Blending Boundaries: A Computational Approach to How Bilinguals Reconcile Cross-Linguistic Categorization

Aditi Singh[*,1] **(as11778@nyu.edu)** **Maya Taliaferro**[*,1] **(mjt10029@nyu.edu)**

Grace Lindsay[1,2,3] **(gm3239@nyu.edu)** **Esti Blanco-Elorrieta** [1,2] **(eb134@nyu.edu)**

[1] Department of Psychology, New York University
[2] Department of Neural Science, New York University
[3] Center for Data Science, New York University
[*] Equal lead contribution

Abstract

We categorize the world using labels that aid memory, recognition, and generalization. While some concepts have clear boundaries, others are more fluid, leading to cross-linguistic differences. How bilinguals manage these differences remains unclear. We investigate this by comparing English monolinguals, Mandarin monolinguals, and Mandarin-English bilinguals in a 2AFC task to test whether bilinguals' categorization aligns with monolingual norms or forms an integrated system. Additionally, we develop a neural network model to simulate category boundary formation under varying language exposure. Our model closely mirrors behavioral data, supporting the idea that bilinguals develop a shared categorization system shaped by dominant language exposure. This combined behavioral and computational approach offers new insights into how bilinguals resolve cross-linguistic conflict and the cognitive mechanisms underlying multilingual concept organization.

Keywords: Bilingualism, conceptual category systems, categorization models

Introduction

To navigate the complexity of the world, we categorize both concrete and abstract concepts by assigning them unique labels that facilitate memory, recognition, and generalization(Medin & Smith, 1981; Corter & Gluck, 1992). Traditional theories suggest that categorization is based on shared features, with boundaries formed where similarities end (Rosch & Mervis, 1975; Medin & Schaffer, 1978). However, these boundaries vary in rigidity. Some categories have clear-cut features that make category membership essentially binary (i.e., possessing or lacking a specific feature directly determines inclusion) while others are more fluid (William, 1973; Malt, Sloman, Gennari, Shi, & Wang, 1999). For example, biological categories tend to have distinct and universal boundaries often attributed to natural discontinuities that "cry out to be named" (Berlin, 1992). In contrast, other concepts such as human-made artifacts, tend to have fuzzier boundaries. This difference in boundary rigidity has implications for the variability in labels used across individuals. Categories with well-defined boundaries tend to be less susceptible to label variability because membership in the category is relatively stable and clear. Categories with fuzzier boundaries, on the other hand, tend to be more prone to variability in labeling, as membership can be more influenced by factors like context and functional goals (Ameel, Storms, Malt, & Sloman, 2005; Ameel, Malt, Storms, & Van Assche, 2009; Nosofsky, 1986; Malt & Sloman, 2007).

One such potentially influencing factor is language itself. Different languages may emphasize different features (shaped by cultural and linguistic needs), leading to category segmentations that vary significantly across languages (Berlin & Kay, 1991; Malt, 2024). For example, Kronenfeld, Armstrong, and Wilmoth (1985) examined how English, Hebrew, and Japanese speakers categorize drinking vessels and found that all three groups rated the similarity of objects nearly identically, even though their labeling patterns varied significantly. In English, the critical feature for categorizing a vessel as a "glass" was material, thus making paper and plastic containers categorized as "cups." In contrast, Hebrew speakers relied on shape, and grouped paper and Styrofoam cups with glass vessels under the term "cos." Hebrew and Japanese further differed in the shape associated with their corresponding categories: in Hebrew a typical "cos" had a cylindrical shape without handles, while the most typical Japanese "gurasu" were non-cylindrical stemmed objects made of glass. This example illustrates that different languages emphasize distinct features when defining category boundaries without necessarily altering the way those features are perceived (Malt, Sloman, & Gennari, 2003). Thus, categorization is the product of both an object's perceivable features and the language experience of the individual.

This raises the question of how bilinguals navigate categorization in instances where boundaries are fuzzy and do not overlap across their languages. One possibility is that bilinguals maintain separate conceptual systems for each language, categorizing objects in a "monolingual-like" manner and switching between these systems depending on the language being used. Alternatively, in line with a common mechanism account of bilingualism (Blanco-Elorrieta & Caramazza, 2021), bilinguals might develop a shared and unified conceptual system that integrates features from both languages (Ameel et al., 2009; White, Malt, & Storms, 2017). Under this hypothesis, bilinguals' categorization would converge across languages, resulting in the same categorization independent of which language is being used. Importantly, the nature of this convergence could vary among bilinguals based on factors such as the extent of cultural immersion and the length of exposure to each language. For some bilinguals, one language could have a stronger influence on categorization than the other. For others, both languages might equally shape their categorization principles and conceptual system.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

In this study, we examine: (i) whether objects with continuous rather than dichotomous features lead to fuzzier category boundaries (ii) whether cross-linguistic differences in naming emerge for categories with fuzzier boundaries, (iii) how bilinguals reconcile these differences across languages, and (iv) how language exposure shapes this reconciliation. To address these questions, we analyze how English monolinguals, Mandarin monolinguals, and Mandarin-English bilinguals label everyday objects that gradually shift from one category to another (Figure 1a). Specifically, we first assess naming variability within each language for objects that either have or lack dichotomous distinguishing features. Low variance in naming would suggest clear category boundaries, while high variance would indicate fuzzier boundaries. Next, to test whether language boundaries diverge when category distinctions are more ambiguous, we compare category boundaries for English and Mandarin monolinguals. We expect more similar boundaries for objects with low within-language variance, but greater cross-linguistic differences for those with higher within-language variance. Then, we examine how bilinguals' categorization compares to monolinguals' and investigate whether bilinguals maintain separate conceptual systems for each language or integrate them into a shared system. Finally, we characterize the constraints that shape bilingual conceptual systems, focusing on the nature and duration of language exposure. To achieve this, we develop a cognitively plausible deep-learning model (Figure 3) that tests whether differences in language exposure causally influence bilingual category boundaries. Combining behavioral and computational methods, we demonstrate that fuzzier, continuous boundaries lead to cross-linguistic naming differences, and that bilinguals resolve these conflicts through a shared semantic system shaped by the relative dominance of each language.

Determining Category Boundaries

To identify category boundaries and determine i) whether they are fuzzy or clear, ii) whether they remain stable or vary across languages, and iii) how they differ between monolinguals and bilinguals, we conducted a two-alternative forced-choice (2AFC) task (Figure 1b). Participants classified objects that gradually transitioned between concepts, enabling us to map category boundaries within each language (Figure 1a depicts all continua). All participants provided informed consent and were compensated for their time. The study was approved by the Institutional Review Board.

Stimuli. We chose pairs of related everyday objects (e.g., a plate and a bowl) and created a continuum from one object to the other by manipulating their visual features (e.g., depth) in a gradual way. Some continua had dichotomous features that delimited the category boundary (i.e., a teapot becomes a mug when the spout disappears; 'Control continua') whereas other continua had continuous features (i.e., a plate becomes a bowl at a certain depth; 'Experimental continua'). The images were created as realistic 3D models using Blender, a free and open-source 3D graphics software, and displayed as 500x500px images. Initially, we created 15 continua that were later normed to select the best 6 for the 2AFC task. The norming procedure was as follows. 30 monolingual English (19 females, 11 males; $M_{\text{age}} = 37.3, SD = 11.0$) and 30 monolingual Mandarin (16 females, 14 males; $M_{\text{age}} = 30.9, SD = 6.7$) participants generated labels for each object category. We used the first (step 1), middle (step 4) and last (step 7) steps of each continuum for this procedure. Participants viewed objects one at a time in a semi-randomized order. To minimize priming effects, middle-category objects were presented first, to prevent earlier exposure to unambiguous category exemplars from biasing their judgments. We used the most frequent names for steps 1 and 7 to determine the "true" category labels of each continuum (i.e., "bowl" and "plate"). We used the labeling distributions for the middle objects to i) determine if a third category existed between the two ends of a continuum (e.g., in continuum from "couch" to "chair", a blended image in the middle would be called a "loveseat") and ii) assess whether the category boundary differed across languages (i.e., whether the label change occurred at the same middle step in both languages). We selected continua that did not include a third (in-between) category between both ends and where the most frequent label for the first and last steps accounted for at least 75% of the responses. Based on these criteria, we chose six continua. Three were "Control Continua," where category boundaries overlapped across both languages (Watercan to Bucket, Teapot to Mug, and Hammer to Ax). The other three were "Experimental Continua," where category boundaries differed between the languages: Plate to Bowl, Bucket to Pot, and Spatula to Ladle. See Figure 1a for all continua.

2AFC Task. *Participants.* We recruited 24 monolingual English speakers (13 females, 11 males; $M_{\text{age}} = 24.50$ years, $SD = 7.72$; $M_{\text{proficiency}} = 6.72$; $M_{\text{AoA}} = 0.82$ years), 21 monolingual Mandarin speakers (8 females, 13 males; $M_{\text{age}} = 32.82$ years, $SD = 11.00$; $M_{\text{proficiencyMandarin}} = 6.98$, $M_{\text{proficiencyEnglish}} = 2.2$; $M_{\text{AoAMandarin}} = 0.0 years$, $M_{\text{AoAEnglish}} = 18.75$ years) and 24 bilingual speakers (18 females, 5 males, 1 non-binary/prefer not to say; $M_{\text{age}} = 22.92$ years, $SD = 3.35$; $M_{\text{proficiencyMandarin}} = 6.67$, $M_{\text{proficiencyEnglish}} = 5.56$; $M_{\text{AoAMandarin}} == 1.5$ years, $M_{\text{AoAEnglish}} = 7.76$ years) for this task. Thus, all our bilinguals were sequential bilinguals. Monolinguals who reported significant proficiency in a 2nd language (above "3" out of "7") or bilinguals who were proficient in a 3rd language and/or not proficient in both English and Mandarin were excluded from our dataset. All participants reported normal or corrected-to- normal vision and no history of neurological or language disorders.

Procedure. Participants underwent a familiarization phase where they saw the objects corresponding to the first and last steps of each continuum accompanied by their category labels. Then, they practiced the 2AFC task, seeing only the first and last object of each continuum, with online feedback

on their responses. In the main task, they received no feedback. The experiment included 84 unique images (6 continua x 7 steps x 2 colors), each shown 8 times, totaling 672 trials. Participants saw one image from a continuum and two labels (e.g., "Bowl" and "Plate" for the Bowl-to-Plate continuum) and selected the label that best matched the object. All instructions and communication were conducted in the testing language. Bilingual participants completed the task in both English and Mandarin, with at least 24 hours between sessions and counterbalanced testing order. We excluded participants who did not assign the correct label to the unambiguous exemplars (steps 1 and 7) or did not respond to at least 80% of the trials. We removed 3.36% of monolingual English, 2.87% monolingual Mandarin, 3.24% bilingual English and 2.50% bilingual Mandarin trials due to no responses.

Each trial started with a fixation cross (300 ms), followed by presentation of the target object. After 800 ms the two possible labels appeared under the object and the participant had to pick the best fitting one using a button box before a 2000 ms timeout (see Figure 1b for task design). The inter-stimulus intervals were randomly sampled from a uniform distribution with a range of 400-600 ms. Presentation of each trial was semi-randomized such that no consecutive trials showed objects from the same continuum. The experiment was presented using Psychopy [2023.1.3].

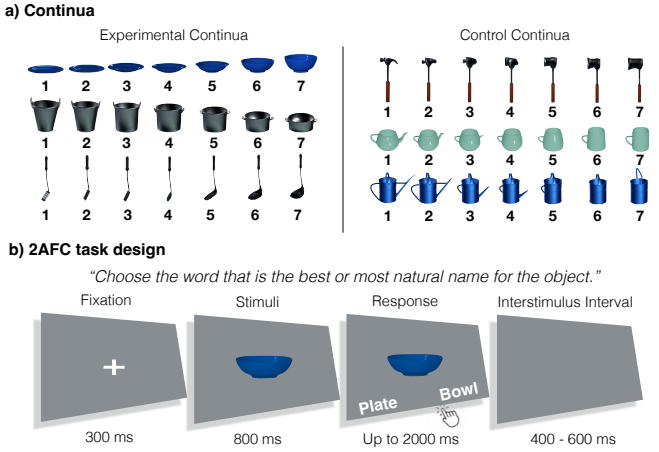

Figure 1: Behavioral Task. Panel a) depicts Experimental (left) and Control (right) continua. Panel b) depicts the two alternative forced choice task performed by participants.

Discrete Features Lead to Stable Boundaries, Continuous Features to Fuzzy Boundaries

Analysis. An initial goal of our behavioral analysis was to examine whether continua that rely on continuous boundaries (i.e., our "Experimental continua") as opposed to dichotomous boundaries (i.e., "Control continua") would result in less stable within-language categorization. To evaluate the clarity of category boundaries, we analyzed the consistency of object labeling within each language group. We defined the category boundary for each continuum as the step where participants were equally likely to assign the label for category A or category B (i.e., crossover step). Then, we conducted Levene's test for homogeneity of variances on the variance in crossover steps based on language group, continuum type, and the interaction between these factors.

Results. Our results support a distinction in boundary stability between "Control" and "Experimental" continua. Specifically, we observed a significant difference in variance between the "Control" and "Experimental" continua, $F(1,540) = 286.76, p < 0.001$. Bonferroni-corrected pairwise comparisons revealed that within all language groups, the participants were much less consistent in labeling the "Experimental" than the "Control" continua ($p < 0.001$ for all groups).

Fuzzy Boundaries Diverge Across Languages

Analysis To determine the overlap of category boundaries across languages, we used linear interpolation and calculated the boundary for monolingual English vs. monolingual Mandarin language groups. Then we compared them through a two-sample Welch's t-test.

Results. Our analysis showed that while in the "Control Continua" there was no difference in the category boundary between Monolingual English and Monolingual Mandarin speakers $t = -0.49, p = 0.62, CI = [-0.07 : 0.04]$; English monolinguals and Mandarin monolinguals significantly diverged in the boundaries for the "Experimental continua" ($t = 4.82, p < 0.001, CI = [0.07 : 0.18]$), thus showing that it is continuous, fuzzy boundaries that lead to differences in cross-linguistic segmentation of the conceptual space.

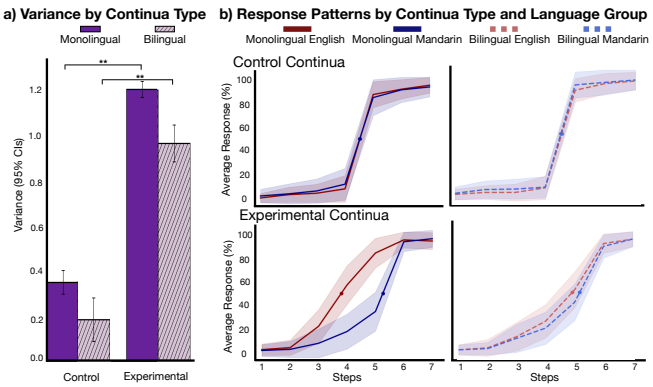

Figure 2: Behavioral Results. Panel a) depicts the differences in variance by continua type and participant type (monolingual or bilingual). Variance is measured by range in 95% CIs around mean crossover step. Panel b) depicts the averaged naming patterns for each language group collapsed across continua type (Control or Experimental). The x-axis represents the item step number, and the y-axis shows the percent of responses assigned to the second category. Shading around lines indicate 95% confidence intervals.

A Shared Conceptual System in Bilinguals

Analysis. Having established the lack of overlap of the "Experimental" boundaries across languages, the primary goal of our behavioral analysis was to test competing theories about bilingual concept categorization. Specifically, we sought to determine whether bilinguals maintain separate, monolingual-like systems for each language or develop a shared, blended system that integrates features of both languages. Using linear interpolation, we calculated the boundary for bilinguals in each language and compared them through a series of two-sample Welch's t-tests. The comparisons included: (a) bilingual English vs. monolingual English, (b) bilingual Mandarin vs. monolingual Mandarin, and (c) bilinguals in English vs. bilinguals in Mandarin.

Results. Our results provide compelling evidence for a shared semantic system even in sequential bilinguals. While the results from monolingual English and Mandarin speakers showed that these languages have distinct categorical boundaries, the bilinguals' crossover steps were not different in Mandarin and English ("Experimental continua": $t = -0.29, p = 0.19, CI = [-0.02 : 0.08]$). Taken together, these findings confirm that although categorization for "Experimental continua" differs across languages, it remains unified in bilinguals. This suggests that bilinguals do not keep separate semantic systems that align with the distinctions that each language makes, but rather they have a unified system that integrates both of their languages.

Contrary to previous findings (Ameel et al., 2009; White et al., 2017), the convergence observed in our bilinguals did not reflect an even, bidirectional influence of both languages. Instead, our bilinguals' category boundaries were significantly different from the monolingual English boundaries ($t = -3.81, p < 0.001, CI = [-0.15 : -0.05]$) but not the monolingual Mandarin boundaries ($t = -0.04, p = 0.97, CI = [-0.05 : 0.05]$). Follow-up two-sample Kolmogorov–Smirnov tests clarified that the bilingual distribution was more similar to that of monolingual Mandarin ($D = 0.08$, $p = 0.281$) than monolingual English speakers ($D = 0.13$, $p < 0.01$). As previously described, bilingualism is a multilayered phenomenon shaped by a complex combination of factors (e.g., Age of Acquisition, Proficiency, and Exposure to each language) and it remains unknown which of these factors influence—or even determine—the structure of bilinguals' conceptual space. One hypothesis is that Bilinguals' conceptual system is highly malleable, with its boundaries shifting based on the probabilistic distribution of exposure to different segmentation patterns. Under this hypothesis, Exposure to each language could shape bilinguals' conceptual system. To explore this possibility and account for the pattern in our behavioral data, we created a computational model that simulates how category boundaries emerge in bilinguals under varying language Exposures.

A Computational Account of the Shared Conceptual System

To ensure a bilingual model's relevance to our behavioral data and allow for reasonable parallels from computation to cognition, we designed our model to align with fundamental principles of object recognition and categorization. (Richards et al., 2019; Saxe, Nelli, & Summerfield, 2020). Based on feature-based cognitive theories of categorization that break down objects into their defining characteristics (Rosch & Mervis, 1975; Medin & Schaffer, 1978; Nosofsky, 1986), we relied on a neural network model architecture, which is inherently suited to map low-level features to higher-level abstractions (Bishop, 1995). Specifically, our model learned to take in an image and predict its category by first identifying key object features and then incorporating language as a general contextual cue before prediction. The bilingual model was trained on images similar to the behavioral experiment and their accompanying labels in Mandarin or English. Then, we tested the model on our 2AFC behavioral experiment, and it successfully reproduced the behavioral results by predicting a shared bilingual conceptual boundary. Matching this behavioral benchmark allowed us to subsequently use this model to unpack the computational principles underlying integration of conceptual information and to precisely characterize the influence of an important sociolinguistic factor, Exposure, which is hard to characterize through (often inexact) longitudinal behavioral studies.

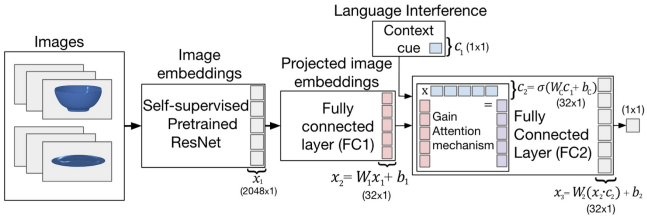

Figure 3: Model Architecture, using the Plate to Bowl continuum as an example. Each model was ultimately trained on the experimental and control continua separately.

Model Architecture. We built a model consisting of a pretrained ResNet-50 for image embeddings, followed by two fully connected dense layers (FC1 and FC2) that, when put together, transformed the high-dimensional visual information from the training images into lower-dimensional abstract representations of categories (e.g., "Bowl" or "Plate") as a binary response (Figure 3). We will refer to FC1 and FC2 as non-linear layers because they were connected through a rectified linear unit (ReLU) activation function. Each non-linear layer had a weight W and bias b parameter, regulating how much influence different features in that layer's input exerted on its output. These model parameters were initially randomized, but as the model computed errors of its category predictions during training, its weight and bias parameters updated through the process of gradient descent during backpropagation from FC2 to FC1. By the end of model training

and parameter optimization, FC1 and FC2 identified key patterns in the training images and assigned greater weight to crucial features for classification.

The pre-trained, self-supervised ResNet-50 captured low-level image features without pre-existing category biases and served as the input to FC1, denoted by x_1. FC1 then transformed x_1 into a lower-dimensional feature representation (x_2) via a non-linear transformation: $x_2 = \text{ReLU}(W_1 x_1 + b_1)$, where W_1 (weights) and b_1 (bias) were parameters specific to FC1. FC1's output (x_2) was passed to FC2, which reconciled visual information from x_2 with a binary language context cue (c_1) that specified whether the model should respond in English or in Mandarin. FC2 then generated an attention score vector $c_2 = \sigma(W_c c_1 + b_c)$ using c_1, and weighted the feature embedding x_2 depending on the language cue. The result of this mechanism, similar to gain attention and feature-recalibration machine-learning methods (Hu, Shen, & Sun, 2018), was a linguistically influenced, refined feature embedding: $x_3 = x_2 \cdot c_2$. This context cue mirrored our behavioral paradigm in which bilingual participants were asked to complete the task in a specific language. Finally, the model's output from FC2 was a (language-adjusted) binary prediction of which category the input image belonged to.

Training & Testing Data. We created two monolingual training datasets—one "English," and the other "Mandarin"—which contained the same training images, but with different image labels that matched either our monolingual Mandarin or monolingual English behavioral responses. This was to parallel the fact that a bilingual speaker would likely not be explicitly trained on a bilingual conceptual representation. We took each of the 3 Experimental and 3 Control continua from our behavioral experiment and expanded them from 7-step to a 10-step continua for our training dataset. We then took each image in these 10-step continua and created variations in color, angle, and position to ensure that the model wasn't simply learning a one-to-one mapping between images and labels and was instead generalizing to abstract categories. To further test generalizability, we excluded the most ambiguous step (fourth) from all training datasets. Each experimental continuum dataset contained 945 training images (7 colors × 3 angles × 5 positions × 9 steps), while each control continuum dataset had 972 (9 colors × 4 angles × 3 positions × 9 steps). This slight difference in the number of positions arose from the symmetrical structure of two of the Experimental continua, which made left-to-right rotations indistinguishable. We also created a separate 7-step dataset for testing only, which came to a total of 168 images (3 colors × 2 angles × 4 positions).

Model training procedure. We trained all models using an Adam optimizer with a 1-cycle learning rate scheduler and a standard cross-entropy loss function. Typical values ranged from [0.0005–0.005] for the learning rate and weight decay, and [0.1–0.5] for label smoothing. For monolingual models, hyper-parameters (learning rate, weight decay, label smoothing) were fine-tuned per continuum to maximize fit. The bilingual models then inherited hyper-parameters from its monolingual counterparts. We trained each model in epochs, where one epoch referred to one complete run of all the training images. We tracked the model's training trajectory by plotting its estimated probability for each category across epochs. We report all models' estimated percent probability with respect to category B (e.g., "Bowl" in Plate to Bowl continuum) such that 0% indicates certainty in category A ("Plate") and 100% indicates certainty in category B ("Bowl"). If completely unsure about an image's category, the model's estimated percent probability would be close to 50%. Each epoch was subdivided into batches of 64 training examples at a time; at the end of each batch, the model updated its weight and bias parameters through gradient descent. Since the bilingual model received a random mix of images with conflicting category labels from the monolingual English and Mandarin datasets, we implemented a gradient projection-based balancing rule (Yu et al., 2020) to encourage the model to learn from *both* datasets. Specifically, when gradients from English and Mandarin datasets pointed in opposite directions (i.e., where the labels across languages didn't overlap, which resulted in a negative dot product), the gradient-balancing mechanism adjusted one gradient by removing its component along the other gradient's direction.

Finally, to control each bilingual model's Exposure to each language, we fixed the ratio of monolingual English and monolingual Mandarin training examples in each batch, in each epoch. The model continued to have a batch size of 64, but it would unequally sample from the two monolingual datasets to create batches until one dataset was depleted–after which, a new epoch would begin. After this, the model reshuffled and resampled to begin the next epoch. This batch sampler controlled the model's exposure to each language's category boundaries. We tested three scenarios: (1) greater Mandarin exposure (90% Mandarin -10% English split), (2) greater English exposure (10% Mandarin -90% English split), and (3) balanced Exposure (50% Mandarin -50% English split). If Exposure determines bilinguals' category boundaries, then adjusting the training ratio should systematically shift the model's bilingual boundary. We chose very unequal ratios (10%-90% and 90%-10%) to simulate unbalanced Exposure, as the model learned to perform successfully on this simple 2AFC task.

Model evaluation. We verified that our model architecture could learn conceptual boundaries by separately training a monolingual English and a monolingual Mandarin model and evaluating whether we replicated the behavioral results. Then, we trained the bilingual models on the monolingual boundaries—now verified by the monolingual models. All models were trained for 50 epochs without early stopping, and we qualitatively determined Epoch 20 to mark the end of the models' training trajectory as the models' softmax predictions stabilized after this point. We did this instead of relying on validation accuracy as we expected our bilingual model to learn a boundary it was not explicitly trained on. Over train-

ing, models transitioned from softmax percent probabilities near 50% (uncertain predictions) to more certain predictions of each category (0% or 100%).

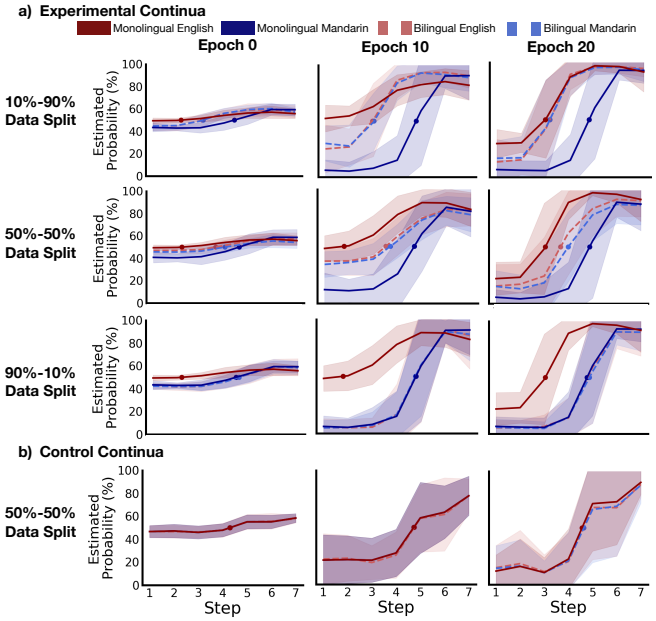

Figure 4: Averaged model predictions of estimated probabilities for each step along the 7-step behavioral continua. Confidence intervals used the average softmax probability estimates across all continua at each step. For example, in the Plate to Bowl continua, estimated probability for category "Bowl" is plotted, along with 95% confidence interval in a lighter shade. (a) shows the experimental continua, for the 10% Mandarin - 90% English training data (top), 50% Mandarin - 50% English training data (middle), and 90% English - 10% Mandarin training data (bottom); (b) shows the control continua for the 50%-50% training data split. Confidence intervals use the average softmax probability estimates across all continua at each step

Language Exposure Shapes Conceptual Space

While all bilingual models converged to a single categorical boundary across languages, replicating our behavioral results, the computational model showed that the nature of this boundary was causally shaped by exposure in each language. The unbalanced models (90%-10% and 10%-90%), converged to a category boundary that was indistinguishable from the dominant language boundary (Figure 4a; 10% Mandarin - 90% English bilingual boundary vs. monolingual English boundary: ($t = 1.5070, p = 0.1387$); 90% Mandarin - 10% English bilingual boundary vs. monolingual Mandarin boundary: ($t = 0.9613, p = 0.3414$). In contrast, the balanced model (50%-50%) converged to a crossover point that was midway between the monolingual English and Mandarin boundaries and was significantly different from both of them (bilingual Mandarin boundary vs. monolingual Mandarin boundary; $t = -8.1070, p < 0.001$, and bilingual English boundary vs. monolingual English boundary; $t = 6.1029, p < 0.001$). This model thus suggests that while bilinguals solve incongruences in cross-linguistic categorization through a shared conceptual space, where exactly the boundaries of this space lie is malleable and can be modulated by language Exposure (i.e., familiarity with a label in a given language).

Additionally, evaluating our model's performance at various epochs during the training process allowed us to explore a well-known developmental phenomenon: bilingual individuals tend to take longer to learn labels than their monolingual counterparts. Our findings suggest that this delay is due to exposure to conflicting labels for the same object. In the model with overlapping category boundaries across languages (Control Continua; Figure 4b), the bilingual model learned boundaries at the same rate as the monolingual model. However, in the Experimental Continua, the bilingual model required more training epochs to establish a clear category boundary compared to its monolingual counterpart.

Discussion

This study provides new insight on how bilinguals handle conflicting category representations across languages through both behavioral and computational approaches. We found i) that objects with continuous, fuzzier boundaries were prone to distinct cross-linguistic categorization and ii) that bilinguals reconcile these cross-linguistic categorization differences by relying on a shared semantic system that integrates both languages (see also (Ameel et al., 2009; White et al., 2017)). Importantly, this supports a common framework of bilingual cognition (Blanco-Elorrieta & Caramazza, 2021), which posits that bilinguals rely on a unified semantic space across languages, where the semantic features for each lexeme will be drawn from a shared pool, but where semantic representations of translation equivalents can be associated with distinct features. Further, our computational model found that language exposure causally modulates category boundaries. Specifically, our modeling suggests that bilingual conceptual spaces are highly flexible, adapting dynamically based on the probabilistic distribution of linguistic input. By adjusting the ratio of language exposure in each training dataset of the bilingual model, we observed systematic shifts in the model's category boundaries, demonstrating that these boundaries are not fixed but instead causally determined by the relative dominance of each language. Although this model was designed solely to examine the impact of Exposure on the bilingual conceptual space, one possibility is that once a conceptual system is established (as determined by the first acquired language), the second language inherits those boundaries and categorizes objects accordingly. If this were the case, Age of Acquisition would be another key factor determining bilingual categorization patterns. Future modeling work could disentangle these sociolinguistic factors by simulating variations in Age of Acquisition, along with Exposure.

References

Ameel, E., Malt, B. C., Storms, G., & Van Assche, F. (2009). Semantic convergence in the bilingual lexicon. *Journal of memory and language*, *60*(2), 270–290.

Ameel, E., Storms, G., Malt, B. C., & Sloman, S. A. (2005). How bilinguals solve the naming problem. *Journal of memory and language*, *53*(1), 60–80.

Berlin, B. (1992). *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies* (Vol. 185). Princeton University Press.

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution*. Univ of California Press.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford University PressOxford. Retrieved from http://dx.doi.org/10.1093/oso/9780198538493.001.0001 doi: 10.1093/oso/9780198538493.001.0001

Blanco-Elorrieta, E., & Caramazza, A. (2021). A common selection mechanism at each linguistic level in bilingual and monolingual language production. *Cognition*, *213*, 104625.

Corter, J. E., & Gluck, M. A. (1992). Explaining basic categories: Feature predictability and information. *Psychological bulletin*, *111*(2), 291.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 7132–7141).

Kronenfeld, D. B., Armstrong, J. D., & Wilmoth, S. (1985). Exploring the internal structure of linguistic categories: An extensionist semantic view. *Directions in cognitive anthropology*, 91–113.

Malt, B. C. (2024, January). Representing the world in language and thought. *Topics in Cognitive Science*, *16*(1), 6–24. Retrieved from http://dx.doi.org/10.1111/tops.12719 doi: 10.1111/tops.12719

Malt, B. C., & Sloman, S. A. (2007). Category essence or essentially pragmatic? creator's intention in naming and what's really what. *Cognition*, *105*(3), 615–648.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.

Malt, B. C., Sloman, S. A., & Gennari, S. P. (2003). Universality and language specificity in object naming. *Journal of memory and language*, *49*(1), 20–42.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological review*, *85*(3), 207.

Medin, D. L., & Smith, E. E. (1981). Strategies and classification learning. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(4), 241.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., ... Kording, K. P. (2019, October). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. Retrieved from http://dx.doi.org/10.1038/s41593-019-0520-2 doi: 10.1038/s41593-019-0520-2

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Saxe, A., Nelli, S., & Summerfield, C. (2020, November). If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, *22*(1), 55–67. Retrieved from http://dx.doi.org/10.1038/s41583-020-00395-8 doi: 10.1038/s41583-020-00395-8

White, A., Malt, B. C., & Storms, G. (2017). Convergence in the bilingual lexicon: A pre-registered replication of previous studies. *Frontiers in psychology*, *7*, 2081.

William, L. (1973). The boundaries of words and their meanings. *New ways of analyzing variation in English*.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., & Finn, C. (2020). *Gradient surgery for multi-task learning.* Retrieved from https://arxiv.org/abs/2001.06782