-instances used in the models simulation presented in Table 1.

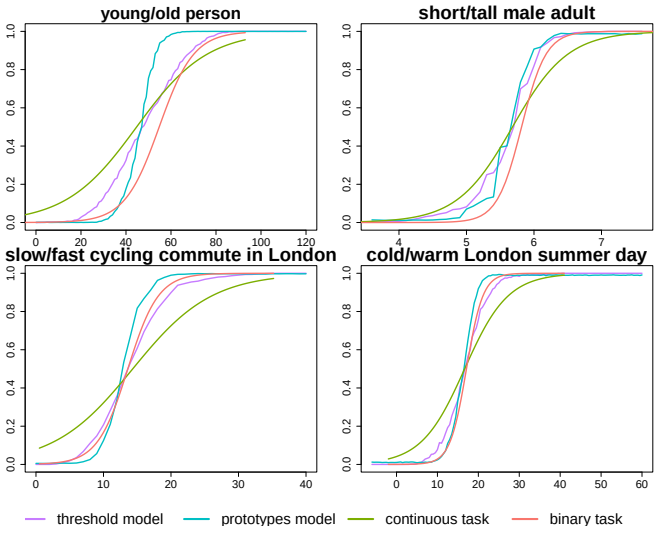

Figure 2: Predicted degree of membership curves from the two models and regression curves from the two categorization tasks, for the four pairs of adjectives.

Threshold-Based Model A similar approach was taken in creating the threshold-based predicted membership degrees. We sample from the reported thresholds values and assign zero to X-instances lower than the sampled threshold and one otherwise, averaging all zero/one values from the 10,000 iterations to predict the degree of membership for each X-instance. Threshold values were sampled from the reported thresholds for the positive adjectives, the reported threshold for the negative adjectives, and added values between the two reported thresholds, if they are unequal. X-instances are the same in the two models.

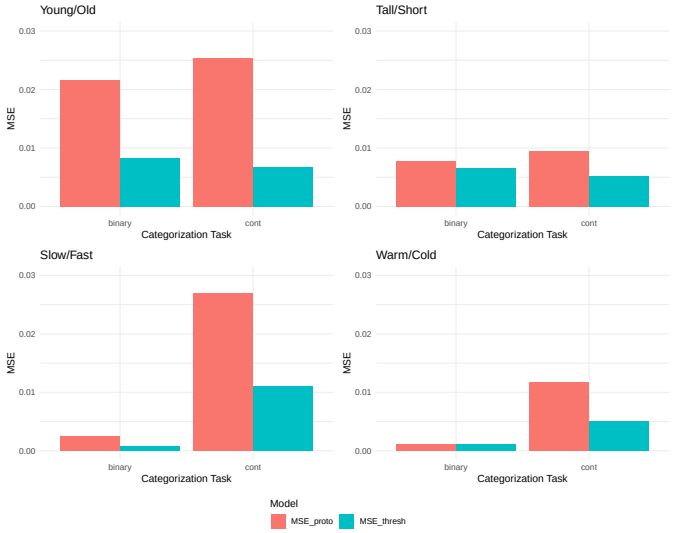

Figure 3: Mean squared error values when comparing predicted degrees of membership from the two models with logistic regression curves fitted to the data from the binary and continuous categorization tasks, for the four pairs of adjectives. Better fit of the model to the data is represented by lower MSE values.

Results

Group-Level Analysis

A logistic regression was fit to the aggregated data from the continuous and binary categorization tasks. Predicted degrees of membership from the prototype model were computed based on aggregated participants' responses to the prototype generation task. Duplicate values in participants' responses were kept in the aggregated data to give higher weights to more frequent values. The same was done for computing degrees of membership by the threshold model, based on aggregated responses to the threshold generation task. The two regression curves of the aggregated data from the binary and continuous tasks, together with the predicted membership values from the prototype-based model and the threshold-based model are presented in Figure 2 for each of the adjective pairs. To compare the differences between the two models' predictions and categorization data, we compute the mean squared error (MSE) between the two curves from the models and each of the regression curves from the categorization tasks. Figure 3 shows the resulting MSE values. Across all adjective pairs, the threshold model predicts better the degree of membership from the categorization tasks, both the binary and continuous tasks, as it shows lower values of MSE than the prototype model. For *slow-fast* and *warm-cold*, the two models predict better the data from the binary task rather than the continuous, while in the case of *young-old* and *tall-short*, there are no clear differences across tasks.

Since we are interested in the underlying cognitive model for categorization with gradable adjectives, we can not rely

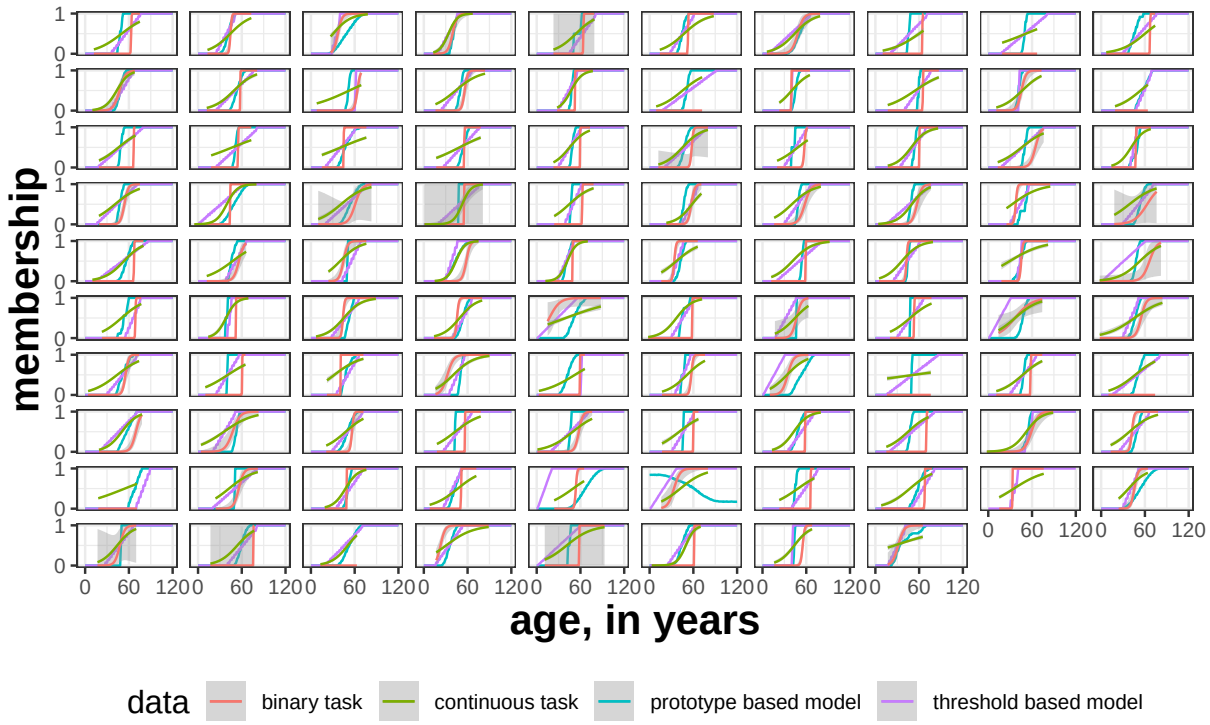

Figure 4: Four predicted degrees of membership curves from the two models and the two categorization tasks, per participant for the adjective pair *old-young*.

on group-level data analysis solely, but look at the individual level data. In addition, it is plausible that there are individual differences in the representation of gradable adjectives, both in the adjectives that individuals assign to the same instance (e.g., while person A might classify person B as tall, C might think he is short) and in the underlying models they use to categorize instances in the world with gradable adjectives.

Individual-Level Analysis

A logistic regression curve was fit to each participant's data from the binary and continuous categorization tasks. Degrees of membership from the prototype and threshold models were computed based on each participant's reports. Resulting curves per participant for the adjective pair *old-young* for *a person* are presented in Figure 4. Overall, prediction curves from the two models overlap to a high extent with the curves from participants' categorization tasks data.

To measure the degree to which the curves overlap and to estimate which of the models better predicts the behavioural categorization data in the individual level, we compute mean squared errors (MSE) between the predicted degrees of membership generated by the prototype and the threshold models and the behavioural data from the binary and continuous categorization tasks for each participant. Figure 5 shows density distribution of resulting MSE values for all participants, and the difference in MSE values from the two models (right-hand panels) for the adjective pair *old-young*. Degrees of membership generated by the threshold-based and prototype-based models highly predict the degrees of membership elicited from the categorization tasks directly for most of the participants, as the peaks of the MSE density distributions that compare the prototype and threshold model to data from the two categorization tasks are almost at zero. For the case of *young* and *old* this can be seen in the left and middle panels of Figure 5[2]. There is more variation across participants in how well the two models fit to the data from the binary categorization task compared to the continuous task. This is expressed in wider distributions of the binary categorization data (upper panels). The same trend is visible in all adjective pairs tested (see Figure 6).

As for the differences between the two models in their ability to predict behavioural data, in predicting degrees of membership from the binary categorization task, the threshold-based predictions match predictions from the prototype model for most of the participants, as the peak of the density distribution presenting the difference in MSEs for the binary task data is at zero. The right tail of the density distribution is thicker than the left tail, in all adjective pairs, suggesting that for more participants the threshold model predicted better the behavioural data from the binary task compared to the threshold model. Looking at how well the models predict membership degrees from the continuous categoriza-

[2]MSE density distributions from the three other adjective pairs will become available on OSF

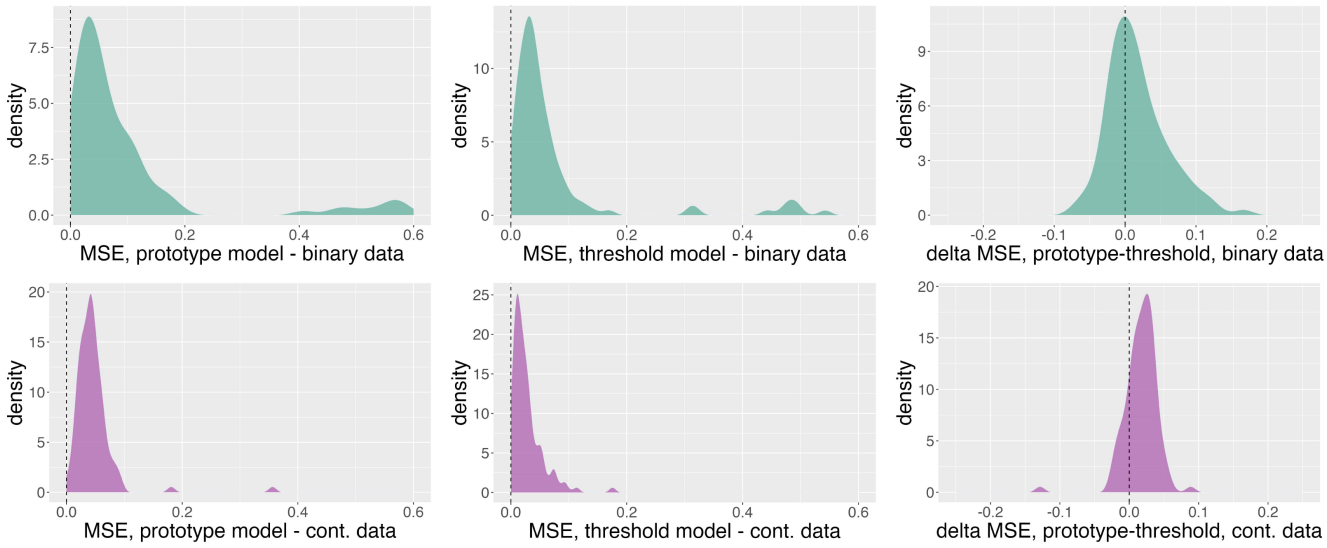

Figure 5: Density distributions of mean squared errors between the predicted degrees of membership by the prototype-based model (left panels), threshold-based model (middle panels) and data from the binary categorization task (upper panels) and the continuous categorization data (bottom panels) for the pair of adjectives *old-young*. Panels on the right present the difference in MSE values between the prototype and threshold model in predicting the data from the two categorization tasks. The vertical dashed line at zero reflects the perfect overlap of the curves. Positive values represent a better fit of the threshold-based model to the behavioural data.

tion data, there is a slightly greater advantage for the threshold model as for most of the participants prediction from the threshold model overlapped more with data from the continuous task. This is clearly visible from density distribution for *old-young*, and to a lesser degree in the other adjective pairs (Figure 6). Nevertheless, the peak of the density distribution is very close to zero, with a narrow shape, suggesting that the differences between the models are not substantial.

Discussion

Our results suggest that the threshold model and the prototype model predict well behavioural data from categorization tasks for the four gradable adjective pairs we tested. When comparing the two models in how well they fit the data, our results show that there is a high similarity in the predictive power of the two models. We see that, for most of the participants, there is very little difference in the mean squared errors from the two models. We see a slight advantage of the threshold model in predicting the behaviuoral data from the continuous categorization task.

Note that the threshold and prototype models simulation were based on participants' explicit reports of threshold and prototype values. It is interesting to see that, in general, participants were good at estimating their implicit values for categorization, as the models output overlapped quite well with the behaviuoral data.

It is not our objective in this paper to compare across adjective pairs, as differences in data between adjective pairs may reflect pragmatic differences or differences related to the comparison classes. We look at four pairs of adjectives in order to be better able to generalize observations from the tested adjectives to gradable adjectives.

It is important to note that the conclusions drawn rely on descriptive statistics and visual inspection. A future objective is thus to formalize these two models within a common framework, such as a computational model, and to assess them using model comparison tools. This assessment may involve measures such as likelihood-ratio tests, model fit indices, or Bayes factors.

Acknowledgments

This research was funded by the Dutch Research Council (NWO) under Gravitation grant Language in Interaction, grant number 024.001.006. A.S. is supported by an Amsterdam Brain and Cognition (ABC) project grant (ABC PG 22).

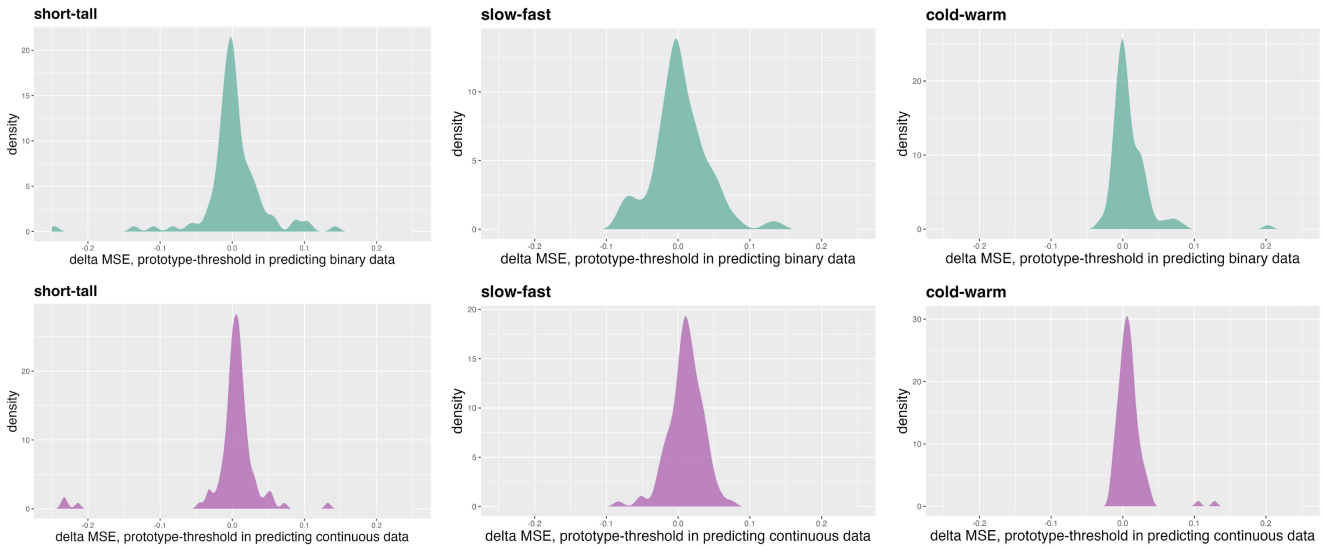

Figure 6: Density distributions of differences in mean squared errors for all adjectives, in predicting the binary categorization data (upper panels) and the continuous categorization data (bottom panels). Positive values represent a better fit of the threshold-based model to the behavioural data.

References

Decock, L., & Douven, I. (2014). What is graded membership? *Noûs*, *48*(4), 653–682.

Douven, I. (2016). Vagueness, graded membership, and conceptual spaces. *Cognition*, *151*, 80–95.

Douven, I., Wenmackers, S., Jraissati, Y., & Decock, L. (2017). Measuring graded membership: The case of color. *Cognitive Science*, *41*(3), 686–722.

Gärdenfors, P., & Williams, M.-A. (2001). Reasoning about categories in conceptual spaces. In *Ijcai* (pp. 385–392).

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, *30*, 1–45.

Kruschke, J. K. (2008). Models of categorization. *The Cambridge handbook of computational psychology*, 267–301.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory* (Vol. 23, pp. 587–610).

Pezzelle, S., & Fernández, R. (2023). Semantic adaptation to the interpretation of gradable adjectives via active linguistic interaction. *Cognitive Science*, *47*(2), e13248.

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. In *Semantics and linguistic theory* (Vol. 24, pp. 23–41).

Ramotowska, S., Haaf, J., Van Maanen, L., & Szymanik, J. (2022). Most quantifiers have many meanings.

Schmidt, L. A., Goodman, N. D., Barner, D., & Tenenbaum, J. B. (2009). How tall is tall? compositionality, statistics, and gradable adjectives. In *Proceedings of the 31st annual conference of the cognitive science society* (Vol. 31, pp. 2759–2764).

van Tiel, B., Franke, M., & Sauerland, U. (2021). Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences*, *118*(9), e2005453118.

Verheyen, S., & Égré, P. (2018). Typicality and graded membership in dimensional adjectives. *Cognitive Science*, *42*(7), 2250–2286.