# Recurrent Models of Number as a Cognitive Technology

**Satchel Grant (grantsrb@stanford.edu)**
Stanford University Department of Psychology,
Building 420, 450 Jane Stanford Way, Stanford, CA 94305

**James McClelland (jlmcc@stanford.edu)**
Stanford University Department of Psychology,
Building 420, 450 Jane Stanford Way, Stanford, CA 94305

## Abstract

The causal relationships between verbal count labels and number concepts are difficult to study in human subjects due to physical and ethical limitations. Some previous work has focused on tribes with limited words for exact number, but these studies are lacking in their ability to determine causal relationships between subjects' verbal and numeric abilities. In this work, we seek to model the interactions between verbal count labels and spatiotemporal numeric task performance using recurrent neural network models in grid worlds analogous to established counting tasks. Using these models, we establish connections and analogies between our models and existing human phenomena. We then explore the effects of different types of language labels on the models' ability to count. We find evidence to suggest that exact verbal count labels do improve models' abilities to count to high numbers in memory based counting tasks, but the labels need to be grounded in the act of spatiotemporally counting. Additionally, we find support for the hypothesis that improving the approximate number system through cultural experience is at least partially responsible for differences in human's counting abilities.

## Introduction

One version of the Whorf-Sapir hypothesis makes the claim that language has a causal influence on the concepts available to a human mind (Whorf, 1956; Sapir, 1929; Pinker, 2007). This version of the theory suggests that without language for a concept, the concept is mentally inaccessible. Although this version of the theory has been disproved in many settings, some experiments still give credence to the hypothesis (Deutscher, 2011). For example, the Russian Blues experiment showed that speakers of languages with different color categorizations will perceive a greater color distance across category boundaries than within (Winawer et al., 2007). Another example comes from the Guugu Yimithirr tribe in Australia, who have no words for egocentric directions. One must use cardinal directions (e.g. North, East, etc.) to properly convey directional information in their language. Members of the tribe showed a bias for cardinal directions when completing various spatial tasks that could be solved equivalently using cardinal or egocentric directions (Levinson, 1997; Majid, Bowerman, Kita, Haun, & Levinson, 2004). Depending on what direction native speakers are facing, they even change the direction of their gestures when describing memories (Deutscher, 2011).

In the domain of counting, a person's lack of precise number words has been shown to be predictive limited performance on a set of tasks involving exact quantities. On these tasks, that are trivial for English speaking adults, the Amazonian tribe known as the Pirahã—who's language lacks exact number words—fail to reliably perform the tasks past the target quantity of 4 (Gordon, 2004). The Pirahã exhibit a constant coefficient of variation (CofV) in their task performance, indicating that they likely switch from an exact to an approximate representation of the target quantity beyond the count of 4 (Crollen, Castronovo, & Seron, 2011; Krueger, 1982, 1984). There are similar findings for an Amazonian indigenous group known as the Mundurukù who have count words for 1 through 5. Members of this group were shown to perform poorly in a number subtraction task. The Mundurukù's number words are similar to the Pirahã in that they are imprecise with respect to the numbers they refer to (Pica, Lemer, & Izard, V., & Dehaene, 2004). Additionally, recent work with a Colombian tribe known as the Tsimanè has demonstrated that the limits of a participants' verbal count list predict an upper bound on their performance in exact number tasks (Pitt, Gibson, & Piantadosi, 2022). Lastly, Miller, Smith, Zhu, and Zhang (1995) have observed developmental differences in the length of children's count lists depending on the systematicity of their language's count list.

In each of these studies, we are left wondering whether the participants' poor performance is caused by the makeup of their verbal count list or caused by a lack of relevant counting experience? It is easy to imagine that a culture that does not place great emphasis on tasks that require exact count words would never develop exact count words. Thus the cause of poor exact number understanding would be cultural rather than lingual. Alternatively, however, the culture could have no idea what exact number tasks are, simply because they have not yet learned a set of count words that would enable them to conceptualize exact number tasks.

Frank, Everett, Fedorenko, and Gibson (2008) devised an experiment to determine if the key to English speakers' success at exact number tasks is their ability to use an exact count list as a cognitive tool to store and retrieve

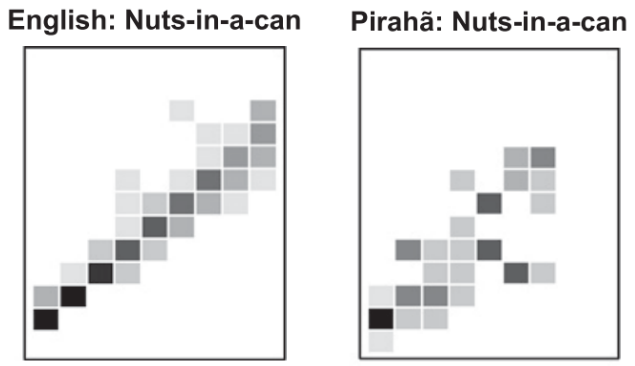


Figure 1: English and Pirahã performance on an memory based counting task. Left Panel: English speakers' performance during verbal interference on numbers 4-12. Right Panel: Pirahã performance without interference on numbers 4-10. Panels taken from (Frank et al., 2011).

numeric information. To test this, they had English speaking participants (who possessed exact number words) perform a verbal interference task while attempting to complete exact quantity tasks. This significantly hindered their performance. In a later study, Frank, Fedorenko, Lai, Saxe, and Gibson (2011) compared a verbal and a visuospatial interference task to control for the possibility that verbal interference merely harmed performance through increased cognitive load. They found that the verbal interference task caused a larger decrease in performance than the comparable visuospatial interference task (Frank et al., 2011). They used this as evidence to show that exact number words are used as a cognitive tool by English speakers, suggesting that if only the Pirahã had words for exact numbers, they would be able to perform the tasks.

The work of Frank, Everett, et al. (2008) and Pitt et al. (2022) seems to suggest that cognitive access to exact number labels is causal in our ability to manipulate exact quantities. This conclusion, however, is not sufficient to account for all findings, nor is it the only interpretation of the existing literature.

In the case of Frank et al. (2011), while the verbal interference does significantly decrease English speakers' performance, the interfered English speaking performance is still visibly better than that of the Pirahã (see Figure 1). Additionally, there is always the possibility that parts of the neural pathways involved with verbal interference task are used separately in exact numeric reasoning. Thus, the English speakers' observed decrease in performance is not necessarily due to a lack of access to verbal number labels. It could be explained by multiple mental processes demanding the same neural resources. Additionally, there have been cases of people who lost their ability to use language due to brain lesions, but maintain their ability to do exact arithmetic (Butterworth, 1999).

Additionally, the finding that the Tsimanè's exact number task performance rarely exceeds their verbal count list is not sufficient to conclude that they are using their verbal count list to perform the task. Nor is the result sufficient to conclude that a verbal count list causally improves performance on tasks involving exact quantity. Many of the Tsimanè's spatiotemporal task performance indicated that they switched to solving the task using approximate rather than exact numbers multiple numeric steps before reaching the end of their verbal count list (Pitt et al., 2022). This suggests that they were not explicitly using their count list to perform the task. Otherwise, they would have likely used the list to its full extent. This suggests the possibility that there is a deeper cognitive mechanism that is jointly granting the subjects' abilities to produce exact number labels as well as perform exact number tasks.

Lastly, it is important to consider what is precisely meant by "access to exact number words." There are examples of children proving their ability to recite a count list but failing at counting tasks within the bounds of their recited count lists Wynn (1997). It would seems likely that some sort of conceptual grounding of the count list is necessary to use it. Is it possible to learn a grounded count list without also developing concepts for exact numbers? How much counting success can we attribute to exact number labels rather than exact number concepts?

In an attempt to address these issues, we explore the effects of different types of number labels in recurrent neural networks trained on exact quantity tasks analogous to those used in (Gordon, 2004), (Frank, Everett, et al., 2008), and (Pitt et al., 2022). Neural network models offer numerous advantages such as the ability to control exact verbal knowledge and the ability to control for exact spatiotemporal (aka cultural) experience during training. Artificial neural networks also allow us to explore causal relationships by controlling the order in which models learn different concepts. We can ask questions such as: how does an agent perform at spatiotemporal exact quantity tasks when it has already learned labels for exact number quantities? Is it important for labels to be grounded in spatiotemporal experience? How quickly does an agent learn verbal labels when it has already learned concepts for exact quantity spatiotemporal tasks? Ultimately, are cultural factors or lingual factors the reason that the Pirahã cannot count?

The contributions of our work are as follows. We first use the models' verbal and spatiotemporal performances to establish the following connections to existing human studies:

1. Our models perform better using exact number labels than using Pirahã quantity words when learning to complete spatiotemporal counting tasks
2. Models trained with systematic exact number labels learn to count to higher target-quantities faster than non-systematic exact number labels
3. All models exhibit a relatively flat CofV suggesting some

reliance on an approximate number system (ANS)

4. Pretraining models with spatiotemporal counting experience speeds up the time it takes them to learn English count labels

We then explore experiments that would otherwise be impossible or unethical to perform on humans to determine the interactions between numeric labels and counting performance. Our results support previous findings that numeric labels can be used as a tool for spatiotemporal counting. We also find, however, evidence suggesting that cultural, grounded experience is more important than a system of count labels for learning to count. And lastly, our results corroborate some form of linguistic determinism, at least during impressionable stages of learning.

## Methods

### Counting Games

We built a grid world game with two variants to mimic the games presented to the Pirahã in Gordon2004. Each game consists of a grid of pixels in which the objects are represented as unique floating point values at their respective grid coordinates. The game objects consist of an agent, target-items, response-items, a response-item dispenser, and a button to end the game. The agent has 5 possible actions: stay, move left, move up, move right, move down, and interact/press.

Each game variant begins with the target-items sequentially appearing in the bottom half of the grid (visibly divided) during which time the agent can move but is unable to interact. The total number of target items is randomly selected from a specified range at the beginning of each episode. Once all target-items have been displayed for an episode, a signal pixel turns on to indicate to the agent that it can now interact with objects in the environment. The agent's goal is to match the number of target-items with response-items by interacting with an item dispenser. Each interaction with the dispenser causes a response-item to appear neatly along a row in the upper half of the grid. The agent is able to navigate to the response-items and carry them back to the dispenser to delete them. Once the agent believes the number of response-items matches the number of target-items, they must navigate to the ending button and press press it to end the episode. During the entire episode, the agent is unable to access the lower half of the grid (where the target-items reside).

In the **complete-visibility** variant of the game, the target-items remain visible for the duration of the episode. In the **incomplete-visibility** variant, each target-item is sequentially flashed for a single frame at the beginning of the episode. After its initial flash, each target-item disappears for the rest of the episode.

### Models/Training Details

Each model consisted of a Convolutional Neural Network (CNN) to extract visual features; an LSTM conditional feature extractor for the model to determine what variant of the game is being played and what its objectives are; a language output LSTM that uses the CNN and conditional features as input, and feeds its recurrent output into a Multilayer Perceptron (MLP) for number label prediction; and an action output LSTM that uses the recurrent state of the language LSTM as its input, and feeds its output into a two layer MLP for action label prediction. See Figure 3 for a visual diagram. $L_t$ is the language prediction and $a_t$ is the action prediction. The loss is calculated as the mean of the cross entropy loss for both $L_t$ and $a_t$ on ground truth class labels.

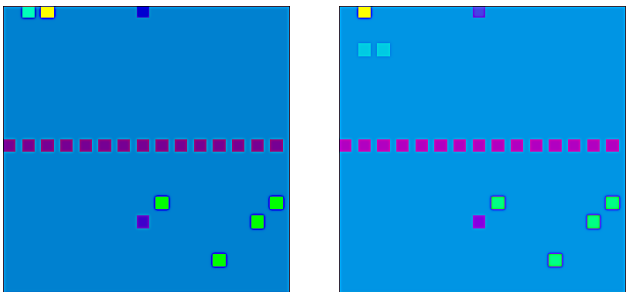

Figure 2: Two images taken three time steps apart from the same episode of the complete-information variant. The bright pixels in the lower half of each image are the target-items. The dark pixel in the center of the lower half of the grid is a visual indication that all target-items have been displayed for the episode. In the upper half of the grid, the bright yellow pixel is the agent, the dark pixel near the middle of the image is the ending button, and the visible pixel in the left image that becomes covered by the agent in the right image is the item dispenser. The two pixels below the agent in the right image are response-items—the result of the agent interacting with the dispenser twice.

The visual grid of the environments was 13 row units by 15 column units. Each visual unit within the grid consists of 3x3 pixels. This made for a visual observation input of size 39x45 pixels for each time step. The CNN consisted of two layers with 18 and 38 channels respectively. The kernel size was 2x2, the padding was 0, and the stride was 1 for each layer. All LSTMs used the same hidden state dimensionality of 256. In the case of the conditional $LSTM_c$, the vector $Z$ in the diagram was simply the hidden state after sequentially processing word embeddings of the conditional statement. The vector $Z$ could have equivalently been a one hot encoding, but we used the conditional statement for flexibility in future work. The input to the language $LSTM_L$ consisted of a concatenation of the flattened output of the CNN (length of 60458) with the recurrent state of the conditional $LSTM_c$, $Z$. The hidden state $h^L$ was then used as the input to the action $LSTM_a$.

Each MLP consisted of a linear weight matrix (256x768), a dropout layer with 35% chance of drop, a ReLU activation layer, a single trainable scaling parameter $\gamma$ and shifting

parameter β applied to all activations, and a linear weight matrix to produce the output. Each LSTM used layernorm on the hidden state before each pass through the LSTM.

The language **pre-trained** model variants were pretrained to watch an expert play the game and predict counting labels while it played. The models were trained to predict a count label of the cardinality of the set of target-items so far displayed as the target-items initially appeared at the beginning of each episode. The models were then trained to predict the cardinality of the number of response-items each time the cardinality changed. After 20 epochs of verbal pre-training, the pre-trained models were trained the same way as the joint-trained models. The spatiotemporally pre-trained model variants were trained similarly to the verbal label pre-trained models except that their pre-training was to predict the actions of an expert on the game without using count labels.

The **joint-trained** model variants were trained to predict counting labels in the same fashion as the pre-trained variants while also predicting the actions of an expert agent at every time step. The joint-training continued for 60 epochs for both the joint-trained models and the pre-trained models.

In grid-world navigation tasks, there can be multiple optimal actions for a given game-state. In these cases, the expert selected its action uniformly from the set of optimal actions.

Models were trained over 8 seeds. 3 seeds each used an RMSprop optimizer with a learning rate of 5e-5. The remaining 5 seeds used an Adam optimizer with a learning rate of 1e-4 except for the No-Language and Inequality variants which continued to use the RMSprop training scheme (this choice was made because the No-Language and Inequality models failed to reliably train using the Adam optimizer). All training sessions sampled fresh data from the games for every epoch. We used batch sizes of 128 over the course of 1000 time steps making 128,000 data samples per epoch. 64 samples in each batch consisted of the complete-visibility variant while the other 64 samples consisted of the incomplete-visibility variant. The hidden state of the LSTMs was reset to a vector of zeros at the beginning of each episode. Over the course of an epoch, data was processed in temporal order in sequences of 7 time steps. The next loop started at $t+1$ as opposed to $t+7$. All models were trained using weight-decay with a value of 1e-3 and a dropout probability of 0.35 between the MLP dense layers of the network. All models were trained on Stanford's CCN cluster using a single NVIDIA Titan Xp GPU.

### Language Labels

We explored different types of language labels and their effects on training speed, final performance, error, generalization to unseen data, and the CofV. It is important to note that the language systems were only trained to predict count labels. They were not trained to predict the end of the episode. They were, however, trained to use the final frame of the episode to predict the count label corresponding to the number of response-items on the grid. There was no implicit training of the models to predict the end of the episode. For all variants, numbers beyond the maximum label are all grouped into the maximum training label when calculating validation performance.

The different label types are:

1. **No Language**: this model variant was always trained without verbal labels in the joint-trained scheme. This model type does not have a verbal label pre-training variant.

2. **English**: a single, unique label for each number in the training distribution (typically 0-11).

3. **Pirahã**: a set of 4 possible labels corresponding to the tribe's distribution of words for increasing set size (Frank, Everett, et al., 2008). The labels for each episode were selected stochastically at the beginning of each epoch according to the statistics recorded in (Frank, Everett, et al., 2008). See Figure 4 for exact label probabilities. Quantities beyond those listed in Figure 4 use the probability associated with a quantity of 10 in Figure 4.

4. **Inequality**: a set of 2 labels—"fewer" or "equal"—corresponding to a comparison of the number of response-items on the grid relative to the number of target-items. The language model does not train on the target-item quantities as there is no sensible comparison during the initial target-item display animations.

5. **Base 4**: a set of 5 labels in a variable length sequence for each numeric prediction. The first 4 available labels correspond to the numerals required for a base 4 numeric system—"0", "1", "2", and "3". The 5th label is a [STOP] token for the model to indicate the end of the sequence. i.e. the number 9 would have a ground truth sequence of **2**, **1**, **[STOP]** whereas the number 3 would be **3**, **[STOP]**.

6. **Random**: a single, unique label for each number in the training distribution uniformly sampled for each time step at the beginning of each epoch.

7. **Duplicates**: a set of two unique labels for each number in the training distribution. Essentially the same as the English labels, but each label has a duplicate. Each of the two correct labels has equal probability of selection, sampled at the beginning of each epoch.

## Experiments/Results

Data for each individual model was collected on freshly sampled data across ten episodes for each target quantity in each game variant. Unless otherwise stated, we trained each model variant across 8 seeds. Each "final model" was the best performing model selected from any training epoch based on validation spatiotemporal response performance.

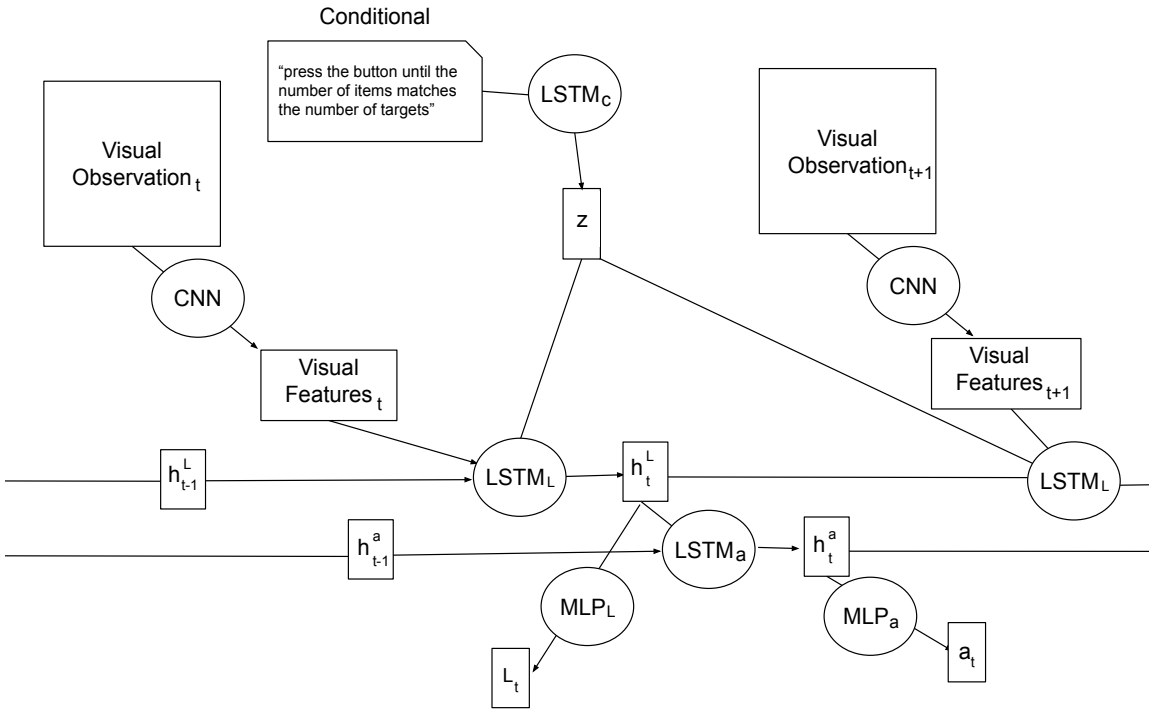


Figure 3: Diagram of the model

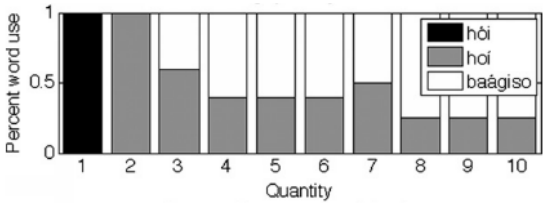


Figure 4: The Pirahã count word distribution. Labels are selected at the beginning of each epoch according to the proportion of word use from this figure. We include a label for a zero quantity with 100% probability of selection. Figure taken from (Frank, Everett, et al., 2008)

Table 1: Final spatiotemporal validation performance of each model variant on target quantities 1-11.

| Label Type | Pre-Trained | Joint-Trained |
|---|---|---|
| No Lang | | $94.7 \pm 0.7$ |
| Inequality | $89.7 \pm 0.9$ | $90.5 \pm 0.9$ |
| **English** | $98.2 \pm 0.3$ | **99.0** $\pm 0.2$ |
| Piraha | $87 \pm 1.0$ | $89 \pm 1.0$ |
| **Base4** | $98.6 \pm 0.2$ | **98.8** $\pm 0.3$ |
| Duplicates | $93 \pm 1.0$ | $93.6 \pm 0.3$ |
| Random | $82 \pm 2.0$ | $79 \pm 2.0$ |

Table 2: A comparison of final joint spatiotemporal performances that include and exclude the held out target quantity, 12. The column **Avg Over 1-11** is the same as the **Joint-Trained** column in Table 1.

| Label Type | Avg Over 1-11 | Avg Over 1-12 |
|---|---|---|
| No Lang | $94.7 \pm 0.7$ | $92.4 \pm 1.0$ |
| Inequality | $90.5 \pm 0.9$ | $87.1 \pm 1.3$ |
| English | $99.0 \pm 0.2$ | $90.6 \pm 1.9$ |
| Pirahã | $89 \pm 1.0$ | $85.0 \pm 1.5$ |
| Base4 | $98.8 \pm 0.3$ | $92.3 \pm 1.7$ |
| Duplicates | $93.6 \pm 0.3$ | $85.8 \pm 3.5$ |
| Random | $79 \pm 2.0$ | $77.9 \pm 2.5$ |

The models' performance on the held out target quantity 12, which was out of the training distribution, was included when selecting the final models. Unless otherwise stated, the reported figures use Joint-Trained model variants and the target quantity 12 is included in the calculated statistics. Error bars standard error measurements calculated over all trials from all seeds pooled together.

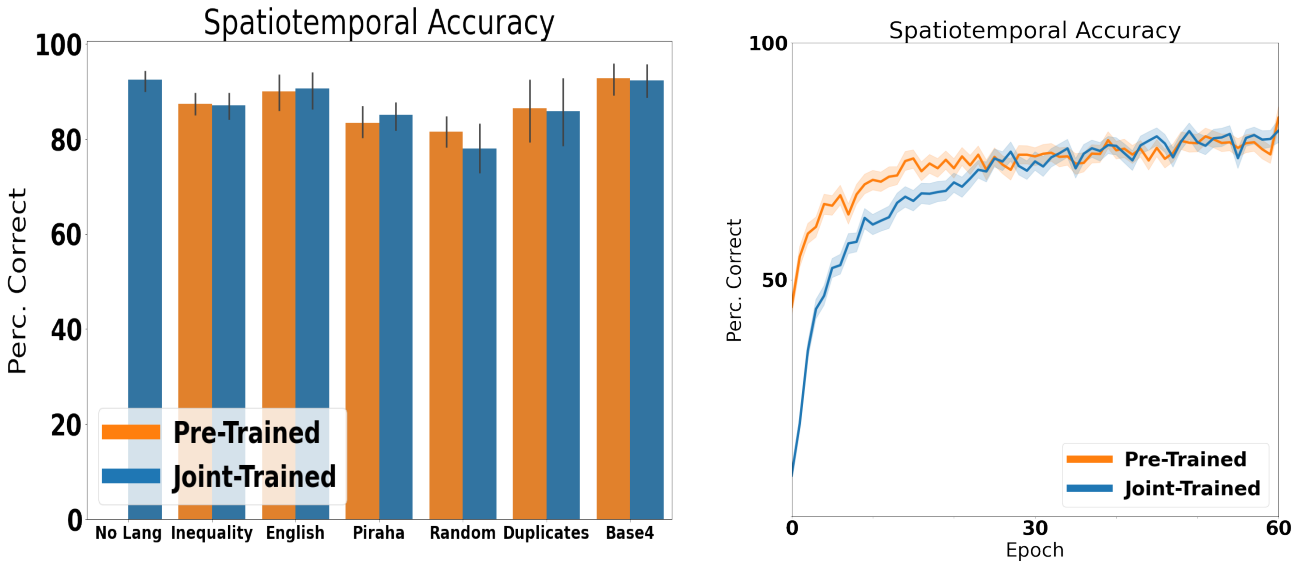


Figure 5: Left Panel: Final spatiotemporal performance of the label variants averaged over 10 validation episodes for each epoch. Blue shows performance of models that were jointly trained to predict count labels and perform spatiotemporal counting tasks. Orange shows performance of models that were pre-trained to predict verbal labels and subsequently trained the same way as the joint-training. Right Panel: all model types averaged together separated by their pre-training condition (excluding the No-Language model, although including the No-Language model did not change the findings). This shows that the pre-training causes an early boost in validation performance but eventually the joint-trained models catch up.

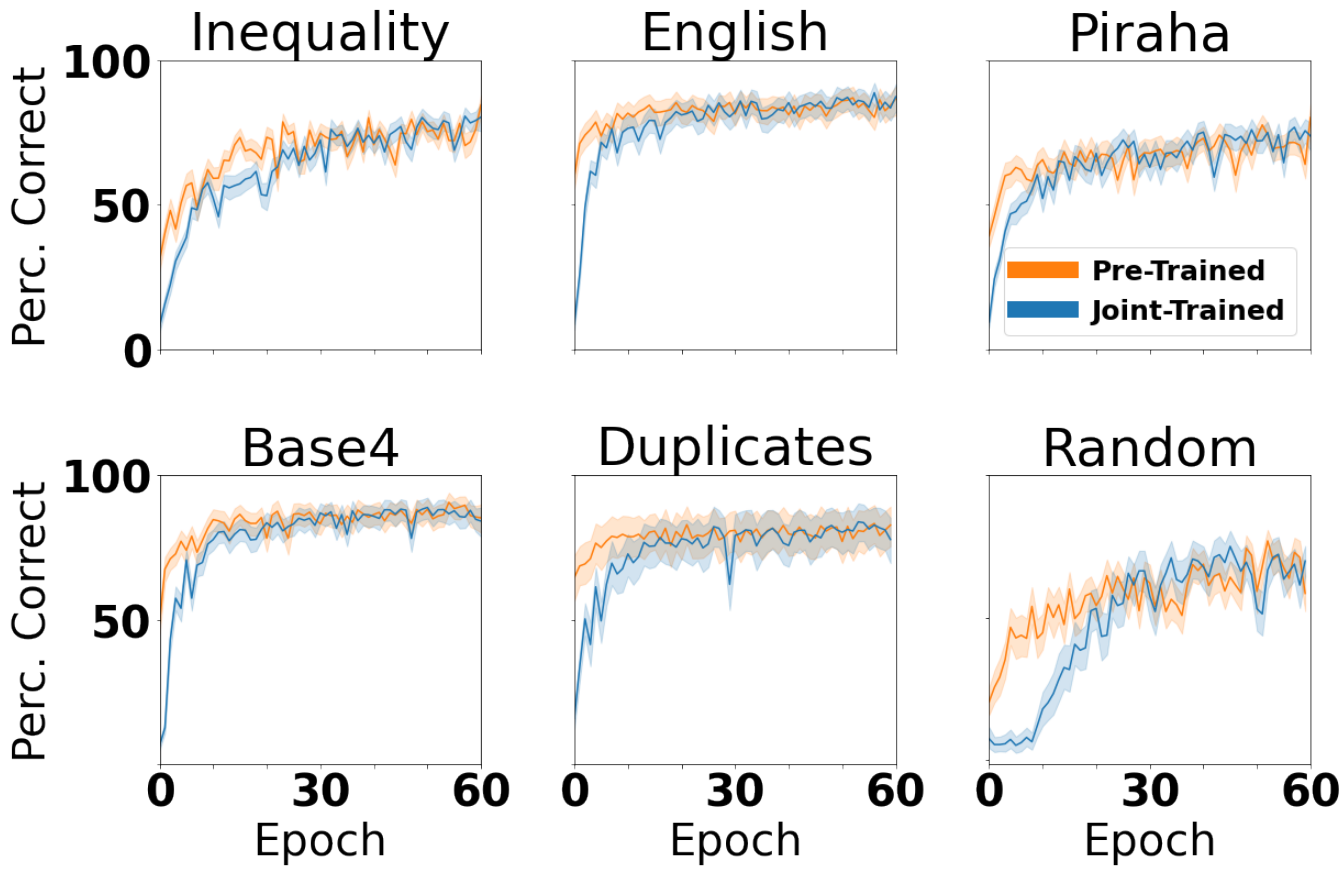


Figure 6: Performance curves over the training period faceted by count label type.

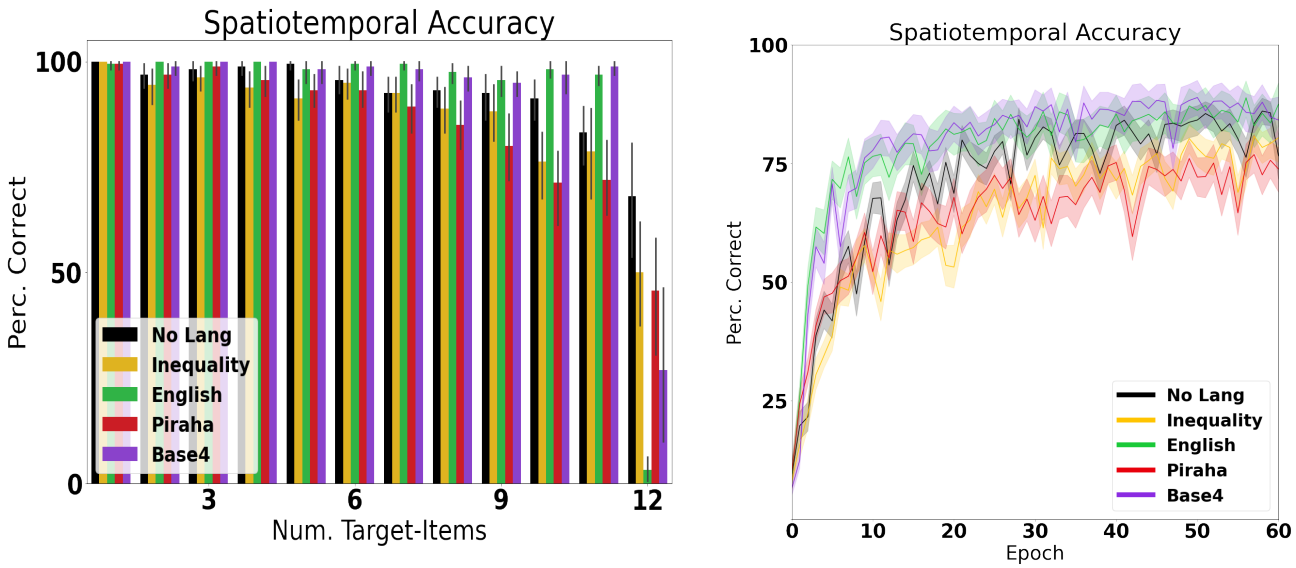


Figure 7: Left Panel: Final performance of each of the joint-trained model variants averaged over 10 episodes for each target quantity. A correct response is defined by the same number of target-items as response-items on the grid at the end of the episode. Right Panel: spatiotemporal counting performance over the course of training.

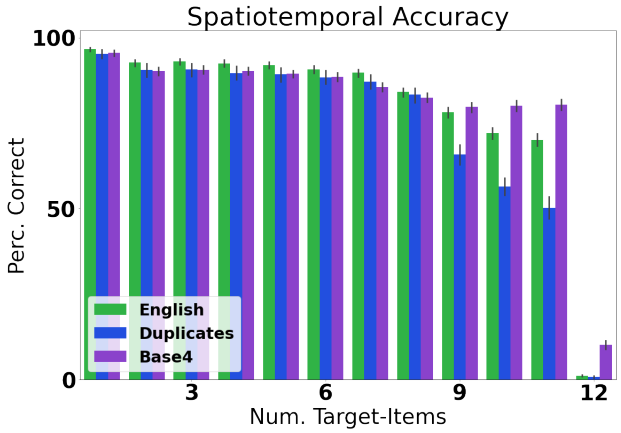


Figure 8: Joint-Trained exact count label model variants validation performance averaged over all epochs in the training period.

## Connections to Human Studies

**Exact Number Labels** From Figure 7 and Table 1 we can see that both of the exact count label variants (English and Base 4) produce the best final performance. This result captures studies of the Pirahã where Gordon (2004) and Frank, Everett, et al. (2008) found the Pirahã to be worse at counting than English speakers. From a computational perspective, this is to be expected considering the large amount of overlap between visual features required for exact count labels and spatiotemporal response selection (Mu, Liang, & Goodman, 2019).

We can see that the English and Base 4 variants are about equal in final performance with final accuracies of 99.0% and 98.8% respectively. Looking at Figures 8 and 17, however, we can see that there is a noticeable performance difference over the course of training. The performance difference is most clear at the largest target quantities 10 and 11 where the Base 4 models clearly perform better than the English models. We can also see in Figures 7 and 13 that the Base 4 models generalize better to the out-of-distribution target quantities.

These results capture findings in developmental literature comparing children who speak Chinese vs English (Miller et al., 1995; Dehaene, 1999) or Welsh vs English (Dowker & Roberts, 2015). Both the Welsh and Chinese number systems are explicitly systematic. In Chinese (Mandarin), they list the quantity of each of the tens places for numbers greater than ten. For example, the literal translation for the numbers 11 and 23 in Chinese are ”one ten one” and ”two ten three” respectively. Miller et al. (1995) found that children who grew up learning to count in Chinese learned to count to 40 as much as a full year before comparable English speaking children. ? (?) found that Welsh children perform better on non-verbal numberline estimation tasks. Both these cases, however, appear to have diminishing effects as the children get older (Nuerk, Cipora, Domahs, & Haman, 2020). Mark and Dowker (2015) found that Chinese speaking children were better than English speakers at number comparison tasks in 1$^{st}$ and 2$^{nd}$ grade, but were not significantly different in 3$^{rd}$ and 4$^{th}$ grade. Dowker and Li (2019) found that by the age of 7, Chinese speaking children were not significantly more accurate than English speaking children at number comparison tasks. Our results match these findings nicely.

From Figure 9 we can see that the English and Base 4

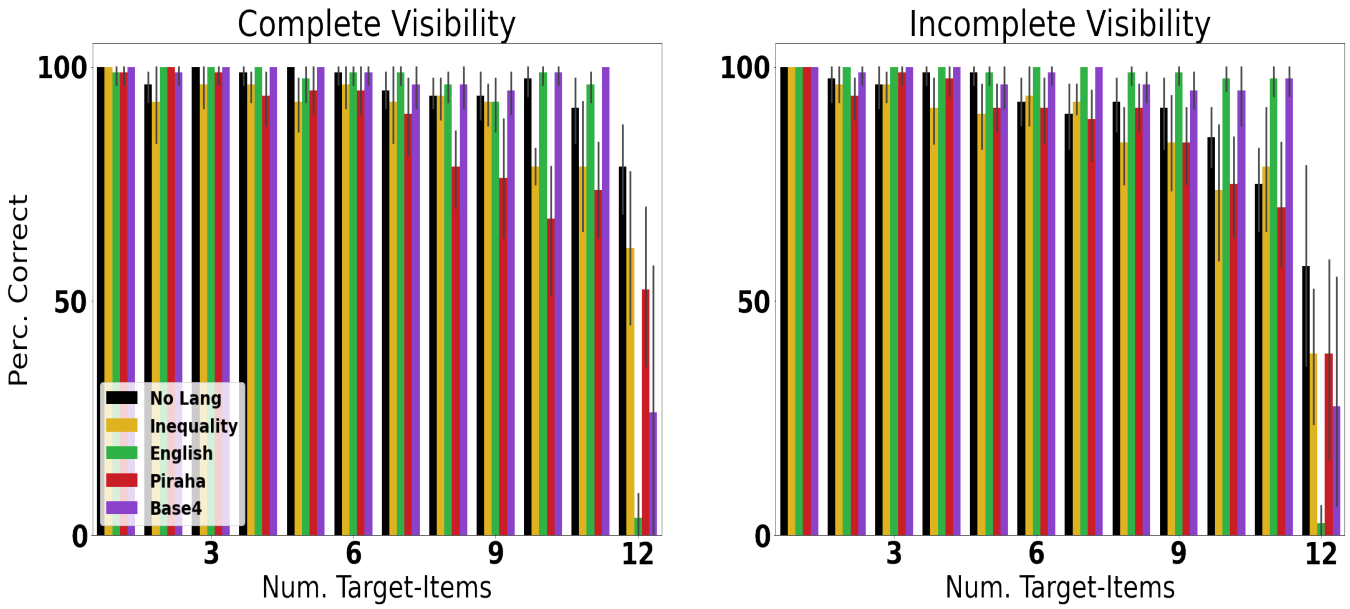


Figure 9: Comparison of performance for the complete- and incomplete-visibility game variants. The incomplete-visibility variant sequentially flashes each target-item for one time-step at the beginning of the episode whereas the complete-visibility variant leaves the target-items visible on the screen for the entire episode.

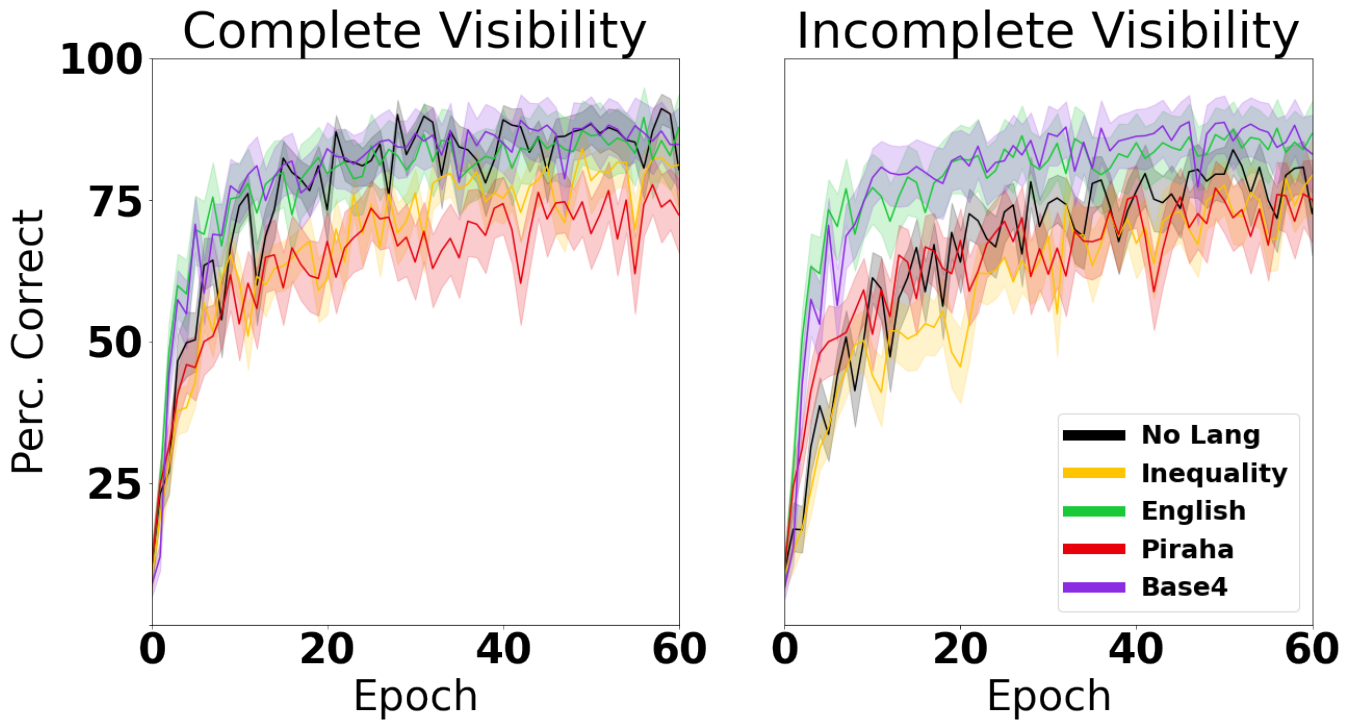


Figure 10: Left Panel: Model performance of the complete visibility task progressing over training. Right Panel: Model performance of the incomplete visibility task progressing over training.

models have even stronger performance relative to the other model types in the incomplete-visibility game variant. This game variant not only requires the agent to count the target items, but also requires the agent to remember the target quantity until the end of the episode. These findings suggest that the precise number labels are somehow improving the network's ability to store information over time compared to the inexact label types. This corroborates the findings of Frank, Fedorenko, and Gibson (2008) in which they showed that the Pirahã and English speakers during verbal interference had better performance in complete-visibility tasks than they did in incomplete-visibility tasks. As (Frank, Fedorenko, & Gibson, 2008) point out, this is likely because an agent can solve the task by pairing items together rather than performing exact counting.

**Coefficient of Variation** As an additional connection to the work of Gordon (2004) and (Frank, Fedorenko, & Gibson, 2008), we examined a well known phenomenon in which the standard deviation of humans' quantity estimation tends to scale with the size of the quantity being estimated (Dehaene, Izard, Spelke, & Pica, 2008). The coefficient of variation (CofV) is a measure of how the standard deviation of a measure scales with the mean of the measure. It is calculated as follows:

$$CofV = \frac{\sigma}{\mu} \tag{1}$$

Where $\sigma$ is the standard deviation and $\mu$ is the mean of the measure. In this work, we use the standard deviation of the model's responses and divide it by the ground truth response.

We can see from Figure 11 that all the models exhibit a relatively flat CofV (with the potential exception of the English models on the held out value 12). From this result, we can establish a connection to the large body of work showing that humans exhibit a linear scaling of error with target quantity when performing number estimation (Gordon, 2004; Crollen et al., 2011; Frank, Everett, et al., 2008; Le Corre & Carey, 2007). This connection suggests, however, that the models are approximating the numeric quantities rather than explicitly enumerating them.

The models' flat CVs can potentially be explained by the work of Cheyette and Piantadosi (2020), who's theoretical framework suggests that a constant CofV results from minimizing the expected error over a distribution that is Zipfian. Zipfian distributions are defined as $p(k) \propto \frac{1}{k^j}$ where $k$ is the target quantity and $j$ is greater than 0.

In our counting games the target quantities encountered during training do not strictly follow a Zipfian distribution. They do, however, follow a distribution that places greater weight on smaller numbers. Each time the game displays $k$ target items, the model necessarily needs to count all numbers less than $k$ as well. If we take $X_q \in \{0,1\}$ to be a random variable representing whether the model interacted with a number $q$ in the counting game in a given episode, then we can calculate $P(X_k = 1)$ to be equal to 1 minus the probability of selecting a target quantity $k$ less than the value $q$:

$$P(X_q = 1) = 1 - P(X_q = 0) = 1 - \sum_{k=1}^{q-1} \frac{1}{N} = \frac{N+1-q}{N} \tag{2}$$

where $N$ is the maximum possible target quantity over the course of training and $q \leq N$. This means that the model interacts with the number 1 on every training episode, and it interacts with the largest number $N$ on $\frac{1}{N}$ of the episodes (in almost all trainings in this work, $N$ was equal to 11).

We did try training a set of Pirahã joint-trained models on environments in which target quantities, $k$, were sampled from Zipfian distributions where the exponent $j$ ranged from the default value (0) to 2. Note that when sampling $q$ using a Zipfian distribution in the counting games, the resulting frequency of interaction with a value $q$ are not strictly Zipfian. The resulting frequencies are proportional to the sum of all target quantity frequencies greater than or equal to the value $q$:

$$P(X_q = 1) \propto \sum_{k=q}^{N} \frac{1}{k^j} \tag{3}$$

Given that smaller values of $k$ in equation 3 have greater weight and smaller values of $q$ include these larger terms in the sum, this means that equation 3 will place greater weight on smaller values of $q$ than a traditional Zipfian distribution would. All this is to say, when $j > 0$, larger numbers have proportionally smaller frequencies of occurrence than they would in a traditional Zipfian distribution.

From Figure 12 we can see a comparison of CofVs for different values of $j$, and from Figure 18 we can see the corresponding performance curves. The results suggest that models trained with smaller values of $j$ have better performance. Furthermore, models trained with larger values of $j$ potentially have more sloped CofVs.

We suspect that these performance differences are caused by the relatively small batch sizes used in stochastic gradient descent while training the models. In general, a model trained with back-propagation needs batch statistics that are representative of the statistics of the data. If one class has a low probability of occurring, then there could be multiple batches in a row that don't include data from the low frequency class. This can severely harm the model's ability to find an optimal solution if it steps too far, too many times, without information about the low frequency class. Increased batch sizes may alleviate this issue by helping the gradient "see" an accurate representation of the data distribution (Johnson & Khoshgoftaar, 2019).

**Generalization to Unseen Quantities** We explored training models while excluding the number 6 from the possible target quantities. We also held out all verbal label training for the number 6 in episodes where the target

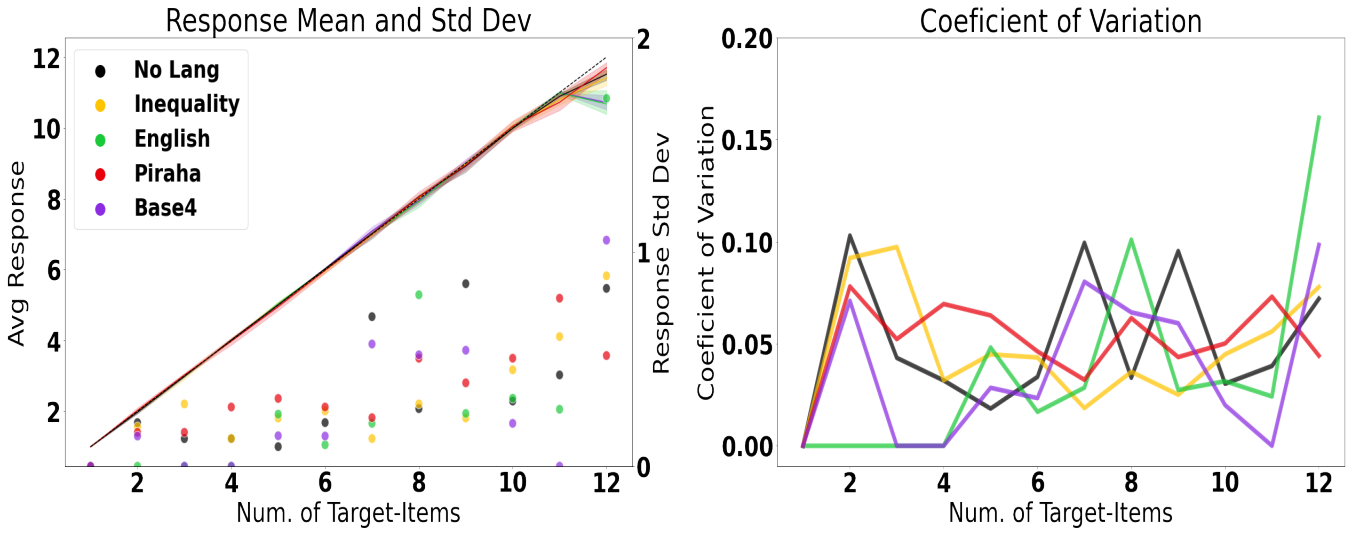


Figure 11: The CofVs of the final joint-trained models for each label type. Left Panel: The lines plotted against the left axis show the average model response (number of response-items on the grid at the end of each episode) plotted as a function of the target quantity. The dots plotted against the right axis show the standard deviation of the responses. Right Panel: The coefficient of variation of the responses.

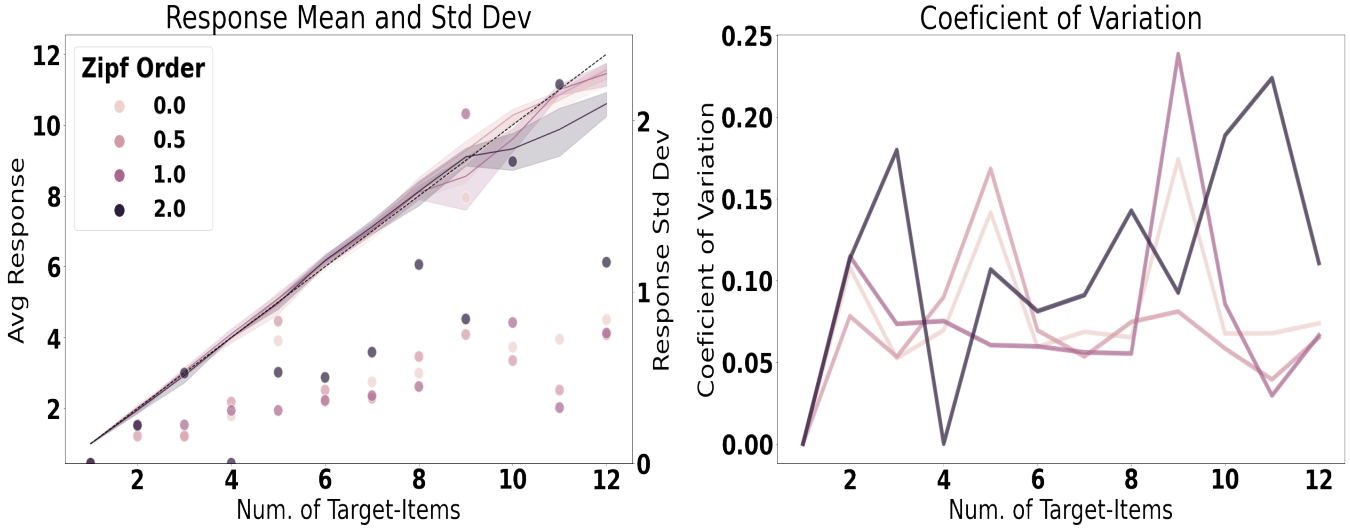


Figure 12: Comparison of trainings with different target quantity sample distributions. All models are Pirahã joint-trained variants with target quantities $n$ sampled with probability $p(n) \propto \frac{1}{n^j}$. $j$ is the value of the hue in the panels. A $j$ of 2 hindered the models' performance.

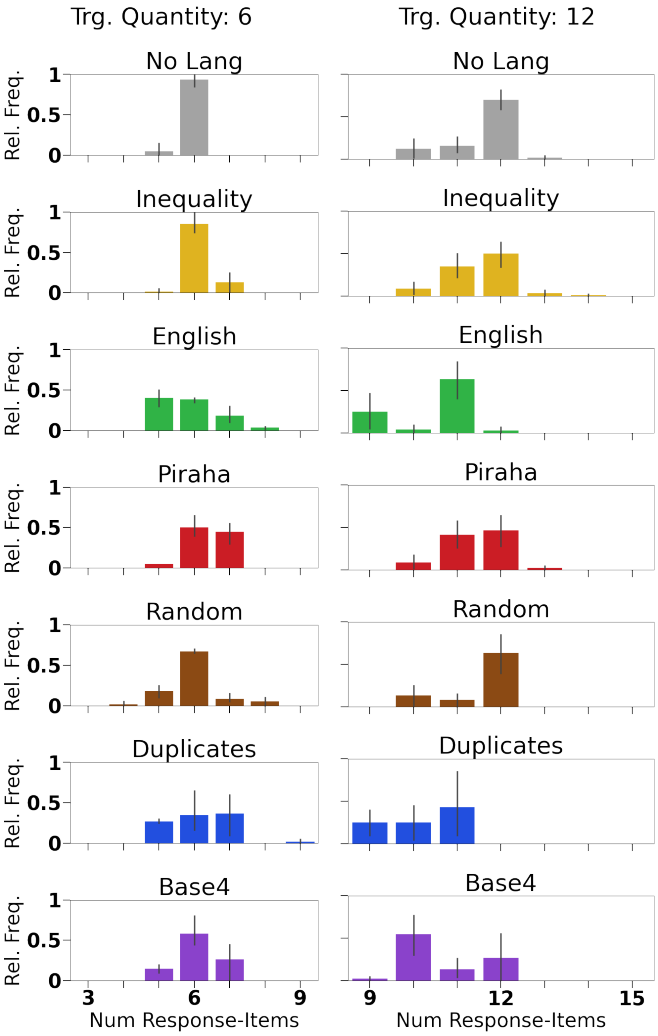


Figure 13: Spatiotemporal response distributions on held-out target quantities 6 and 12. The models used for the left-side panels with the target quantity of 6 were trained such that 6 was never included in their verbal label training, nor was 6 used as the target quantity for any training episodes.

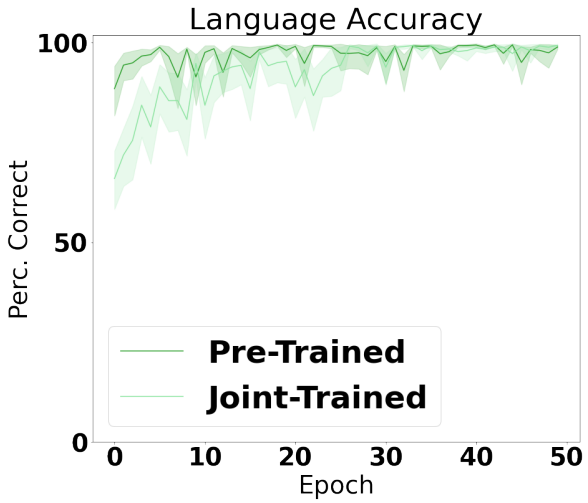


Figure 14: English model variants pretrained on the spatiotemporal counting task learn their language labels in fewer epochs than without pretraining. Dark green represents the language label accuracy for joint-trained English model variants pretrained on the spatiotemporal counting tasks without access to exact count labels. Light green is the language label accuracy of joint-trained English label models without spatiotemporal pre-training.

quantity exceeded 6. From Figure 13 we can see that final performance for the held out quantities was worse for the English and Base 4 model variants than the Inequality, Random, and No-Language models.

Using these results, a parallel can be drawn to recent work showing that the Tsimanè had difficulty completing counting tasks on quantities that extended beyond the limits of their exact verbal count list (Pitt et al., 2022). We can see from Figure 13 that the English, Duplicates, and Base 4 models fail to count beyond the limits of their trained count list. Notably, the Base 4, who's verbal count system implicitly includes 6 and 12, generalizes best out of the three exact label types. This corroborates Pitt et al. (2022)'s findings in that an agent's available count list places a limit on their spatiotemporal counting abilities.

Using this analogy, however, makes the high performance of the inexact count label variants more surprising. We speculate that the count labels act as a sort of scaffolding for the model's potential solutions. More precise count labels constrain the model to learn more precise counting concepts, but the constraints may be so restrictive that the agent overfits to solutions within its accessible count labels. Whereas the more nebulously defined count labels (or no count labels at all) provide less constraint on the solution space, but also allow more general solutions.

**Spatiotemporal Causality** We performed an experiment of pretraining the English and Base 4 models to perform the spatiotemporal counting tasks without language labels and

then observed how this would impact the models' ability to learn their respective numeric labels.

From Figure 14 we can see that the spatiotemporal pre-training improved the models' learning speed. This is congruent with the work of Fang, Zhou, Chen, and McClelland (2018) who found pre-training a model to sequentially touch items in a spatiotemporal setting improved the model's verbal counting abilities. This also captures Alibali and Dirusso (1999)'s findings that children who were better able to touch a sequence of objects were also better at verbally counting them.

From a computational perspective, this result is to be expected. A model that has learned to count in a spatiotemporal sense has also necessarily learned a feature extraction function that extracts some form of a numeric representation from raw pixels. The models' action selection function is likely a more complex function—conditioned on navigation, interaction, and quantity representations—than its verbal label system—which is only conditioned on the quantity representation. As such, a model that is pre-trained to spatiotemporally count is spared the computational cost of learning a target quantity feature extraction function while learning numeric labels, thus speeding up training.

We note that the increased verbal training speed can alternatively be explained by non-numeric visual experience with the environment, rather than a learned spatiotemporal understanding of number concepts. An experiment to control for this would be to pre-train the model on a non-numeric game within the same environment. If this training scheme improved the model's training speed on count labels, it would suggest that the driving causal factor is simply spatiotemporal experience with the environment. We leave this experiment for future work. We do, however, explore this hypothesis in more depth, later in the paper, when determining the causal effects of verbal count labels on spatiotemporal training speeds.

## Model Explorations

Probably the most valuable aspect of this modeling approach is our ability to use the models to explore experiments that would otherwise be impossible or unethical in the real world. In this section we present experiments that may provide insight into universal principles of cognition.

**Number as a Cognitive Technology** The Inequality model variants were trained to verbally label whether the number of response-items were less than or equal to the number of target-items on the grid. This model variant was designed to control for the effects of learning exact numeric labels that are independent of the agent's ability to use a count list. The analogous real-world situation would be one in which a person had learned to equally divide a resource, like berries or nuts, into equal groups without using exact count words.

From Figure 7 we can see that the Inequality and No-Language models both had a final performance that was worse than the English and Base 4 models for the target quantities 6-11. This indicates that exact count labels do improve performance at larger target quantities. Given that the effect is exaggerated in the incomplete-visibility game variant (see Figure 9), the performance gap is likely caused by the English and Base 4 models' improved ability to encode and/or remember the target quantity at the beginning of an episode.

This interpretation is strengthened by noting that the Inequality variant shows improved generalization to held-out quantities compared to the English and Base 4 models (see Figure 13). This is consistent with the idea that the models rely on their labeling systems to perform the task. Another potential interpretation is that the held out numbers, 6 and 12, are contained within the bounds of the Inequality language labels. This potentially makes it easier for the Inequality variant to generalize to unseen numbers. Similarly, for the Base 4 model, the held out quantities are implicitly within the boundaries of its available count labels and system. Perhaps this is why we see slightly better generalization from the Base 4 variant than the English variant.

We expect, however, that part of the relatively poor overall performance by the Inequality models can be explained by lack of a training label signal at the initial stages of each episode. To remind the reader, we did not train the Inequality models to produce labels during the initial reveal of the target-items. We believe it would be insightful to compare the Inequality models against other label variants trained without label signals during the initial target-item display. We leave this comparison to future work.

**Counting Without Words** We were initially surprised by the high performance of the No-Language models (see Table 2 and Figure 7). The high performance goes against intuitions that exact language is necessary for counting to high numbers. There are, however, real world examples of counting without words. Some chimpanzees, for example, have been trained to recall the spatial location and order of a sequence of numbers up to nine (Matsuzawa, 1985, 2009). In these cases the chimps work with unique numeric symbols making it unclear if they are actually counting the sequence or merely differentiating and ordering a set of symbols. A potentially more relevant finding may be the abilities of individuals with brain lesions who have lost their ability to do lingual tasks, but still posses numeric faculties such as the ability to do arithmetic (Butterworth, 1999).

The fact that the No-Language models perform so well relative to the performance of the English and Base 4 models suggests that the human mind can, to some degree, learn to count without explicit labels. This interpretation would help to explain the remaining difference in performance between the Pirahã and English speakers found in Frank, Everett, et al. (2008) (see Figure 1).

There are ways to mentally encode the cardinality of a set without using verbal count words. For example, we invite the

readers to try encoding a quantity using rythmic humming, or by visually clustering items into sets of 3 or 4. We are not suggesting that this is how the models are performing the task, but these are possibilities of how the task could be solved without verbal labels.

The No-Language models, however, have relatively poor starting performance as seen in Figure 7 and have a worse performance gap in the incomplete-visibility game variants. We also informally note that the training of the No-Language (and Inequality) models was more difficult than the other model variants. For some seeds, models trained with these label variants required a more precise learning rate and optimizer for reasonable performance. We do not present a thorough analysis of this informal observation, but we do wish to draw the reader's attention to the idea that this could be another advantage of language in the human mind. Perhaps grounded language can provide stability during training for a more robust learning process.

We believe, however, the most likely interpretation of the high performance of the No-Language models can be explained by the plasticity of approximate number systems (ANS). Park and Brannon (2013) showed that humans who practiced number estimation developed better a more accurate ANS and improved their mathematic abilities relative to controls. We speculate that the No-Language (and non-exact numeric language) models are using and improving an ANS throughout training that becomes surprisingly accurate by the end of training. Under this hypothesis, it is unclear what the limits of an ANS are. With infinite neurons, training time, and data, can an ANS match the performance of exact number systems? Here, we informally note that in the preliminary stages of the project, training with larger models seemed to reduce the observed final differences between the different numeric label training variants. We leave a more thorough exploration of this to future work.

**Ungrounded Label Pretraining** To explore the causal effects of exact count words on the models' spatiotemporal counting performance, we wanted to see how an ungrounded count list would impact models' spatiotemporal counting performance. To address this, we created a pre-training scheme in which the model received zeros as its visual input and was trained to repeatedly output its count list for the quantities 1-10. The count loop was reset at the beginning of every episode.

We found that this pre-training scheme did not improve the models' training on the spatiotemporal counting tasks (see Figure 15). This result captures human data from the work of Wynn (1997) in which they found that some children, who proved their ability to recite a count list, would still fail at relevant spatiotemporal counting tasks. Our findings corroborate the idea that access to a count list is not sufficient to perform counting tasks. It seems that a grounded understanding of number is more important than a system of symbols to spatiotemporally count.

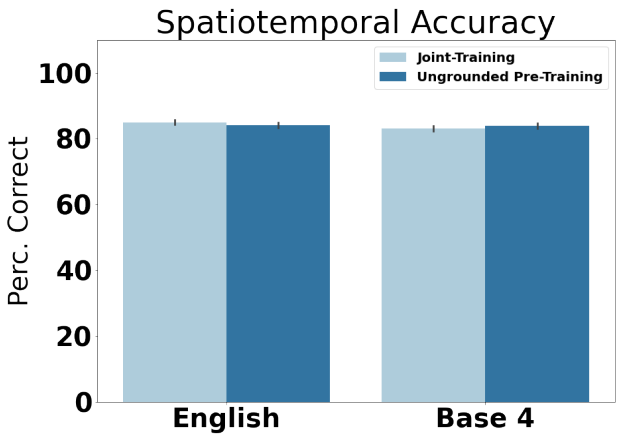


Figure 15: Comparison of ungrounded verbal label pretraining. Light Blue: models pre-trained to sequentially produce verbal count labels using blank visual inputs. The count labels looped back to their smallest value once the model reached the end of the count list. Dark Blue: models without pre-training.

From a computational perspective, we speculate that this can be explained by the difference in potential ways to learn a system of symbols vs grounding a system of symbols in spatiotemporal experience. The way in which the model's neurons learn the verbal count system will have an impact on its ability to use the verbal count system in its spatiotemporal experience. We suspect that the number of potential solutions to learning how to repeat a system of symbols is far greater than the number of ways to learn a system that can be used in connection to the spatiotemporal experience. Without the constraint of connecting the ungrounded count list to a spatiotemporal experience, the chances that it learns a useful count list are low. To put it more simply, we speculate that learning a sequence of labels is a far easier task than connecting those labels to spatiotemporal experience.

**Grounded Label Pretraining** To further explore the ideas mentioned in the ungrounded label pretraining, we wanted to explore how grounded labels would impact the models' performance. We found that pre-training models with grounded count labels caused quicker learning in most cases but did not seem to cause a better final accuracy. We can see from Figures 5 and 6 that verbal pre-training results in an earlier learning curve for all model variants. Eventually the joint-trained models catch up for an equivalent final accuracy.

We included the Random label variant as a control to determine if it was the grounded count labels that caused the performance boost, or if the causal factor was general, spatiotemporal experience in the environment. From Figure 6 we see that the Random label models did, indeed, have a shifted performance curve, which suggests that the numeric labels were not the causal factor in early performance improvements. This, combined with the results from the

ungrounded label pre-training, implies that learning a verbal count list is less important than shaping grounded perceptions of the environment when learning to spatiotemporally count.

Extrapolating this result to the human literature, this implies that cultural experience, rather than exact count words, plays an important role in participants' ability to perform the exact counting tasks. We can't, however, fully ignore the improvements that come from trainings that use exact count labels. The combination of these two results—exact count labels cause better final performance at larger quantities (within count labels), and pre-training with exact count labels does not causally improve training—forces us to reconsider the role of exact count words in counting.

It appears that exact count words help the models learn to count, but only when the count words are learned in tandem with spatiotemporal counting. We speculate that the count words serve as a sort of scaffolding for the models' learned concepts. It is as though the verbal labels are a constraint that helps the model cognitively organize a concepual solution to the problem. In computational terms, the verbal labels can potentially be thought of as a regularizer on the optimization process (Mu et al., 2019). In psychological terms, the labels seem to encourage more precise concepts and thus more precise solutions.

**Duplicate Labels** To better understand how the probabilistic label selection affected the performance of the Pirahã model variants, we included the Duplicates variant that was trained using two equivalent count words for each possible target quantity in the training. During training, the ground truth count labels were selected with 50% probability from the two possible choices.

From Table 1 and Figures 19 and 20 we can see that the Duplicates models have slightly worse performance (93.6) than the English models (99.0). Given that the Duplicates performance curve in Figure 19 flattens out towards the end of training, we do not believe the performance decrease is caused by insufficient training time or data. We believe it is more likely caused by the noisy gradient signal stemming from the probabilistic label selection.

The labeling scheme in the Duplicates variants causes correct responses to be labeled as incorrect in 50% of the training samples. An improved training scheme to avoid the noisy gradient would be to use a multi-label loss function. This could additionally be used for the Pirahã variants. This would, likely, be more analogous to the real world in which there is no training punishment for using one of many equivalent verbal labels. Although, we admit, this is an assumption about the way the brain learns.

## Known Limitations

In this section we address some limitations that would not be obvious from the figures and data presented up to this point. We noticed in the preliminary stages of the project that hyperparameter choices (such as learning rate, dropout, and l2 regularization) could change relative performances of each of the model variants. Informally, we note that the exact count label variants generally performed best, but the final results were subject to change depending on the chosen parameters. In future work, we may compare models across a greater range of hyperparameter settings.

We also found that the Inequality and No-Language variants performed best in environment settings where the agent needed to manually carry the response-items from the dispenser to the grid (as opposed to simply interacting with the dispenser for a single frame in the game). It was important to use game variants in which the counting labels could progress frame-by-frame rather than being spaced out over multiple time-steps. We believe attention mechanisms could alleviate this limitation.

## Discussion

In this work, we have found that precise number labels can improve a model's ability to perform spatiotemporal counting tasks, and we have found that a system of exact count labels can improve an agent's speed of learning and ability to generalize. We demonstrated that all of the models have a constant CofV which implies they are somewhat relying on an approximate number system. This also connects our models to many findings in humans (Gordon, 2004; Crollen et al., 2011; Frank, Everett, et al., 2008; Frank, Fedorenko, & Gibson, 2008; Le Corre & Carey, 2007). We showed that models with precise count labels have poor generalization to quantities outside of the numbers they were trained on. We connected this result to recent findings in the Tsimanè tribe. We demonstrated that models pre-trained on the spatiotemporal counting tasks learned English count labels faster than models without pre-training. And lastly, we showed that exact number labels improved performance more than numeric labels that require a semantic understanding of exact quantity without explicitly enumerating each numeric value (the Inequality variant).

Extrapolating beyond existing human literature, we demonstrated that agents pre-trained to learn language labels improves their learning speeds on counting tasks. The causality, however, appears to be mainly from grounded experience with the environment rather than the count labels themselves, as demonstrated by the ungrounded English label pre-training and the grounded Random label pre-trainings. Furthermore, our No-Language agents demonstrated a surprisingly good ability to count without any count labels. When viewed in conjunction with the relatively high generalization performance of the Inequality models, and the poor generalization performance of the exact label variants, it appears that, in some scenarios, count labels can actively hinder an agents' ability to generalize their counting.

The poor performance of our Inequality models seems to suggest that differences in counting abilities between cultures is mainly caused by the counting labels rather than cultural differences. In this case, our data serves to strengthen the

findings of existing literature (Frank, Everett, et al., 2008; Pitt et al., 2022). The relatively strong performance of the No-Language models, however, enhances the picture, suggesting that there is a non-negligible cultural/experiential component to a participant's ability to count. This is not necessarily counter to the idea of using number as a cognitive technology, but another explaining factor for the observed differences in the Pirahã and verbally inhibited English humans' ability to count (Frank, Everett, et al., 2008). This interpretation is also congruent with findings that people can improve their ANS with training, and an improved ANS causes improved mathematical performance (Park & Brannon, 2013).

Our models suggest that the reason the Tsimanè often reach the limits of their spatiotemporal counting abilities before they reach the limits of their verbal count list is because knowledge of verbal count words does not guarantee a connection of the count words to the participants' spatiotemporal counting experience. This conclusion is also congruent with the work of Wynn (1997), who found that children who proved their ability to recite a count list could still fail at relevant spatiotemporal counting tasks.

At this point, we wish to provide a perspective that differentiates between "language", "concepts", and "labels". We define a concept as a compact, discrete, generalization of real or imagined experience. It is not guaranteed that a concept is accurate or useful, it is simply a compact, discrete representation. In terms of a human or model, a concept is a collection of neuronal activations. Concepts can be attached to unique symbols (verbal or otherwise) which we have been referring to as verbal labels in this work. Languages are systems that connect labels to one another. Languages offer the flexibility to precisely communicate and navigate relations between concepts. They grant the ability to combine grounded concepts to compose unlabeled concepts in terms of labeled concepts.

It is possible for an agent to produce labels from a language without connecting these labels to any of its existing concepts. We have been referring to this as ungrounded language throughout this work. Additionally, there can exist concepts that have not been attached to verbal labels (the vast majority of concepts fall into this category). And lastly, there are systems that can implicitly offer connections between concepts with labels and concepts without labels.

The results of our models suggest that concepts are somewhat malleable to the influences of imposed labels. This is demonstrated by the fact that each of the training label variants has different performances. Our results also suggest a system of labels can be used as a tool for greater cognitive achievement. This was demonstrated by the relatively high performance of the exact count models at larger target quantities in the incomplete-visibility, memory based tasks. And lastly, our results suggest that agents prefer to think within the bounds of their available verbal labels. This was demonstrated by the poor performance of the exact label variants generalizing to unseen quantities.

The findings from our models indicate that some form of linguistic determinism does, in fact, exist. As noted by Deutscher (2011), the nature of linguistic determinism likely depends on what is forced by the language labels rather than what is encouraged by the language labels. From a computational perspective, language might be understood as scaffolding for the learning process—strongly encouraging conceptual solutions to be consistent with the accessible system of labels—but not completely restrictive in an agents ability to have concepts outside of their available language. The more important aspect for cognitive performance is the existence of concepts than the existence of labels or languages. Language simply enhances the agents abilities, providing an additional tool for conceptualizing and performing tasks.

A problem with our existing experiments is that we do not know to what degree the exact label models are using their count labels to perform the counting tasks vs using the count labels to constrain and improve their conceptual learning of the tasks. To address this, we could train the models with separate conditional tasks for their verbal and their spatiotemporal systems. In this setup, we could train the models to learn verbal tasks other than counting. This would be a more analogous setup to Frank, Everett, et al. (2008) and Frank et al. (2011) in which participants' lingual systems were capable of precise number words, but were occupied with a task other than counting during the spatiotemporal counting task.

Using a setup in which the agents' verbal and spatiotemporal systems are capable of being decoupled, we would be able to better explore how the active use of count labels—rather than the optimization constraints that come with count labels—would affect an agent's ability to spatiotemporally count. Furthermore, we could introduce non-counting tasks in the spatiotemporal environment to further determine if the performance shift in verbal label learning that we observed in Figure 14 was caused by the model learning to spatiotemporally count or merely due to visual experience with the environment.

We would also like to explore how attention might enable the models to sync their count labels with their actions across greater time horizons. This would alleviate any information timing issues that can arise in recurrent models. Expanding our model architectures to use Transformers or other forms of attention might make them more generalizable to environment dynamics (Vaswani et al., 2017).

## Conclusion

Using deep neural networks trained to spatiotemporally count, we explored how numeric labels can affect the model's performance and ability to learn counting tasks. We found that using this training setup, the model exhibits many similarities to existing Cognitive literature. Furthermore, we used the models to explore what might happen in experiments

that would otherwise be impossible or unethical to create in the real world.

Our results were mixed. We found that exact number words are causal in improving an agent's ability to perform counting tasks, but the causality was only observed when the verbal counting systems were trained in tandem with the spatiotemporal systems. And we found that agents with no language are potentially better at counting than previously thought, suggesting a cultural/experiential influence is at play in preexisting work.

Given the promise of our modeling approach we hope to continue this research with new training variants, model architectures, and tasks to better explore the effects of language on mathematical cognition.

## Acknowledgments

Many thanks to the Stanford Department of Psychology and Stanford's PDP Lab for funding this research. Also thansk to the members of the PDP Lab for their support of this research at lab meetings. And thanks to Mike Frank for acting as a sounding board, suggesting relevant related work, and his feedback during my FYP discussion meeting.

## References

Alibali, M. W., & Dirusso, A. A. (1999). The function of gesture in learning to count: More than keeping track. *Cognitive Development*, *14*(1), 37–56. doi: 10.1016/S0885-2014(99)80017-3

Butterworth, B. (1999). *The mathematical brain / brian butterworth* [Book]. Macmillan London.

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature Human Behaviour*, *4*(12), 1265–1272. Retrieved from http://dx.doi.org/10.1038/s41562-020-00946-0 doi: 10.1038/s41562-020-00946-0

Crollen, V., Castronovo, J., & Seron, X. (2011). Under- and over-estimation: a bi-directional mapping process between symbolic and non-symbolic representations of number? *Experimental Psychology*, *58*, 39-49. doi: 10.1027/1618-3169/a000064

Dehaene, S. (1999). The number sense: How the mind creates mathematics. *British Journal of Educational Studies*, *47*(2), 201–203.

Dehaene, S., Izard, V., Spelke, E., & Pica, P. (2008). Log or linear? distinct intuitions of the number scale in western and amazonian indigene cultures. *Science*, *320*(5880), 1217-1220. doi: 10.1126/science.1156540

Deutscher, G. (2011). *Through the language glass: Why the world looks different in other languages*. Arrow Books.

Dowker, A., & Li, A. M. (2019). English and Chinese children's performance on numerical tasks. *Frontiers in Psychology*, *9*(FEB), 1–11. doi: 10.3389/fpsyg.2018.02731

Dowker, A., & Roberts, M. (2015). Does the transparency of the counting system affect children's numerical abilities? *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00945

Fang, M., Zhou, Z., Chen, S., & McClelland, J. L. (2018). Can a recurrent neural network learn to count things? *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, 360–365.

Frank, M. C., Everett, D. L., Fedorenko, E., & Gibson, E. (2008). Number as a cognitive technology: Evidence from pirahã language and cognition. *Cognition*, *108*(3), 819-824. doi: https://doi.org/10.1016/j.cognition.2008.04.007

Frank, M. C., Fedorenko, E., & Gibson, E. (2008). Language as a cognitive technology: English-speakers match like pirahã when you don't let them count..

Frank, M. C., Fedorenko, E., Lai, P., Saxe, R., & Gibson, E. (2011). Verbal interference suppresses exact numerical representation : online lexical encoding as an account of cross-linguistic differences in cognition..

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*(5695), 496–499. doi: 10.1126/science.1094492

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1). Retrieved from https://doi.org/10.1186/s40537-019-0192-5 doi: 10.1186/s40537-019-0192-5

Krueger, L. E. (1982). Single judgments of numerosity. *Perception & Psychophysics*, *31*(2), 175–182. doi: 10.3758/BF03206218

Krueger, L. E. (1984). Perceived numerosity: A comparison of magnitude production, magnitude estimation, and discrimination judgments. *Perception & Psychophysics*, *35*(6), 536–542. doi: 10.3758/BF03205949

Le Corre, M., & Carey, S. (2007). One, two, three, four, nothing more: An investigation of the conceptual sources of the verbal counting principles. *Cognition*, *105*(2), 395–438. doi: 10.1016/j.cognition.2006.10.005

Levinson, S. (1997, 06). Language and cognition: The cognitive consequences of spatial description in guugu yimithirr. *Journal of Linguistic Anthropology*, *7*, 98 - 131. doi: 10.1525/jlin.1997.7.1.98

Majid, A., Bowerman, M., Kita, S., Haun, D. B., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108–114. doi: 10.1016/j.tics.2004.01.003

Mark, W., & Dowker, A. (2015). Linguistic influence on mathematical development is specific rather than pervasive: revisiting the chinese number advantage in chinese and english children. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00203

Matsuzawa, T. (1985). Use of Numbers by a Chimpanzee. , 7–9.

Matsuzawa, T. (2009). Symbolic representation of number in chimpanzees. *Current Opinion in Neurobiology*, *19*(1), 92–98. doi: 10.1016/j.conb.2009.04.007

Miller, K. F., Smith, C. M., Zhu, J., & Zhang, H. (1995). Preschool origins of cross-national differences in mathematical competence: The role of number-naming systems. *Psychological Science*, *6*(1), 56–60. Retrieved 2022-06-10, from http://www.jstor.org/stable/40062877

Mu, J., Liang, P., & Goodman, N. D. (2019). Shaping visual representations with language for few-shot classification. *CoRR*, *abs/1911.02683*. Retrieved from http://arxiv.org/abs/1911.02683

Nuerk, H., Cipora, K., Domahs, F., & Haman, M. (2020). *On the development of space-number relations: Linguistic and cognitive determinants, influences, and associations*. Frontiers Media SA.

Park, J., & Brannon, E. M. (2013). Training the approximate number system improves math proficiency. *Psychological Science*, *24*(10), 2013-2019. Retrieved from https://doi.org/10.1177/0956797613482944 (PMID: 23921769) doi: 10.1177/0956797613482944

Pica, P., Lemer, C., & Izard, V., & Dehaene, S. (2004). Exact and Approximate Arithmetic in an Amazonian Indigene Group Author ( s ): Pierre Pica , Cathy Lemer , Véronique Izard and Stanislas Dehaene. , *306*(5695), 499–503.

Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. Viking.

Pitt, B., Gibson, E., & Piantadosi, S. T. (2022, feb). Exact Number Concepts Are Limited to the Verbal Count Range. *Psychological Science*, 095679762110345. doi: 10.1177/09567976211034502

Sapir, E. (1929). The status of linguistics as a science. *Language*, 207–214.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Whorf, B. L. (1956). *Language, thought, and reality: Selected writings of benjamin lee whorf*. MIT press.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the National Academy of Sciences*, *104*(19), 7780-7785. doi: 10.1073/pnas.0701644104

Wynn, K. (1997). Competence models of numerical development. *Cognitive Development*, *12*(3), 333–339. doi: 10.1016/S0885-2014(97)90005-8

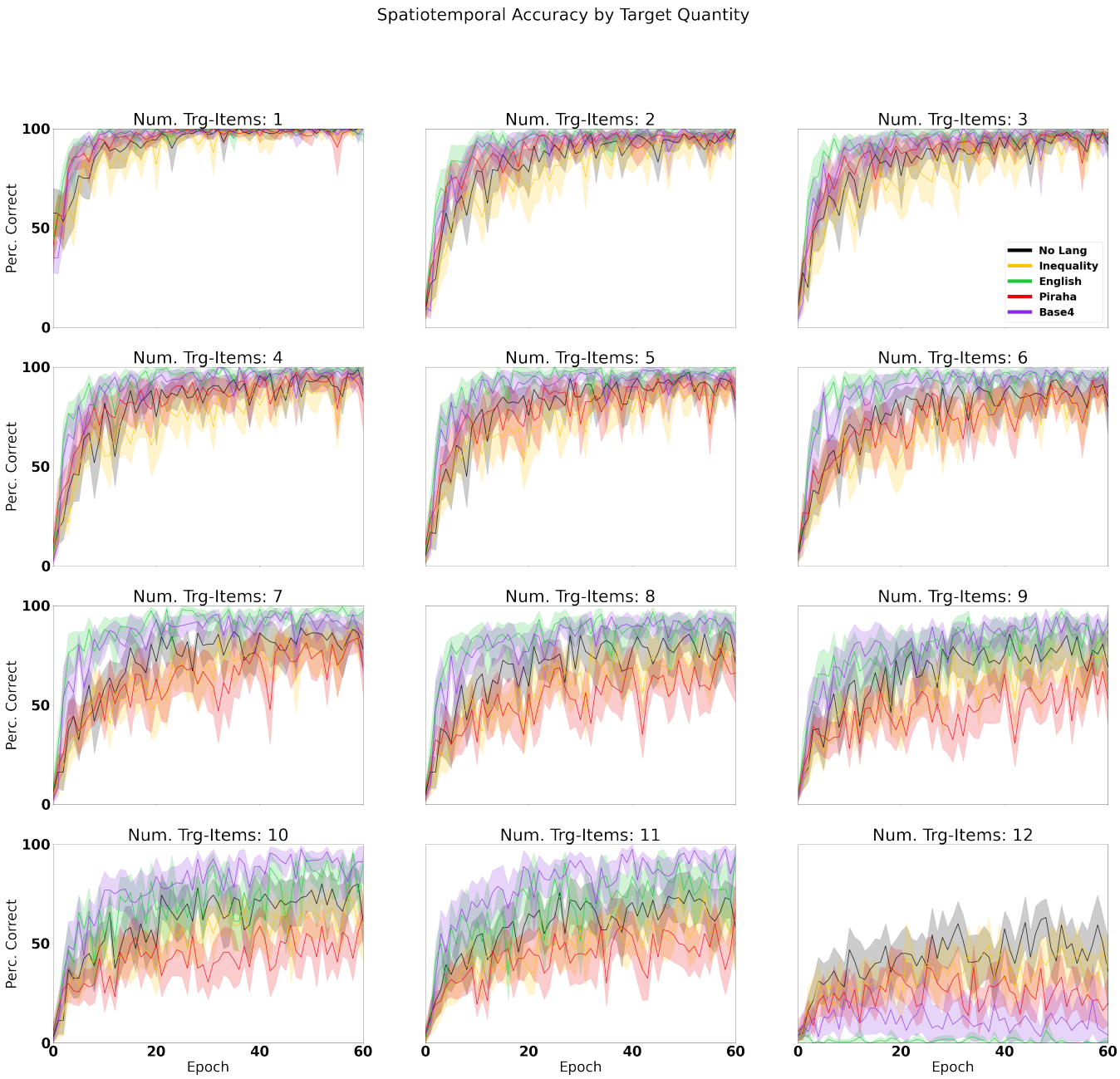


Figure 16: Spatiotemporal performance curves across the course of training faceted by target quantity.

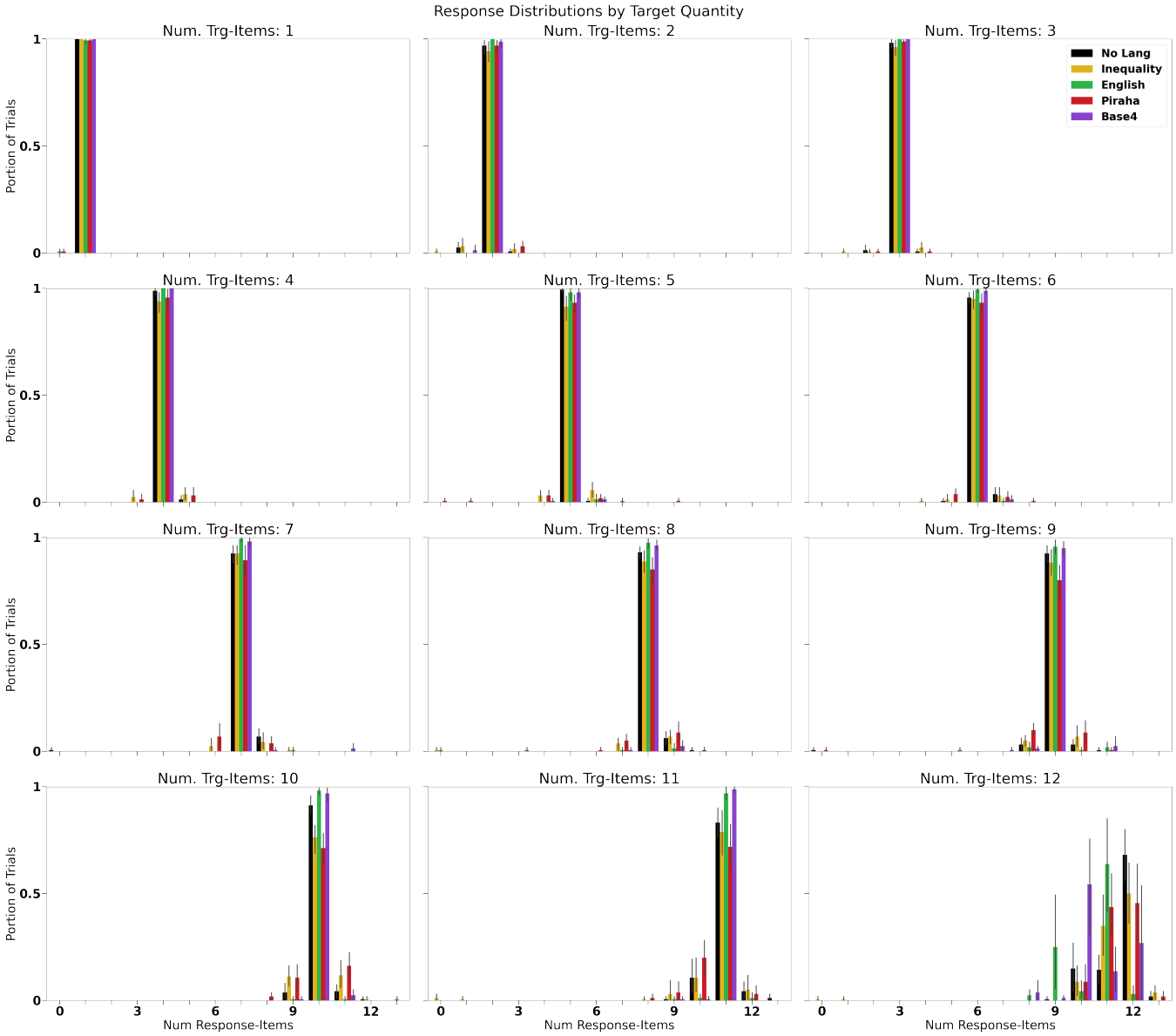


Figure 17: Final model response distributions for different target quantities.

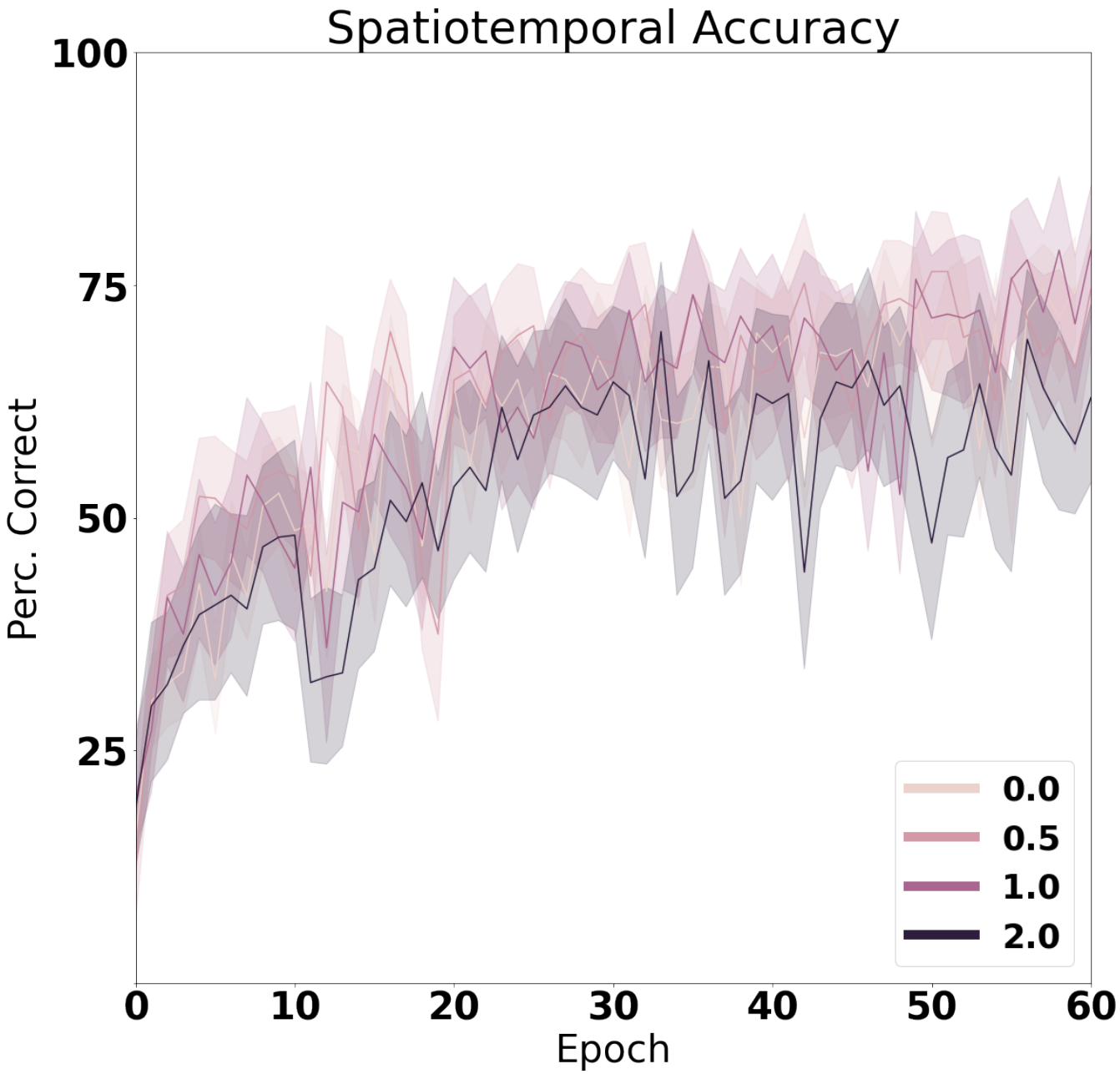


Figure 18: Comparison of trainings on Pirahã joint-trained models with target quantities $k$ sampled during training with probability $p(k) \propto \frac{1}{k^j}$. The hue in the panel denotes the value $j$ used during training.

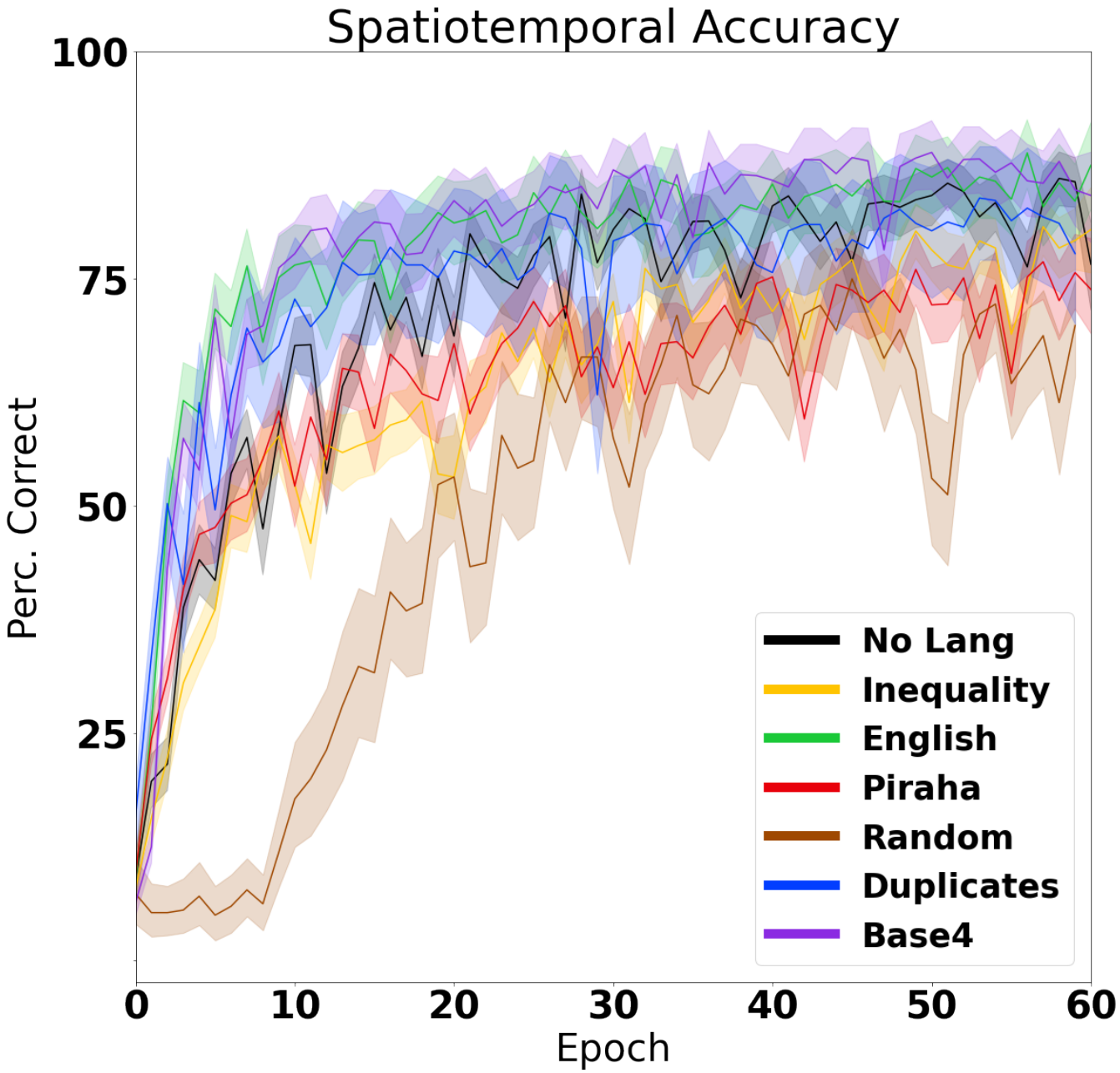


Figure 19: Model performances over the course of training. The performance is calculated as the average number of episodes in which the number of response-items matched the number of target-items at the end of the episode.

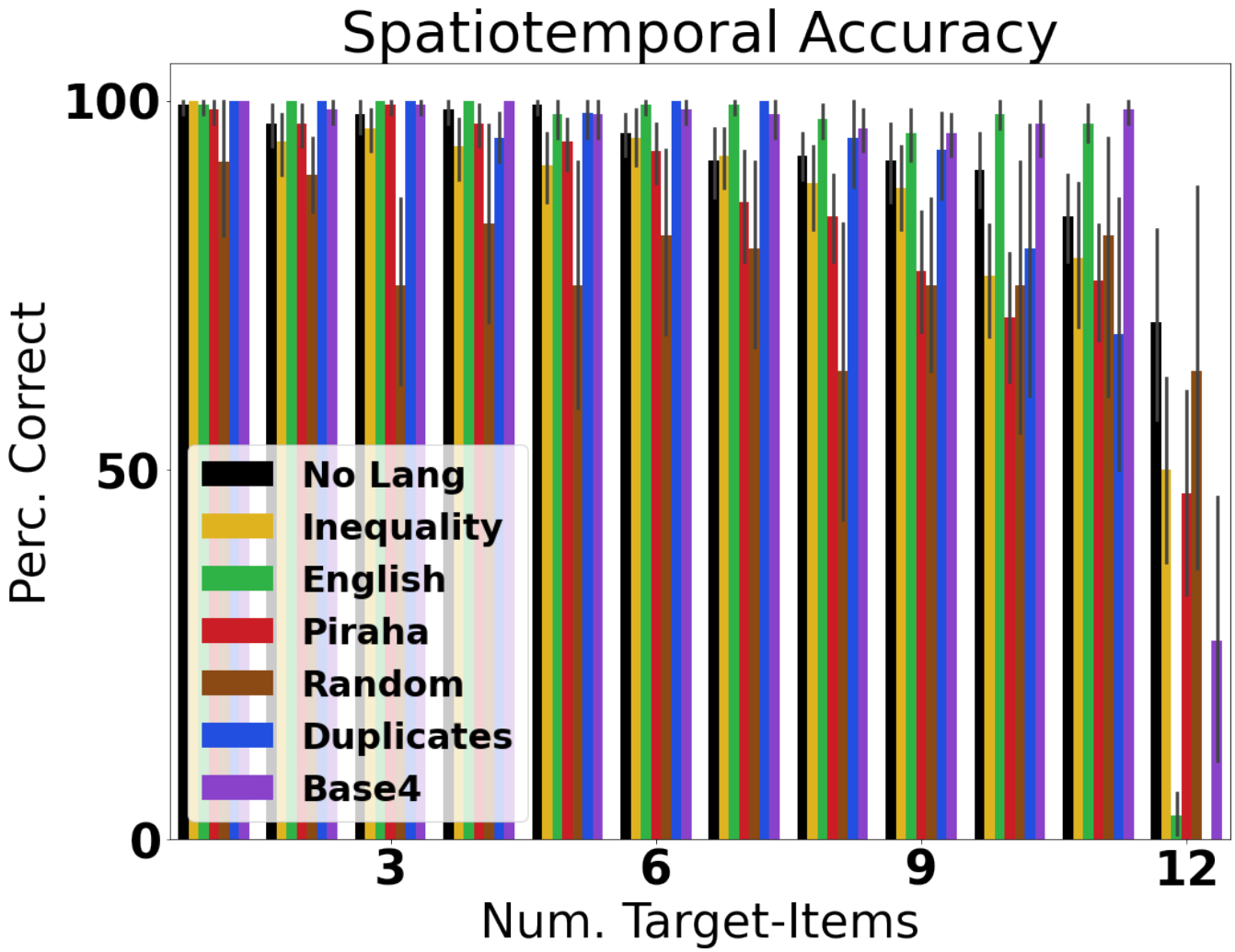


Figure 20: Final model performances. The performance is calculated as the average number of episodes in which the number of response-items matched the number of target-items at the end of the episode.