Graphical Abstract

Bidirectional Influences of Grounded Quantification and Language in Acquiring Numerical Cognitive Abilities

Satchel Grant, James L. McClelland

Highlights

Bidirectional Influences of Grounded Quantification and Language in Acquiring Numerical Cognitive Abilities

Satchel Grant, James L. McClelland

- It is possible for recurrent neural network models to learn to perform exact numeric matching tasks in the absence of learned number words.
- Networks with limited experience with or without number words exhibit the more approximate behavior seen in such tasks in adults members of a culture without number words and young children in numerate cultures.
- Aspects of human behavior and neural representations of number emerge in the models purely from learning to perform the numeric matching tasks.
- Number words causally decrease the amount of experience needed for models to learn to perform exact matching tasks, and learning to perform such tasks causally decreases the experience needed to learn to use number words correctly.

Bidirectional Influences of Grounded Quantification and Language in Acquiring Numerical Cognitive Abilities

Satchel Grant[a], James L. McClelland[a]

[a]*Stanford University Dept. of Psychology, Stanford, 94305, CA, USA*

Abstract

We explore the role of language in cognition within the domain of number, revisiting a debate on the role of exact count words in numeric matching tasks. To address these issues, we introduce a virtual environment to simulate exact equivalence tasks like those used to study the numerical abilities of members of the Pirahã tribe, who lack exact number words, in previous works. We use recurrent neural networks to model visuospatially grounded counting behavior with and without the influence of exact number words. We find that it is possible for networks to learn to perform exact numeric matching tasks correctly up to non-trivial quantities with and without the use of exact number words. Importantly, however, networks with limited counting experience with and without language capture the approximate behavior exhibited by adult members of the Pirahã and young children learning to count in cultures with number words. Our networks also exhibit aspects of human numerical cognition purely through learning to solve the tasks: a flat coefficient of variation and a compressed mental number representation. We explore the causal influences of language and actions, showing that number words decrease the amount of experience needed to learn the numeric matching tasks, and learning the task actions reduces experience needed to learn number words. We use these results as a proof of principle for expanding our understanding of number cognition, and we suggest refinement to our collective understanding of the interactions between language and thought.

Keywords: Whorf hypothesis, numerical cognition, language and thought, Pirahã

Which came first, language or thought?

Each version of the Sapir-Whorf hypothesis makes some form of the claim

that a person's language has a causal influence on the way they think (Whorf, 1956; Sapir, 1929; Pinker, 2007). The theory posits that the language we use causally influences our thought patterns and ability to think about the world. Over the past century many experiments have given credence to the idea. For example, native Russian speakers, who have different words for light and dark blue, respond more quickly to blues that they name differently than to blues with the same name, while native English speakers show no such difference for the same colors (Winawer et al., 2007). Another example can be found in the Guugu Yimithirr tribe in Australia, who have no words for egocentric directions (e.g. left or right) – their language only has words for cardinal directions (e.g. North, East, etc.). Members of this tribe tend to use cardinal directions when performing spatial tasks that could be solved equivalently using cardinal or egocentric directions (Levinson, 1997; Majid et al., 2004).

Relevant findings that have been used to argue for the Sapir-Whorf hypothesis have also been found in the domain of number cognition. A tribe in the Amazon, known as the Pirahã, does not have exact number words. They have only 3 number-related words corresponding roughly to "one or very few", "a few", and "more than a few". Adult members of the Pirahã also do not reliably produce exact matches to given quantities in non-verbal numerical matching tasks, in which the participant is asked to reproduce a given target set of items (Gordon, 2004; Frank et al., 2008a; Everett and Madora, 2012a). There are similar findings for another Amazonian indigenous group known as the Mundurukù who only have roughly translated numeric words for 1 through 5. Members of this group were shown to perform poorly in numeric subtraction tasks (Pica et al., 2004). More recent work has centered on a Colombian tribe known as the Tsimanè. The Tsimanè tribe consists of members with different levels of experience with numbers. Pitt et al. (2022) showed that none of the individuals they tested could reliably perform non-verbal exact matching tasks when the number of items exceeded the number of items they could verbally count.

Although these examples are consistent with many variants of the Sapir-Whorf hypothesis, we cannot use them to establish a causal direction between language and thought. Others, notably Gleitman and Papafragou (2005), have suggested that cognitive differences may reflect differences in cultural practice, proposing that "Thought is first, language is its expression". For each case that we find ourselves "crying Whorf," (Casasanto, 2005; Deutscher, 2011) it is possible that the participants' cultural expe-

rience shaped their thoughts, which in turn, is reflected in their language. Indeed, Everett (2005) has argued this case extensively for the absence of numerical ability among the Pirahã.

The ability to use an explicit counting strategy may, however, play an important role in ensuring accurate performance in exact matching tasks. Frank et al. (2008a) devised an experiment to determine if English speakers need their verbal count list to perform such tasks. They tested Western educated, English speaking adults on exact matching tasks while simultaneously giving them a verbal interference task. Frank et al. (2008a) found that the verbal interference task significantly hindered the English speakers' performance. A later study went a step further and showed that verbal interference caused a larger decrease in performance than a comparable spatial interference task (Frank et al., 2011), though performance did suffer in both conditions. Frank et al. (2011) offered their findings as evidence that English speakers use their verbal count list (internally or externally) to perform the numeric tracking required for non-verbal matching tasks.

Does the work of Frank et al. (2008b) show that count words are necessary for performing exact equivalence tasks? There are a few lingering possibilities that prevent a definite answer. The first is that it is possible for verbally related neural pathways to overlap with number related pathways while words are not explicitly used for the numerical tasks. It is possible for this to occur while visuospatial pathways remain mostly separate from the numeric pathways. If this were the case, verbal interference might affect numeric performance more than spatial interference, even though words were not explicitly used for the numeric task. Additionally, if we were to accept that educated English speakers currently need their verbal count list to perform these tasks, this does not preclude the possibility that English speakers could learn to perform the tasks without number words. Furthermore, in Frank et al. (2008a), the verbally inhibited English speakers perform noticeably better at the equivalence tasks than the Pirahã despite the English speakers' verbal interference. This could be explained by count words "leaking through" the verbal inhibition, or this difference may reflect the English speakers' increased non-verbal experience with exact numbers.

Other work suggests that knowledge of an ordered sequence of number words (or *count list*) is not sufficient for performing exact equivalence tasks. The recent work of Pitt et al. (2022) on the numerical abilities of members of the Tsimanè tribe in Columbia provides one case in point. The tribe consists of people with different length count lists. They found that all of

the Tsimanè who could count past 40 could also perform an exact numerical matching task. However, all but one of those with count lists of 20 or less were unable to perform the task to the full extent of their count list, and 3 of the 15 such participants failed on all numbers greater than 3 (Pitt et al., 2022).

It is unclear why these individuals fail at the equivalence tasks. Perhaps they are indeed attempting to use their count list to perform the task, but they are prone to making count list errors or misalignments of count words with items, much as young children are when they attempt to count the number of items in a set (Alibali and Dirusso, 1999; Wynn, 1997). The assessment of counting abilities of the Tsimanè involved not only pure count-list recitation but also counting the number of items in a set as they picked up each successive member of the set from a pile and placed it in separate pile. This raises the possibility that some of the Tsimanè had the ability to recite an ordered sequence of words in coordination with the successive items in a set, without appreciating deeper implications of the exact cardinality of a set. This dissociation is frequently observed early in human's acquisition of numerical abilities (Davidson et al., 2012; Condry and Spelke, 2008; Jara-Ettinger et al., 2017). This demonstrates that possessing count words, and even having the ability to coordinate them with items in a set, is not sufficient to consistently perform exact equivalence tasks.

If we are to assume that count words can help people to track quantity, what is the mechanism that causes count words to help? We note that when performing a counting task, the reader will likely be able to use known lists that are not ordinarily used to denote the cardinality of a set. The alphabet is one such list, but many others could be relied on as well. Humans could encode quantity through aligning items with notes in a melody or words in a song; or by visualizing a sequence of written numerals; or by visualizing themselves walking down a familiar street, associating each successive house with an increment in quantity.

With this intuition in mind, we generalize the notion of count words to any *symbolic counting system* (SCS). We define an SCS as a system of *discrete categorical symbols*, isomorphic to a (possibly finite) set of natural numbers, with a (possibly small) set of functions and relations defined on that set. In this work, we are largely interested in the successor function. This is the function that increments one natural number to the next in the sequence. A simpler, more relevant definition of an SCS is any system of symbols that can be used as a precise, book-keeping technology to track the cardinality of

a set. We note here that a human's internal monologue can act as an SCS if the words—or more inclusively, the symbols—are used in a way that is consistent with the principles that allow position in the sequence to reliably correspond to cardinality (Gelman and Gallistel, 1978).

We suspect that symbols offer several advantages for a neural learning system. Symbols can align specific, highly variable instances of a class into a single, generalizable, low variance instance of the class. One might even define a *symbol* as the most compact, generalized representation of a class. We suspect that symbols should allow the neural system to have more modular learning, allowing it to learn a function to map pixels to quantities concurrently with a function to map quantities to actions. Additionally, the production of language or symbolic representations might act as a constraint on learning. This sort of constraint could act as a form of regularization on the learning process (Mu et al., 2019) within a learning system that, in the absence of an SCS, would not otherwise conform its behavior in accordance with such compact representations. We examine this possibility in the context of exact equivalence tasks in the experiments in this paper.

How might an SCS emerge in the first place? This touches on the debate as to whether cognitive number systems are learned or innate. One perspective, (Gelman and Gallistel, 1978; **?**; **?**) holds that the foundations of an SCS-like system may be innate, while a second perspective holds that an SCS arises from engagement with an externally provided, typically language-based, external counting system, building on foundations provided by an innate approximate number system and a system for tracking small numbers of individuated objects (Feigenson et al., 2004). Our approach acknowledges that evolution provides mechanisms out of which our cognitive abilities grow, while emphasizing the experience-dependence and gradual developmental refinement that underlies both numerosity estimation (Halberda and Feigenson, 2008) and exact counting (Alibali and Dirusso, 1999), as captured in prior work using neural-network based learning systems (Testolin et al., 2020; Fang et al., 2018). Here we extend this approach to demonstrate how environmental pressures and cultural practices might give rise to a experience-dependent acquisition and gradual developmental refinement of the ability to perform exact numerical equivalence tasks, capturing both initially approximate and, with further refinement, behavior we might call *approximately exact*, without requiring, but nevertheless benefiting from, an externally provided SCS. Within this context, our stance with respect to the question of the primacy of language and thought opens up the possibility that the influences of language

and cultural practices may be bi-directional. To illustrate this possibility, we show how numeric language and numeric thought might causally assist one another in becoming more precise and accurate.

Our work also relates to the open question as to how humans manage to learn language at such a young age, with relatively small amounts of verbal experience when compared to contemporary computational models that capture language abilities and sensitivity to linguistic structure without building in explicit symbolic representational systems (e.g., (Brown et al., 2020; **?**)). Frank (2023) explicitly addresses this language data gap presenting a number of potential explanations including the following. One explanation posits that there are pre-existing structures, developed through evolution, that enable children to easily learn specifically human cognitive abilities (Spelke and Kinzler, 2007), including, for the case of language, a system of constraints such as those captured in universal grammar (Chomsky, 1972). A slight variation of this idea is that there are representational, architectural and optimization constraints, including explicit symbolic representation (Fodor and Pylyshyn, 1988), that encourage certain types of rapid learning early in human development (Tenenbaum et al., 2011). Our own approach falls among those that emphasize the rich, multi-modal, non-linguistic data that all humans receive from their sensory experience of the world (Gopnik and Wellman, 2012), and the possibility of exploiting this within cultural practices that provide experience which may ultimately give rise to the core constructs found in the languages used by members of these cultures (Everett, 2005; James M. McClelland, 2003), perhaps including number. While acknowledging a potential role for innate constraints including biases that favor symbol-like representations and or language-specific constraints, we do not employ such constraints in the current work. Instead, we use causal learning experiments to demonstrate how non-linguistic experience can enhance acquisition of the relevant language in the case of number. Specifically, we show that learning exact equivalence through a visuospatially grounded exact matching tasks can reduce the amount of language data required to learn a language-based symbolic counting system. We use this result as a proof of principle for the cultural learning hypothesis.

To address the aforementioned issues, we explore the abilities of Recurrent Neural Networks (RNNs) to learn exact equivalence tasks with and without the use of an externally provided SCS. More specifically, we train Long Short-Term Memory RNNs (LSTMs) (Hochreiter and Schmidhuber, 1997) to perform embodied equivalence tasks analogous to those used on the Pirahã

(Gordon, 2004; Frank et al., 2008a). We teach the models to map observations of pixels to sequences of discrete actions that constitute performance of an exact numerical matching task. We then explore how an externally provided SCS in the form of a system of count words can influence the models' abilities to learn and perform exact matching tasks and *visa-versa*.

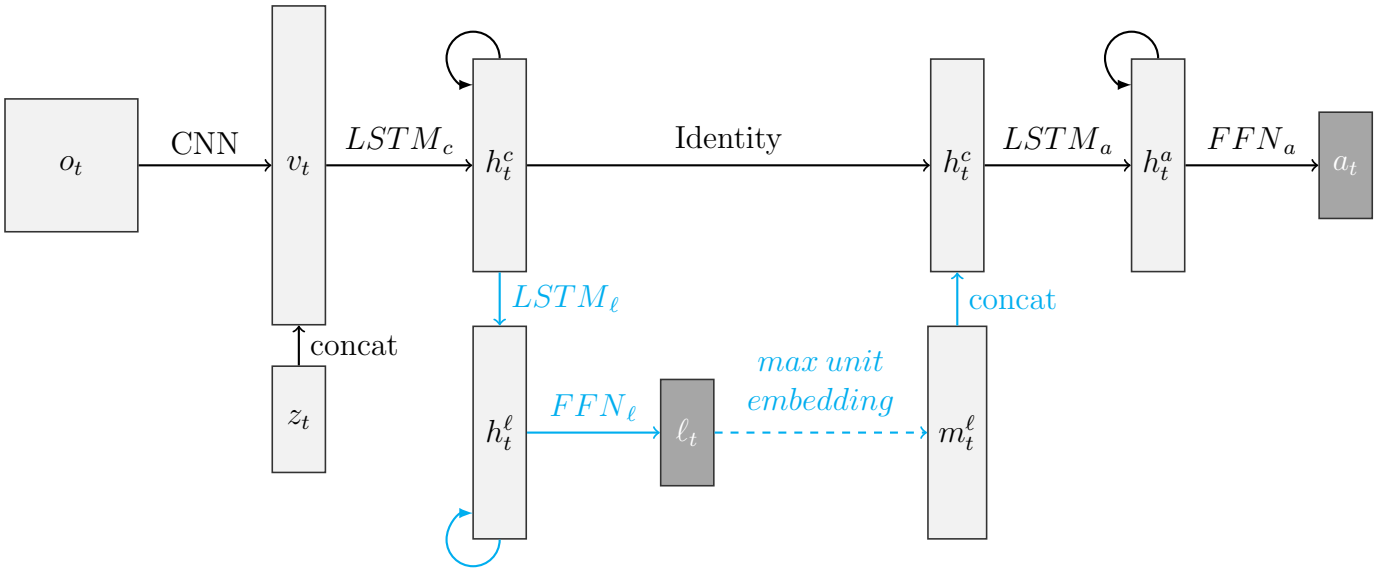

Figure 1: Diagram of the model architecture. Arrows denote functional operations whereas rectangles denote vectors of model activations except for o_t which is a matrix of floating point values. z_t is the conditional task embedding. The action pathway consists of the connection between h_t^c and h_t^a whereas the language pathway, colored light blue, follows the alternative branch stemming downward from h_t^c and connects again from m_t^ℓ to h_t^a. The "max unit embedding" path consists of selecting an embedding corresponding to the unit with the maximum activation. There is one embedding for each possible language output. No gradients are backpropagated through the embedding selection—denoted by the dashed line. A learning signal is generated from both the action prediction and the language prediction denoted by a darker background with white text. The No-Language models are trained without the language pathway altogether.

Artificial Neural Networks (ANNs) provide a means of exploring how an experience dependent learning process can give rise to cognitive abilities. These models allow us to explore questions about what kinds of experience influence acquisition of specific sub-abilities, and whether experiences of different kinds can mutually support each other. We can experiment with these models in ways we cannot so easily experiment with humans, controlling their experiences in ways that would be impractical or unethical in used

with actual human learners. These models allow us to explore causal relationships by choosing the order in which models learn different things. We can address questions such as: how well does the model perform on exact matching tasks when it has not been taught an explicit SCS? How does it perform when it learns an SCS concurrently with equivalence tasks? Is it important for number words to be grounded in visuospatial experience? How much more quickly can the model learn count words when it has already learned visuospatial concepts of exact equivalence? Ultimately, we can attempt to simulate the interactions between cultural experience and language on a human's ability to perform exact matching tasks.

In training RNNs to map pixels to actions on numeric equivalence tasks, we find the emergence of compressed representations of number within the model's latent representations. We note the resemblance of this phenomenon to the compressed mental representation of number that is supported by both behavioral and neurophysiological evidence (Dehaene, 2003; Dehaene et al., 2008). Furthermore, we examine the errors of the models throughout the course of training to observe the emergence and refinement of a human-like Coefficient of Variation (CV) over the course of learning. The CV measures the relationship between the mean of a numeric estimation and the error of the estimation. In human studies, both the verbally inhibited English speakers and the Pirahã exhibit relatively flat CVs (Gordon, 2004; Frank et al., 2008b). The emergence of this phenomenon further establishes the relevance of ANNs as vehicles for modeling the emergence of numerical abilities in humans.

We summarize the contributions of our work as follows.

1. We show that the models we explore can perform exact matching tasks without explicitly learning an SCS.
2. We show the emergence of known human phenomena purely through training the models to perform exact matching tasks.
3. We show that explicitly teaching the models to use an SCS increases their rate of learning on exact matching tasks.
4. We show that the models learn an SCS significantly faster when they have first learned to solve the exact matching tasks suggesting the importance of cultural learning and providing a possible explanation for the language learning data gap (Frank, 2023).

We use these results as a proof of principle for understanding number

cognition and to suggest refinement to our collective understanding of the interactions between language and cultural experience in human cognition.

1. Methods and Materials

1.1. Overview

We use models relying on a convolutional neural network module and several long short-term memory (LSTM) recurrent modules (See Figure 1). We train them to map from inputs consisting of sequences of image frames to action and language outputs within a virtual environment. We construe our models as analogs of human students learning to play an exact equivalence game by learning to imitate the actions of an expert playing the game.

During training, each round of the game involves a demonstration phase, consisting of a sequence of frames showing successive items placed in the image frame by a demonstrator followed by a frame signifying the completion of the sequence. The demonstration phase is followed by the response phase, in which the player produces a sequence of actions each of which causes a response item to appear in the subsequent frame. Once the number of response items matches the number produced by the demonstrator, there is a navigation phase in which the expert navigates to and then presses a 'done' button in the environment indicating completion of the sequence.

We train multiple different variants of the model, varying if and when language accompanies the play of the game; the language type (e.g. English-like, which conforms to the principles of using number to establish cardinality, or Pirahã-like, where number words are not exact), and aspects of the task environment. We intersperse testing episodes as training progresses to assess each model's ability to perform the exact equivalence task and (where applicable) to produce appropriate language.

When language is used, the production of items by both the demonstrator and the player is accompanied by the production of number words, and a special word is produced by the player as it navigates during the navigation phase. During all phases the model produces a language prediction conditioned on each world observation. Unless otherwise stated, the model's language prediction is then used as an additional input into the model's action selection module. The process is as follows: the model first observes the world; it then makes a language prediction; it then uses this prediction to accompany the processed observation as input to its action selection module; the agent then observes the results of its actions in the next observation.

The model's language and the action predictions are trained on expert player data to make better predictions via backpropagation.

When there is no language, the model reduces to the single chain of network components connected with black arrows shown across the top of the diagram in Figure 1. The network must learn to imitate the expert player's actions using only the sequence of images as inputs.

1.2. Environment and Training

The environment can be formalized as a Partially Observable Markov Decision Process (POMDP) (Zhang, 2010) consisting of a tuple (O, S, A, A^*, p_0, p) in which O is a set of images in the form of a 13x23 grid of floating point values. The values within each image correspond to different types of objects within the game. S is a set of game states, A is a set of actions { WAIT, LEFT, RIGHT, UP, DOWN, PRESS }, A^* is a set of optimal actions at each time step, $p_0(s_0)$ is the probability of the initial state, and $p(s_{t+1} \mid a_t, s_t)$ is the probability of a transition from state s_t to s_{t+1} given the action a_t. When language is used, the tuple is augmented with the additional elements L and $L*$ corresponding to the language output and optimal language respectively. Actions and language are represented as one-hot encodings.

During training, the network receives sequences consisting of demonstration, response, and navigation phases, in which the demonstrator produces some number of objects and the expert player reproduces the corresponding number of items before pressing the done button. The demonstration, response, and navigation, proceed as follows. At the beginning of the demonstration phase, the environment initializes an ending button and a counting button with randomly along the top row of the grid. The player's body is always initialized on top of the counting button. The demonstrator/environment samples a target quantity, n, proportionally to a Zipfian distribution, $\frac{1}{n^x}$ where x is the Zipfian exponent. We set x to 1 unless stated otherwise. The possible target quantities for a given trial during training ranges from 1 to 15. Items produced by the demonstrator are then displayed one by one, in the lower half of the image, until n items have been displayed. At each demonstration frame there is a 20% chance that a frame without a new item will be inserted, to ensure that the agent is counting target items rather than counting iterations of the computation without regard to items in the display. The agent is trained to produce the STAY action during the demonstration phase.

We train the model with two different variants of the game, modeled after tasks introduced by Gordon (2004) and used in subsequent studies with the Pirahã. In the persistent visibility task, similar to Gordon's cluster line match task, and a Transient Visibility version similar to his nuts-in-can task. In the Persistent Visibility version, the demonstrated items remain visible after their initial appearance. In the Transient Visibility version, items produced by the demonstrator appear in just a single frame, with no to items appearing at the same location. As with the nuts-in-can task, the Transient Visibility variant forces the agent to have a memory of the total number of target items in order to succeed at the the game.

Once all target items have appeared on the screen, a signal pixel appears in the lower half of the grid to indicate the start of the response phase. In this phase, the player must press the counting button the same number of times as the target quantity. Each press of the counting button causes a new response item to appear neatly in a row in the upper half of the grid. Once the agent has pressed the counting button the correct number of times, it must navigate to the ending button and press it to end the game.

Language labels differ depending on the phase of the game. During the demonstration phase, the labels, thought of as produced by an expert player whose actions the learner is trying to predict, are based on the number of target items that have been displayed up to and including time t. This also applies during skipped frames, where the expert player repeats the current count word for each skipped frame until a new target item is displayed. During the response phase, the language labels correspond to the number of times the player has pressed the counting button, also reflected in the number of response items visible on the grid. Once the player has reached numeric equivalence, and produced the corresponding count word, the navigation phase begins, at which point the expert player produces a "navigate" word while it navigates to and presses the ending button. This ends the trial.

Navigation Baseline Environment. As a control condition relevant for some of the comparisons described below, we introduced another task variant using the same environment without task dependence on quantity. In this task, the environment is the same as the Persistent and Transient Visibility environments, except that the expert player produces a button press for every new target item during the demonstration phase, and is trained to skip the response phase entirely, navigating directly to the ending button and pressing it. There is no language accompanying this task, as it is only used as

an action pre-training baseline. This game was introduced as a control for numeric vs general visual experience.

1.3. Model, Training and Testing

We train models to map, for each time step, from a grid of floating point (pixel) values, o_t, to a vector of logits over possible actions, a_t, and a separate vector of logits over words, ℓ_t. The loss for the action and language predictions at an individual time step for a single training sample is calculated as follows (the subscript t has been removed for legibility):

$$\mathcal{L}_t(a, \ell, \epsilon) = -(1-\epsilon)\log(\frac{e^{a_c}}{\sum_j^{|A|} e^{a_j}}) - \epsilon\log(\frac{e^{\ell_c}}{\sum_k^{|L|} e^{\ell_k}})$$

.

Here a_j is the logit value of action prediction j, ℓ_k is the logit value of the language prediction k at time step t, and a_c and ℓ_c are the logit values of the correct action and language choices made by the expert player at this time step. $|A|$ and $|L|$ are the number of possible actions and words in the language respectively and ϵ is a scaling constant set to 0.5 when jointly training on actions and language. Otherwise ϵ is set to 0 or 1 for action and language pre-training respectively. The loss $\mathcal{L}_\sqcup(a_t, \ell_t, \epsilon)$ is summed over all 36 time steps. We use a batch size of 128. We use stochastic gradient descent and backpropagation through time (BPTT) to train the model over 36 time steps.

Our model architecture is illustrated in Figure 1 Each model variant has a Convolutional Neural Network (CNN) that maps the image, o_t, to a feature vector ν_t, which is then concatenated with an conditional environment vector, z_t, depending on whether the observation is from the Persistent or Temporary Visibility variant. The resulting vector is projected into a single visual latent vector of length 128. It is then fed into the "core" LSTM, $LSTM_c$, which produces an output, h_t^c that is fed into both the action and language LSTMs, $LSTM_a$ and $LSTM_\ell$ respectively.

Models that produce language have an $LSTM_\ell$ that uses h_t^c as input to produce h_t^ℓ which is then fed into a feed-forward network, FFN_ℓ, to produce a language prediction ℓ^t. Models that produce actions have an $LSTM_a$ that uses h_t^c as the input to produce h_t^a which is then fed into a feed-forward network, FFN_a, to produce an action prediction a_t. In variants that involve language, either the word selected by the expert or the word produced by the model (selected by choosing the response word with the strongest activation)

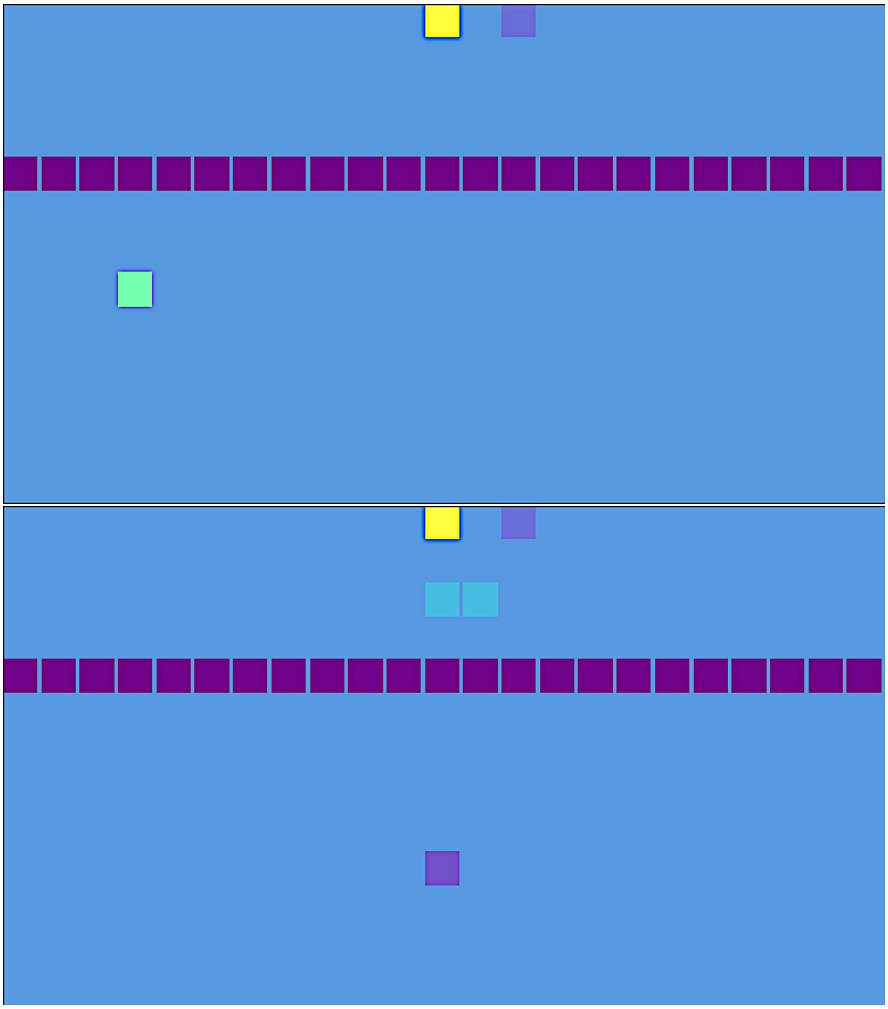

Figure 2: Two images taken four time steps apart from the same episode/trial of the Temporary Visibility task variant. The bright pixel in the lower half of the top image is a target-item. The dark pixel in the center of the lower half of the grid is a visual indication that all target-items have been displayed for the episode. In the upper half of the grid, the bright yellow pixel is the agent, and the dark pixel to the right of the agent is the ending button. The two pixels below the agent in the upper half of the grid in the bottom image are response-items—visual feedback for each press of the counting button.

is mapped to a learned embedding vector, m_t^ℓ, which is then concatenated to h_t^c before being processed by $LSTM_a$. In models that are initially trained without language and then learn language at a later point, the max over the untrained ℓ_t prediction is included as input into their action predictions before learning language. At that point, ℓ_t has not been shaped by gradients, and thus is not equivalent to using language in the model.

The CNNs consists of 2 layers with 9 and 18 channels of 3x3 filters. We use a stride of 1 and a padding of 0. The output of the CNN is flattened before being concatenated to z_t. Each FFN consists of a hidden linear layer that increases the incoming dimensionality by a factor of 3. Each hidden unit is rectified using a ReLU and is followed by a layernorm (Ba et al., 2016). Finally the outputs of the Layernorm are processed by an output linear layer, mapping the dimensionality down to the number of possible output predictions. The predictions are then trained using softmax crossentropy on expert actions or expert language labels.

Each training epoch consists of 768 time steps over 128 environments. Half of the samples in each batch consisted of the Persistent Visibility game variant while the other half consisted of the Temporary Visibility variant. At the beginning of each epoch, we collected new data tuples from the environments using the optimal actions A^*. This creates 768×128 data tuples per epoch, each consisting of an observation, action, task, and word (o_t, a_t, z_t, and ℓ_t). The expected number of steps per trial is determined by the sampling of the target quantities, $p_{skip} = 20\%$, and size of the grid (13x23). In our case, this leads to approximately 20 expected steps per trial. This means there are approximately 4915 trials per epoch. We tested the models on 100 trials for each target quantity from 1-20, stepping through the environment using actions selected by the action with the largest activation. We record a trial as correct when the number of response items is equal to the target quantity. Trials in which the agent fails to press the ending button are automatically terminated after 96 steps.

Models were trained over 3 seeds using an Adam optimizer with a learning rate of $1e{-}4$. Unless otherwise indicated, training continued for a total of 100. All models' weights were initialized by PyTorch's default Kaiming Uniform initialization (He et al., 2015). All training data for each epoch consisted of newly sampled data from the games. The hidden state of each LSTM was reset to a vector of zeros at the beginning of each trial. Over the course of an epoch, the models were trained using backpropagation through time (BPTT) using a sequence length of 36 time steps. After each backward pass,

the next data sequence was shifted one step to forward. This means the next backward pass was performed on steps $t+1$ to $t+36$ as opposed to $t+36$ to $t+2*36$. The models were trained without weight decay and without dropout. All models were trained on Stanford's CCN cluster using a single NVIDIA Titan Xp GPU.

1.4. Model Training Variants

Here we list the full set of variants of the training regimes used to obtain the results described in the results section below.

- **No Language**: trained only to produce action predictions without language production or language input.
- **English**: for each step of the environment, these models were trained to both produce an action and a single, language label for each quantity type (0-15) and navigation step. The max over the model's language predictions is used to select an embedding which is then concatenated to the input to the $LSTM_a$. No gradients were backpropagated through the embedding selection into the language pathway.
- **Pirahã**: similar to the English models except that the ground truth language labels for different quantities consist of a set of 4 possible labels corresponding to the tribe's distribution of words for increasing set size as recorded in Frank et al. (2008a). The labels for each trial were sampled at the beginning of each epoch according to these statistics. We include a deterministic label for quantities of zero and navigation. See Figure 3 for label probabilities. Quantities beyond those listed in Figure 3 use the probability associated with a quantity of 10.
- **Pre English**: these models are the same as the English models except that they were pre-trained to produce language predictions only for 50 epochs before training both the language and action pathways in the same way as the English models. 50 epochs was chosen such that models had converged on 100% language accuracy.
- **Pre Random**: these models are the same as the Pre English models except the ground truth pre-training language labels were shuffled. After the pre-training, these models were trained in the same way as the English models.

- **Correct English**: these models were the same as the No-Language models except that the ground truth English language label was used to select an embedding to concatenate to input to the action $LSTM_a$ during both training and testing. This variant was included to further explore *why* language is beneficial to learning the equivalence tasks.

- **Auxiliary English**: these models are the same as the English models except that they do not feed their language prediction into the action pathway. This variant was included to isolate the effects of using language as a constraint on the learned representations as opposed to use the predicted language labels as a technology.

- **Separate English**: these models are the same as the English models except that instead of having the language and action pathways share vision and early recurrent parameters, the pathways have no shared parameters. i.e. the parameters used in the language and action predictions are kept completely separate. The language prediction is still fed into the action pathway without backpropagating through this operation. This variant was included to isolate the effects of using language as a technology without the effects of constraining the representations.

- **No Actions**: these models are trained to produce English language predictions only, without action predictions.

- **Pre Actions**: these were the same as the English models except that they were pretrained to produce action predictions for 90 epochs without language. 90 epochs was chosen such that models had converged on 100% action accuracy.

- **With Actions**: these are the English models relabeled to emphasize focus on their language performance. We only label them as "With Actions" when viewing the language prediction accuracy in comparisons with other conditions.

- **Pre Navigation**: these models are the same as the Pre Actions models except that their action pre-training takes place on the Navigation environments as described above.

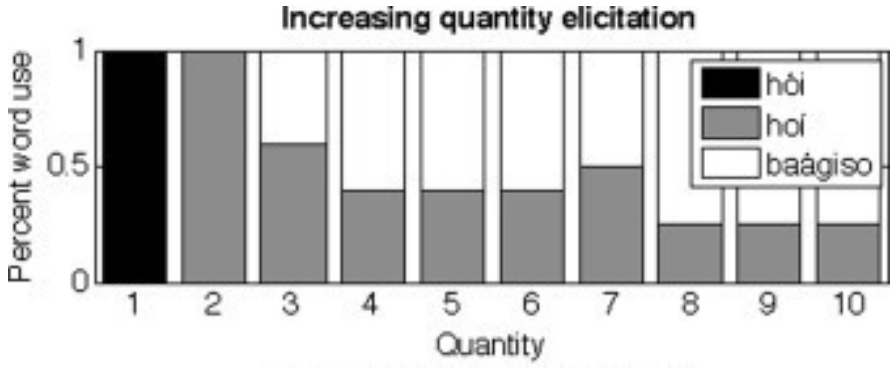

Figure 3: The Pirahã count word distribution. Pirahã model variant language labels are selected at the beginning of each epoch according to the displayed proportions. We include word labels for skipped frames, navigation, and quantities of zero with 100% probability for Pirahã and English variants. Figure taken from (Frank et al., 2008a)

1.5. Measures and Analysis Methods

Principle Components Analysis. For the PCA analysis, we concatenated both the $LSTM_c$ hidden state vector, h_t^c, and the $LSTM_a$ hidden state vector, h_t^a, at each time step over 15 trials for each target quantity from 1 to 20 after 100 training epochs. We then projected each concatenated vector into the two principle components with the largest eigenvalues. These components explained about 55% of the variance. We refer readers to Syms (2008) for an in depth explanation of PCA. We used scipy to implement PCA in code (McKinney, 2010).

2. Results and Discussion

2.1. Learning to Count without an externally provided SCS

We first address the question of whether it is possible for a neural model to learn to perform exact matching tasks without an externally provided system of number words or other symbolic counting system. To test this, we trained the No-Language model variant to perform equivalence tasks without learning or using number words. The final testing accuracy, averaged over all 3 model seeds for target quantities 1-15 was 0.9977 ± 0.0007. using Standard Error Measurement (SEM) over 3 model seeds. Qualitative results over the course of training can be seen in Figures 4.

To establish a stronger connection between these models and human cognition, we explored the ways in which they learned over the course of training.

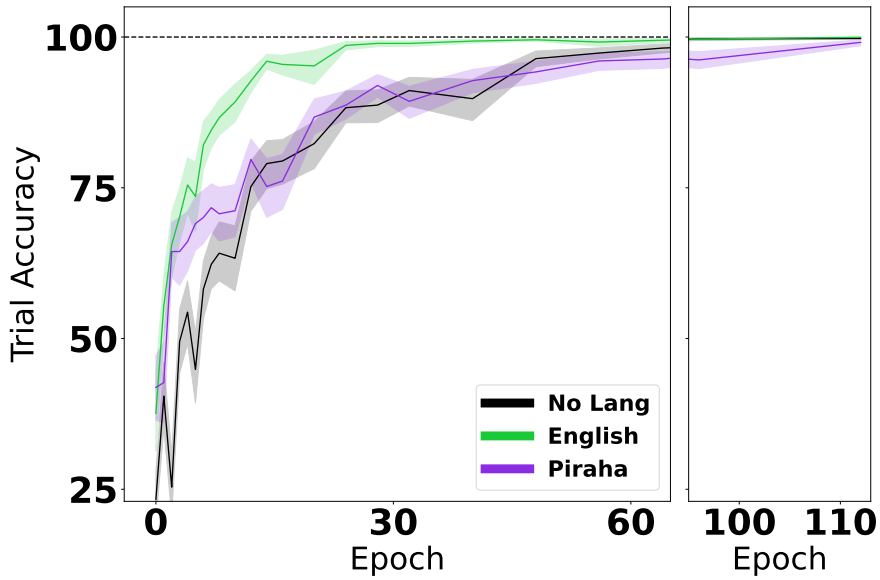

Figure 4: Percentage of successful trials on held-out data averaged over multiple target quantities. Data was collected using unseen trials with target quantities that were encountered during training.

To do this, we looked at histograms of the models' responses during testing as a function of the target quantity, the Coefficient of Variation (CV) associated with these histograms, and visualizations of the latent states of the model using PCA.

Looking at the 2D histograms in Figure 5, we can see that the models learn a conical distribution of response quantities, progressively widening with increasing target quantity. This is similar to the performance of the Pirahã found in the work of Gordon (2004) and Frank et al. (2008a). We qualify the notion of similarity by looking at the log probability of the real Pirahã responses under the probability distribution created by the 2D histogram from Epoch 0 in Figure 5. We compared this value to random samples taken taken directly from the histogram to find that 66.7% had smaller log probabilities, placing the real Pirahã data near the middle of the distribution.

The refinement from Epoch 0 to 3 is similar to the gradual refinement seen in children (Halberda and Feigenson, 2008; Libertus and Brannon, 2010; Wynn, 1998; Xu and Spelke, 2000; Park and Brannon, 2013). We performed a model comparison between the 2D histogram from Epoch 0 and Epoch 3 to validate the claim that accumulated experience may explain the difference

between Frank et al. (2008a)'s human Pirahã and human verbally inhibited English results. For the human Pirahã data, we found that the log probability under Epoch 0 was -83.5 and under Epoch 3 was -95.0. For the verbally inhibited English speaking human data, we found that the log probability under Epoch 0 was -695.5 and under Epoch 3 was -560.1. These results are consistent with the interpretation that less experience explains the human Pirahã data more, and more experience explains the human English data more. These findings suggest that different amounts of experience with exact equivalence tasks, even if completely non-verbal, could partially or even completely explain differences between the Pirahã and verbally-inhibited English speakers on exact equivalence tasks.

To measure this phenomenon more precisely, we use the Coefficient of Variation (CV). The CV is a measure of how the standard deviation of an estimate scales with the mean of the estimate. It is calculated as follows:

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

Where σ is the standard deviation and μ is the mean of the estimate. In this work, we set μ to the ground truth target quantities. It is a commonly observed phenomenon that the standard deviation of humans' numeric estimation tends to scale approximately with the size of the quantity being estimated (Dehaene et al., 2008). This relationship produces a relatively flat CV, although there is some disagreement with the exact nature of the relationship (Testolin et al., 2020).

We can see from Figure 6 that a characteristic CV emerges in all model types for the Zipfian sampling distributions. See Table 2 for linear fits to the models' CVs. From this result we can establish a connection of the models to a large body of work showing that humans exhibit an approximately linear scaling of error with target quantity when performing number estimation (Gordon, 2004; Crollen et al., 2011; Frank et al., 2008a; Le Corre and Carey, 2007). We note, however, that the sampling distribution is different from the model's raw experience with each number. The trials are set up such that the model encounters every number less than the target quantity in addition to the target quantity for any given trial. i.e. the model sees the quantity 1 in every trial regardless of the sampling statistics.

2.2. Pirahã vs English Count Words

We now address the question of whether an externally provided Symbolic Counting System (SCS), in the form of a system of exact number words used in accordance with counting principles, improves a models' ability to perform exact matching tasks. To answer this, we introduced the English model variant which was trained on the expert's actions to perform the exact matching tasks for quantities 1 to 15 while also learning to produce number words and use these words as input to its action selection process. We also trained a set of models with accompanying Pirahã-like number words, in which the expert player's number words were probabilistic, rather than exact in nature, and chosen to match the empirical distribution of the real Pirahã number words as described in *Methods*.

Although all model types had similar final accuracy, the entire developmental period shows a clear performance difference, where the English count words decrease the amount of experience needed by the models for a given proficiency. See Figure 4 for accuracy curves. As a quantitative measure of overall learning efficiency, we used the cumulative test error, measured by summing the error rate calculated on the test data at the end of each epoch. Both the accuracy and error curves rose quickly for the English models, reaching an average cumulative error of 3.1 ± 0.2 (arbitrary units). Both the Pirahã and the No Language models learned much more slowly, with the Pirahã variant at a cumulative error of 5.1 ± 0.1, and the No Language models at 6.5 ± 0.4. Using a one-way ANOVA comparing these three model variants' cumulative error, we find there was a significant difference between at least two of the three variants ($F(16, 8) = 115.426$, $p = 0.0$). Using a Tukey HSD test, we found that the English variant was significantly different than both the Pirahã ($p=0.0$) and No Language ($p=0.0$) variants. See supplemental Table 3 for more details into model comparisons based on cumulative error and Figure 16 for details into the rates at which the models learned each target quantity.

2.3. Causality of Words and Experience

A benefit of using computational models is that we can perform causal experiments to explore how pre-existing number language might affect exact match performance and *visa-versa*. In this pursuit, we created a Pre-English model variant in which we pre-trained the models to learn the grounded English count words without learning actions. After this procedure, we subsequently trained these models in the same manner as the English model

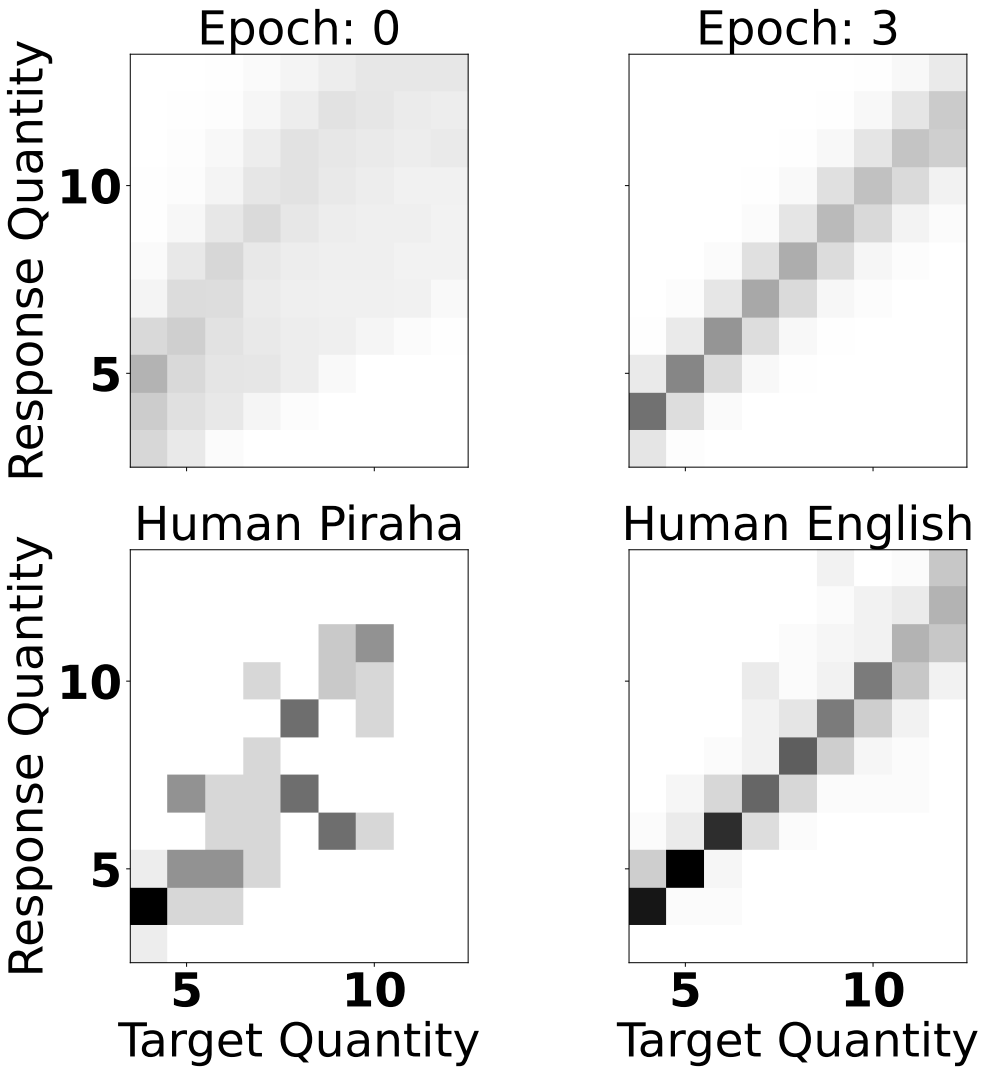

Figure 5: 2D histograms of response quantities as a function of target quantity. Darker squares indicate a greater quantity of trials ended within that bin. The top two panels show the No Language models at the end of different training epochs. The bottom panels show the results from Frank et al. (2008b) with the human Pirahã on the left panel and the verbally inhibited English speakers on the right.

variants—training them jointly on actions and language. This Pre-English variant was meant to give insight into how language might causally influence task performance. As an experimental control, we introduced another

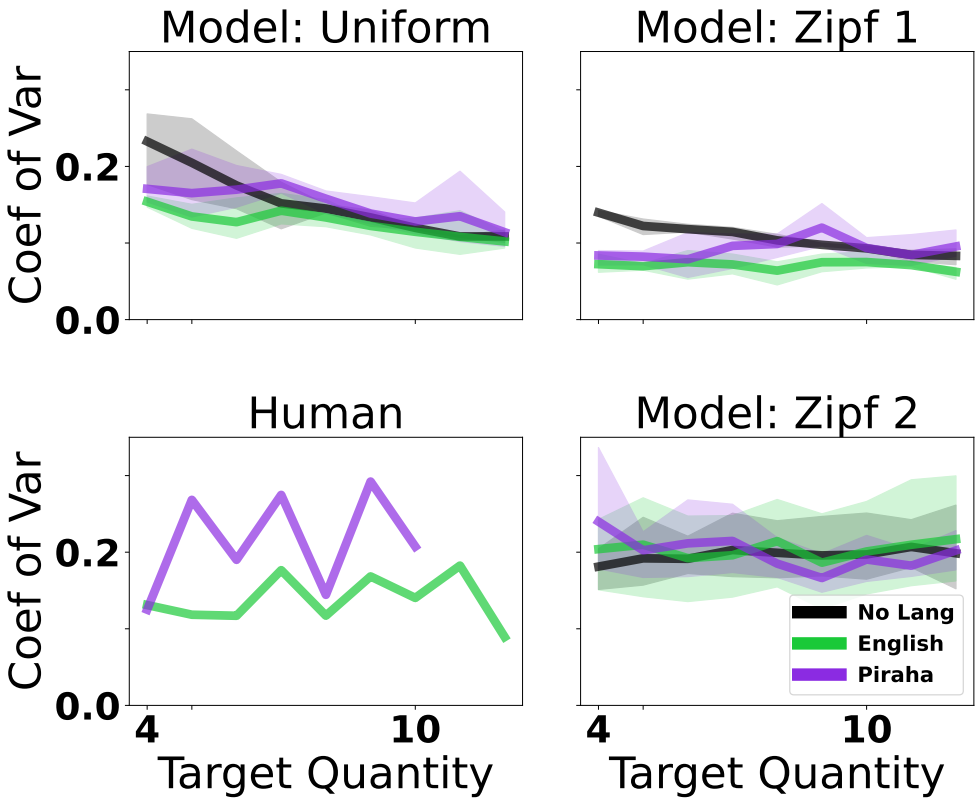

Figure 6: Coefficient of variation (CV) as a function of target quantity for different training distributions. Each of the panels with the title of Model came from trained models with different sampling statistics: uniform sampling and Zipfian sampling with an exponent of 1 and 2. The Model results panels are taken from the first training epoch in the Uniform and Zipf 2 panels, and epoch 3 for the Zipf 1 panel. See supplemental Figures **??-??** for all epochs and Table 2 for linear fits to the CV slope and intercept. The lower left "Human" panel shows the data from Frank et al. (2008b) where the purple line shows the human Pirahã performance and the green line shows the performance of verbally inhibited English speakers. Error bars represent standard error over model seeds.

variant, called Pre-Random, that was pre-trained on shuffled English count word labels. This model served to disentangle the benefits of any form of visuospatial experience from specifically numeric visuospatial experience. The Pre-English variant had a cumulative error of 0.66 ± 0.06 whereas the Pre-Random variant had 3.7 ± 0.2. As a reminder, the English variant had a cumulative error of 3.1 ± 0.2. These results are consistent with the hypothesis that learning count words does reduce the amount of data needed to learn

the exact equivalence tasks. See Figure 7 for more qualitative results.

We further explored causality in the opposite direction, addressing the question of whether learning to perform the actions of exact matching tasks influences the models' ability learn number words. Note that here, we are measuring verbal counting accuracy rather than exact matching task performance. To do this, we created three new model variants to compare to the standard English models. We relabel the standard English variant to With-Actions in Figure 8 to highlight the figure's focus on language performance. The first two variants are the No-Actions variant, which was trained to only make English number word predictions (identical to the standard English variant without producing actions), and the Pre-Actions variant, which was pre-trained to perform the equivalence tasks without language and subsequently trained in the same way as the standard English model—to produce both the language and actions concurrently. Lastly we introduced a baseline variant called Pre-Navigation. This variant was pre-trained to press the counting button on the appearance of each new target item during the demonstration phase, and then skip the response phase, navigating directly to the ending button at the end of the demonstration phase.

As expected, we can see from Figure 8 that pre-training on the task actions in the Pre-Actions model reduces the models' data requirements to learn the English number words relative to the standard, With-Actions baseline. We can see that the numeric experience specifically was the causal factor when we examine the difference between the performance of Pre-Actions and the Pre-Navigation baseline. The Pre-Actions variant achieved a cumulative error of 0.18 ± 0.07, the Pre-Navigation variant achieved 0.43 ± 0.03, and the With-Actns variant achieved 1.37 ± 0.08. Furthermore, we feel it is worth noting that the models all converged on a language solution by approximately 30 training epochs whereas the action performance often took 60-110 epochs for convergence. This potentially suggests that learning number words is an easier task for the models than learning the actions. We leave further exploration of this finding to future work.

2.4. PCA Analysis

To further understand how the models solved the tasks, we explored reduced dimensionality visualizations on the models' latent states. In Figure 9, we show the first two components of a single No Language model's Principle Components Analysis (PCA) on its latent state vectors, in both the Temporary and Persistent Visibility environments (upper and lower panels

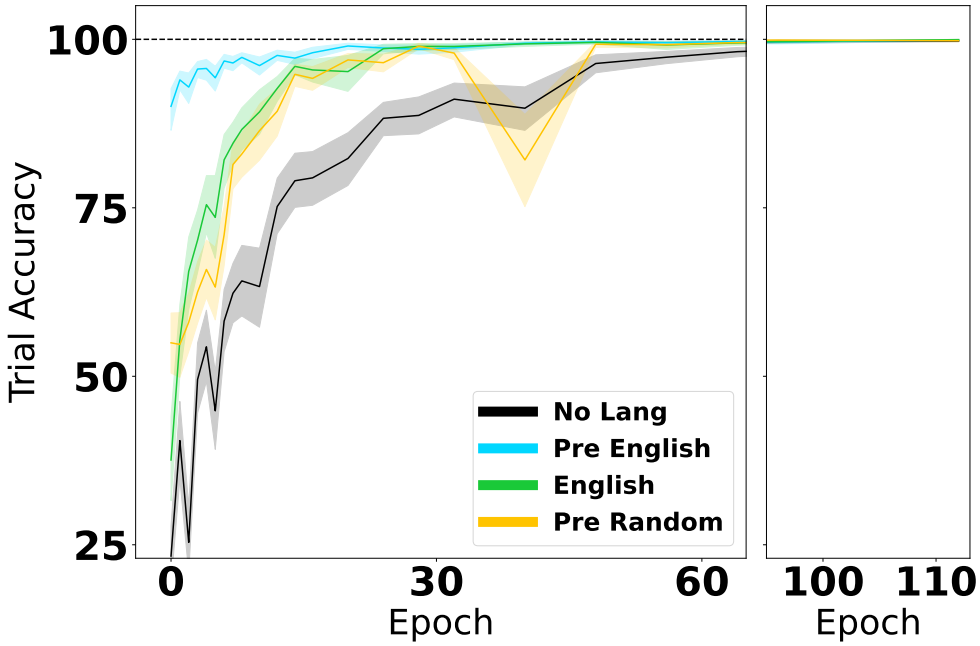

Figure 7: Percentage of successful trials on held-out data for each target quantity at different training epochs. We can see that pre-training models to learn English number words (Pre English) improves the speed of learning relative to no pre-training (English). We included a variant pre-trained on shuffled number words (Pre Random) as an experimental control.

respectively). Each dot corresponds to a single time-step within one of 15 held-out trials for each target quantity. We can see three selected single example trial trajectories plotted in blue, orange, and cyan, overlayed on the projected states. The blue trajectory corresponds to a trial with a target quantity of 3, the orange is a trial that had a target of 6, and the cyan was a trial of 18. For both the Temporary and Persistent visibility cases, the trial trajectories start close together (circled points) and proceed down and then to the right through this space during the demonstration phase. Then during the response phase in the Temporary Visibility variant, the trajectory jumps to the upper left and then walks back down to the left to end somewhat near to the starting position. Trajectories in the Persistent Visibility variant, however, form a conical step pattern during the response phase, in which the states appear to walk around the surface of the cone until they have reached the other side, likely denoting that they have counted the correct amount.

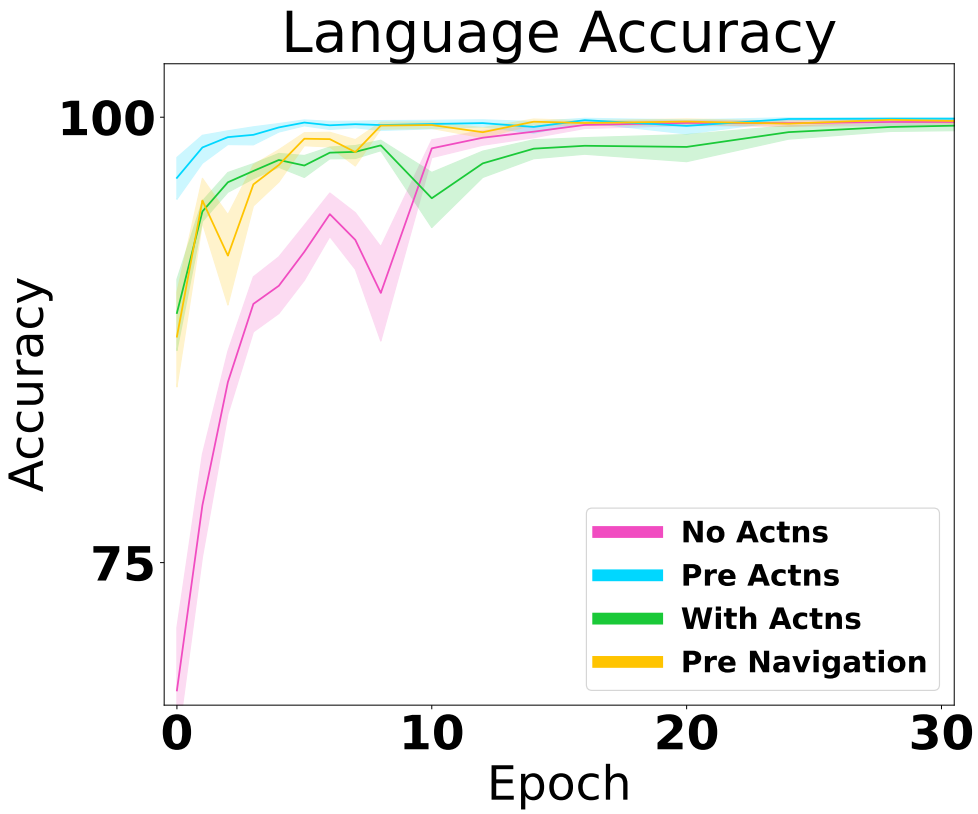

Figure 8: Language performance over the course of training on unseen trials. We highlight the fact that pre-training the models to learn the exact equivalence tasks and subsequently training them on both actions and language (Pre Actns) improves the speed at which they learn the numeric language. This is in comparison to not learning any actions at all (No Actns) and jointly learning actions and language without pre-training. We included a variant that was pre-trained on a navigation task to demonstrate numeric experience is a causal factor.

During the demonstration phase in both environments, it appears as though there is a compression of states with increasing magnitude. To explore connections to human cognition, we included Figure 10 which shows the distances between adjacent states as a function of increasing quantity, $d = h_{t+1} - h_t$, where d is the distance between two recurrent states h_{t+1} and h_t. The color of each point corresponds to the *distance-from-equality*, which is the difference between the number of target items and the number of response items on the grid. We can see that the distance between latent

states decreases non-linearly as the magnitude of the demonstrated quantity increases. Alternatively worded, the model appears to compress its latent space at larger quantities during the demonstration phase. This compression also seems to occur in the Response phase, although it is less apparent in Figure 9, and from the distance-from-equality coloring in Figure 10, it is less clear if the compression corresponds to increasing response quantity in the same way it corresponds to increasing target quantity. The models' compression of the demonstration phase is reminiscent of the logarithmically compressed numberline known to exist in humans (Dehaene, 2003)).

2.5. Language as a Technology

In this section, we explore the question of whether the counting language is mainly beneficial due to the way it shapes the internal, shared representations over the course of learning, or if its benefit is mainly analogous to an assistive technology, effectively providing a cognitive technique for scaffolding the counting process. To examine this question, we compare performance differences between model architectures that use their number words differently. We introduce one model variant, called the Separated English variant, that first predicts a count word and then uses the predicted word as an additional input to its action prediction network; this is similar to the standard English variant except that the action and language pathways leading up to the language label prediction and input are kept completely separate. This allows us to examine the benefits of using language as a technology in the absence of constraint influences on the representations. Next, we introduce the Auxiliary English variant, characterized by jointly learning actions and English language during training without feeding its language predictions into its action selection module. This variant makes language predictions purely as an auxiliary training task, shaping the shared representations. To further isolate the technological effects of a consistent, accurate SCS, without any internal training influences on the model's representations, we introduced a variant called Correct English. This variant is analogous to providing the correct count word to the participant at all stages of training and testing. We do not train the Correct English variant to make any language predictions.

We can think of the Separated English and Correct English variants as using "language as a technology" because these models have the ability to use their predicted number symbols to track and inform their action predictions. We can think of the Auxiliary English variant as a learning constraint that helps shape the learned representations. It is important to note that the

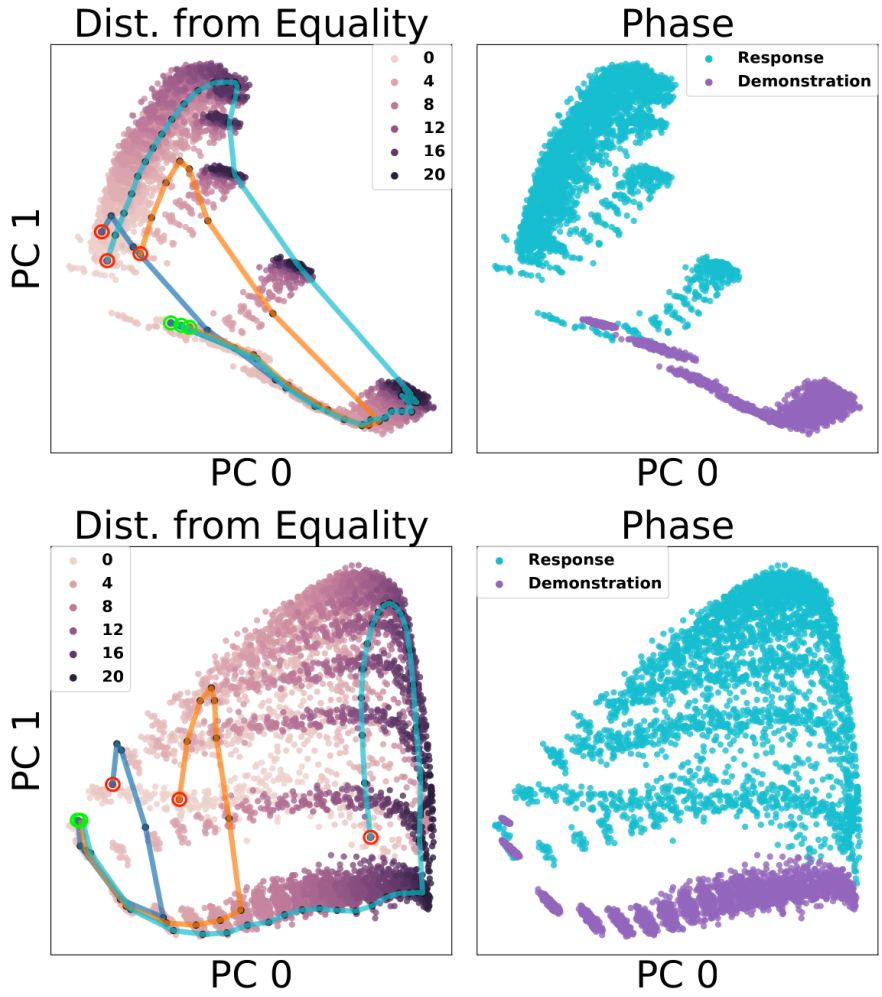

Figure 9: All panels display the first two Principal Components (PCs) of the state vectors over multiple different trials. The h_t^A and h_t^C vectors were concatenated together to create a single state vector for each time t (see Figure 1 for model architecture). All samples were collected from the same No-Language model over 15 trials for target quantities 1-20. The two topmost panels show trajectories from the Temporary Visibility environment. The two bottom-most panels display points from the Persistent Visibility variant. For each data point, the color in the left panel denotes the *distance-from-equality*, defined as the number of target items that have been displayed up to that point minus the number of response items that have been dispensed by the agent. The left panels also show example trajectories for trials with target quantities 3 (blue), 6 (orange), and 18 (cyan). These example trajectories start with a green point and end with a red point. The right panel displays the phase of the experiment for each point in the left panel. The navigation phase and the random timing skips of each trial have been removed for simplicity.

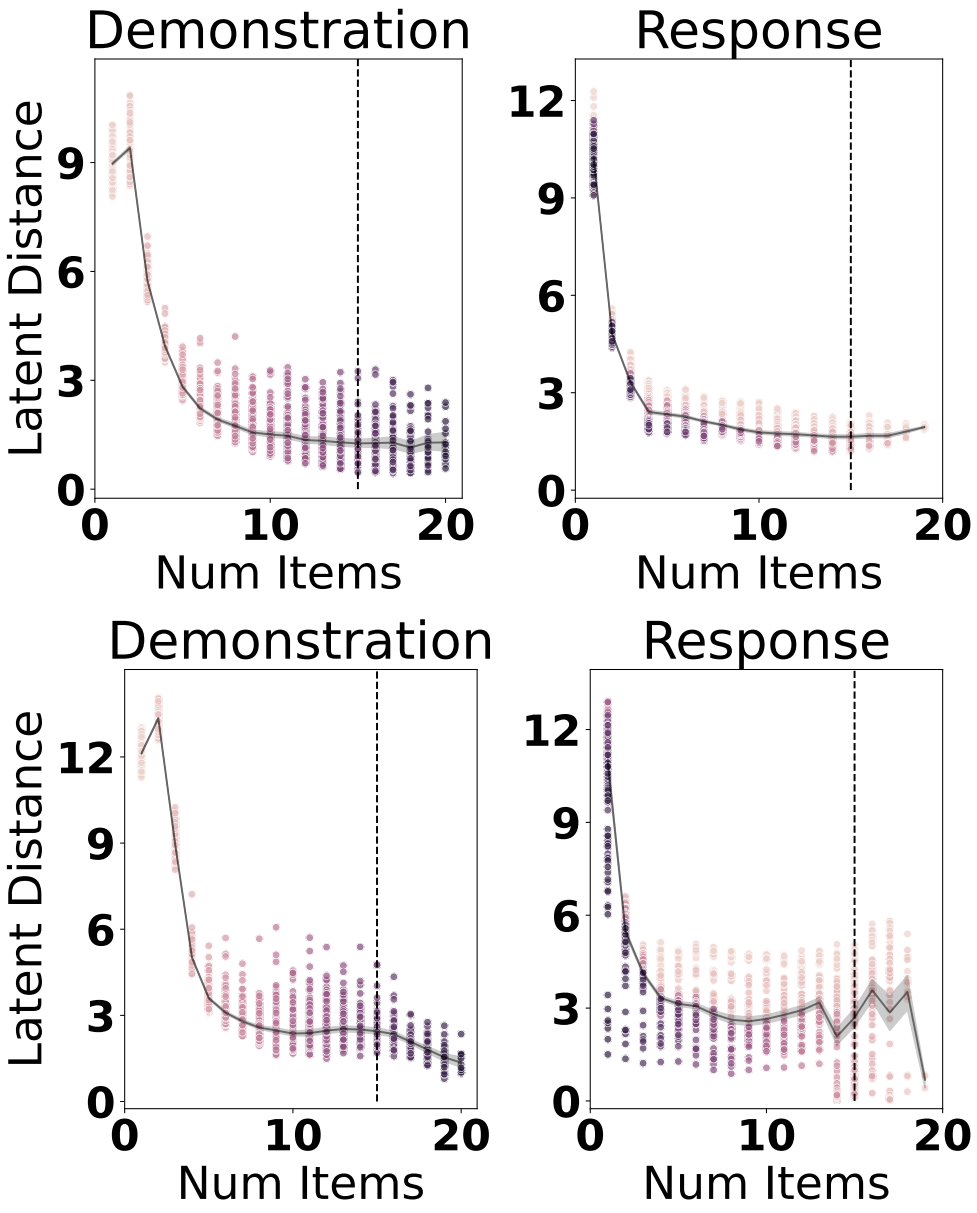

Figure 10: L2 distance between the latent state vectors at time $t-1$ and t. The x-axis represents the number of target items that have been displayed up to time t in the left, Demonstration Phase panels. The x-axis shows the number of response items up to time t in the right, Response Phase panels. The color shows the distance-from-equality, which is calculated as the number of target items minus the number of response items at time t. We see a compression of state space for states associated with larger numbers of items.

benefits of language as a technology and as a constraint are not-necessarily mutually exclusive alternative hypotheses.

The Auxiliary English model had a cumulative error of 4.7 ± 0.2, the standard English model had an error of 3.1 ± 0.2, the Separated English variant had 2.14 ± 0.08, and the Correct English variant had 0.3 ± 0.1. The Correct English variant learned far faster than all other variants, and the Separated English variant learned faster than the English and Auxiliary English variants, with the Auxiliary English variant having the largest cumulative error. We use this as evidence for the interpretation that the English models are indeed using their number words as a technology as predicted by Frank et al. (2008a,b, 2011). We note, however, that there does still seem to be a benefit of using language as a constraint on optimization relative to the No Language model. This is in agreement with findings in non-numeric domains, where the effect of constraining the representations can be more important than using language as a technology. (Mu et al., 2019; Luo et al., 2021). See Figure 11 for qualitative results.

Discussion

As a quick summary of our results, we trained models to perform exact matching tasks with and without teaching them explicit number words. These same models exhibited a relatively flat CV and a non-linear compression of neural state space. The models learned to perform the matching tasks faster when concurrently learning precise number words. The models needed less data to learn the tasks when they were pre-trained with precise number words. Conversely, the models learned to produce precise number words with less data when concurrently trained to perform the matching tasks, and even more so when pre-trained to perform the exact equivalence tasks.

Before this work, there has been standing reason to believe that precise numeric language is necessary to perform exact matching tasks above the count of 3 or 4. Many researchers have speculated that humans are not be capable of exact numeric thought without numeric language (Everett and Madora, 2012b; Everett, 2013b,a; Peters, 2020; Frank et al., 2011; Feigenson et al., 2004). Our work serves as a proof of principle that it is possible for a neural system to learn concepts of exact quantity, beyond the count of 3, purely through learning to perform tasks that depend on exact equivalence. We emphasize that the models learned to solve the task without being explicitly taught to use number symbols. At this point, however, we cannot

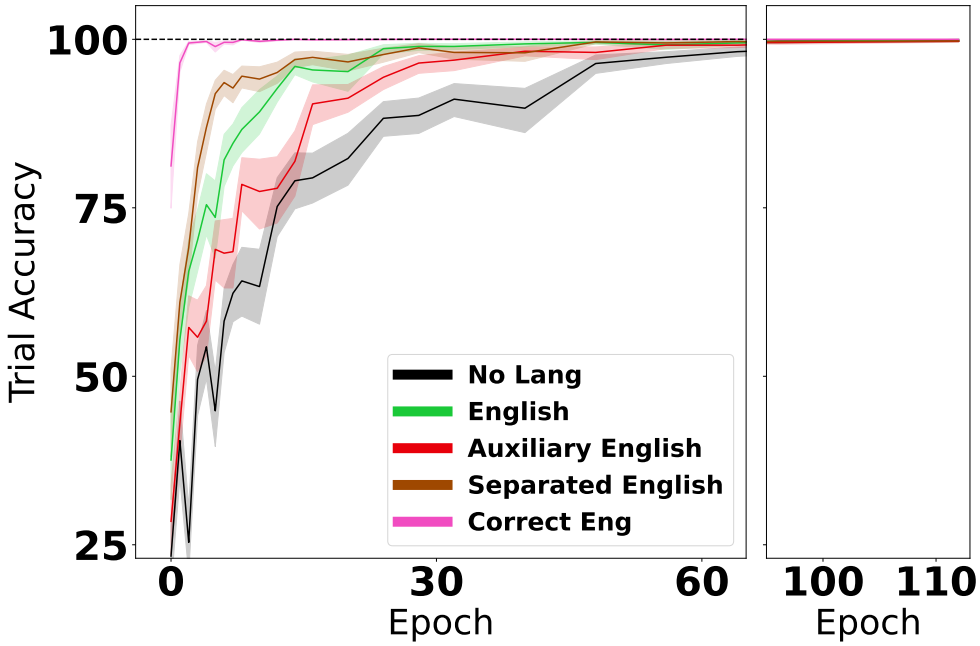

Figure 11: Percentage of successful trials on held-out data averaged over multiple target quantities. We see that both using number words as a contraint on learning, and using number words as a cognitive technology improve the models' data efficiency. It appears, however, that numbers as a technology offer greater improvements.

definitively say whether the models learned to solve the task without using some form of a Symbolic Counting System (SCS). We note that it is possible that the models managed to develop their own latent version of an SCS within their learned representations. We leave exploration of this possibility to future work.[1]

What relevance do these models have to human cognition? We first note that the LSTM Neural Network architecture is a general neural architecture, which readily creates an isomorphism between model activations and biological neural firing rates. This establishes some relevance given that at least

[1]It is not obvious how the models could perform the tasks without some form of symbolic number representation. One potential approach would be to create separate solutions for each possible input. Thus any intermediate computations would not have a fully generalizable, symbolic analog. In this case, any model generalization would fall into one of many possible network solutions.

some human neural activity uses neural firing rates to perform computations (Gerstner et al., 1997). To further establish a connection between the models and *cognition*, we look for the emergence of known human phenomena in the models. Although emergent phenomena do not guarantee relevance of the models, they do increase the likelihood of a meaningful model isomorphism.

In the case of this work, we see the emergence of the relatively flat CV over the course of training. The flat CV is a known characteristic of human number estimation (Frank et al., 2008a, 2011; Revkin et al., 2008). Other modeling approaches have found similar results, often finding a difference between numbers within and outside subitization ranges (Pecyna et al., 2019; Chen et al., 2018; Fang et al., 2018; Creatore et al., 2021). In our work, we show that the statistics of the dataset have an effect on the shape of the CV. Flatter CVs seem to emerge with an increasing Zipfian sampling exponent up to 2. This is consistent with the findings of Cheyette and Piantadosi (2020), who suggested that a flat CV could emerge from the optimal solution to numeric training data that follows a Zipfian distribution with an exponent of 2.

The PCA analysis (Figures 9 and 10) shows that the models compress the distance between latent state representations as the target quantity becomes larger. This is reminiscent of the Weber-Fechner Law found in human neural and behavioral data in which perceptions of numeric magnitude follow a logarithmic compression (Dehaene, 2003). More concretely, the Weber-Fechner law predicts the perceived intensity, p, of a signal, S, as follows: $p = k \ln(\frac{S}{S_0})$. Where k and S_0 are constants specific to the type of stimulus and the perceiver. The emergence of this compression further suggests relevance of the models to human cognition. The compression also potentially gives insight into the strategy that the models use to solve the task. It appears that these models use a different strategy for counting during the demonstration phase than the response phase. We base this conjecture on the distance-from-equality in Figure 10 and the shape of the response phase in Figure 9.

How do we interpret the significance of the No Language models' performance as a model of human cognition? One interpretation is that the models' performance represents a theoretical possibility of human quantity estimation. It is known that human's can improve their numeric estimation abilities with practice (Park and Brannon, 2013; Kim et al., 2018). With enough practice, it may be possible for humans to perfectly estimate quantities beyond common subitization ranges. We note that a practical implementation of

this analogy would require a massive amount of practice in a very restricted task domain. Thus, this interpretation has little practical, pedagogical use. There is some existing evidence that improving one's numeric approximation abilities potentially improves their math proficiency at an early age (Park and Brannon, 2013). The evidence is mixed, however, as to whether the improvements transfer to more advanced levels of mathematics (Kim et al., 2018).

We wish to emphasize that the English model variant demonstrates a cognitive utility of number words beyond the obvious benefits of transmitting knowledge and task details between peers. There are many benefits of communication that we do not wish to address in this work. We instead focus on the intra-agent cognitive effects of language (spoken externally or internally within a single agent). Indeed, we might necessitate that for these models to be a meaningful scientific contribution, it is required that they demonstrate improvement from the inclusion of number words. We see that introducing English number words to the models' training decreases the amount of data they need to learn the exact matching tasks. The advantages that number words provide to the models could offer an explanation as to why there are no known humans who possess the ability to perform exact equivalence tasks without the ability to use number symbols.

It is clear that teaching number words to the models improves their task performance, but why does it help? We suspect that number words, as implemented in the models, provide two potential benefits. The first is that the training signal from the number words may serve as a constraint on the parameter optimization, effectively enhancing the training signal for solving the exact matching tasks. The second potential benefit is that the number word inputs to the action network act as a low variance, generalized signal to solve the tasks. That is, number words serve as a consistent, general feature that can be used to solve the task. The number word signal obviates the need for extracting features of exact quantity from a high dimensional, highly variable visual observation. In summary, we suspect that the number words may 1. act as a constraint on the models' optimization, and 2. provide an easier, generalized, low variance feature to use for action selection.

Both parts of this theory have been previously observed in the machine learning literature (Andreas et al., 2017; Mu et al., 2019; Luo et al., 2021). At a high level, the first part of the theory can be understood as using language as a *post hoc* explanation, to be used for learning (Lampinen et al., 2022). The second part can be understood as using language as a "cognitive

technology" to help solve a task (much as Frank et al. (2008a) proposed). This technology interpretation could potentially be akin to chain-of-thought reasoning (Wei et al., 2022; Prystawski and Goodman, 2023) although we do not explore the connection further in this work.

To empirically explore the effects of using language as an optimization constraint, we introduced the Auxiliary English model variant which was the same as the English variant except for that it did not include its language predictions as input to its action network. We can see from Figure 11 that this Auxiliary English variant learned faster than the No Language variant, but learned slower than the Separated English variant (a variant similar to the standard English models without shared representations between the action and language pathways). These results are consistent with the idea that the language labels offer an optimization benefit without accounting for all beneficial factors.

We empirically explored the second part of the theory by introducing the Correct English model variant. This variant was similar to the standard English variant except that it received the ground truth number labels as input into its action network rather than its own language predictions. We also did not train its language pathway. We can see from Figure 7 that the Correct English variant learns far quicker than both the English and Auxiliary English variants. We can also see from Figure 18 that the English and Correct English models overfit more to the numbers within their training distribution compared to the Auxiliary English, No-Language, and Pirahã variants. This suggests that the English and Correct English variants use/rely on their number symbols to perform the tasks—perhaps progressively more so as the consistency of the number symbols increases.

One interpretation of the human Pirahã data is that the Pirahã perform poorly on the exact equivalence tasks because they have a limited amount of experience emphasizing exact quantity. Our model variants at early stages of training serve as an analogy for humans with limited amounts of experience. Notably, the models at early stages of training make similar errors to the Pirahã. We can see this from early epochs in Figure 5 and from the flatish CV in Figure 6 where the models' errors increase with growing target quantity. Models at early training epochs exhibit stronger performance on small numbers and progressively poorer performance on larger numbers. This effect is pronounced enough that subitization appears to emerge for smaller quantities. This corroborates the work of Cheyette and Piantadosi (2020) who have suggested that subitization can emerge from a single numeric system,

and Testolin et al. (2020) who suggested that subitization and number sense can emerge in a system without pre-built structures for number.

To explore the causal relationship between words and actions, we causally manipulated the models' abilities by pre-training in various ways. These sorts of manipulations allow us to explore questions about how pre-existing knowledge of number words might causally affect a model's ability to learn actions and *visa-versa*. First we focus on the effects of pre-training on grounded number words. Then we discuss the effects of pre-training on actions and how this affects their ability to learn number words. As expected, we can see in the Pre-English variant in Figure 7 that the language pre-training does reduce the amount of data necessary to perform the tasks. To ensure that the effect was caused by the number words as opposed to any experience with the environment, we introduced the Pre-Random variant. This variant was pre-trained on shuffled English number words.

Now, we focus our attention on Figure 8 which shows language performance of various model types differing in their pre-training. We see a beneficial effect on language learning from including actions during training (With Actns), and we see an even greater effect from first pre-training the models to learn the actions followed by jointly training on both actions and language (Pre Actns). We introduced the Pre-Navigation variant as a control to demonstrate that numeric experience is driving the performance gains. These results demonstrate the models' ability to reuse/adapt computations that had been developed for representing quantity in the exact matching tasks. This experiment captures a hypothesized effect of cultural learning on language (Everett, 2005). It acts as a proof of principle that neural systems can convert/adapt representations developed for a visuospatial task for faster language learning. This finding is provocative in that it both suggests a story of how numeric language developed, as well as offering an explanation for the data gap between children and Large Language Models (Frank, 2023). We speculate about the possibility that some early cultures placed a strong emphasis on exact quantity causing the members of the culture to have more robust representations of quantity. This sort of emphasis could have occurred organically in tasks like equal division of natural resources. This is especially plausible in resource scarce environments. Our models exemplify how these cultural pressures could have first formed primitive representations of quantity, which in turn would make it easier to create explicit words/systems for tracking quantity. If this were the case, we would see both language and thought improving one another, creating a self-propagating effect within the

culture.

References

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

Andreas, J., Klein, D., Levine, S., 2017. Learning with latent language. arXiv:1711.00482.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016. Layer normalization. arXiv:1607.06450.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language models are few-shot learners. arXiv:2005.14165.

Casasanto, D., 2005. Crying" whorf". Science 307, 1721–1722.

Chen, S.Y., Zhou, Z., Fang, M., McClelland, J.L., 2018. Can Generic Neural Networks Estimate Numerosity Like Humans? Proceedings of the 40th Annual Meeting of the Cognitive Science Society, CogSci 2018 , 202–207.

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

Chomsky, N., 1972. Language and Mind. Enlarged Edition, Harcourt Brace Jovanovich. URL: https://books.google.com/books?id=QvZrAAAAIAAJ.

Condry, K., Spelke, E., 2008. The development of language and abstract concepts: The case of natural number. Journal of experimental psychology. General 137, 22–38. doi:10.1037/0096-3445.137.1.22.

Creatore, C., Sabathiel, S., Solstad, T., 2021. Learning exact enumeration and approximate estimation in deep neural network models. Cognition 215, 104815. doi:https://doi.org/10.1016/j.cognition.2021.104815.

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.

Davidson, K., Eng, K., Barner, D., 2012. Does learning to count involve a semantic induction? , 162–173.

Dehaene, S., 2003. The neural basis of the weber-fechner law: A logarithmic mental number line. Trends in Cognitive Sciences 7, 145–147. doi:10.1016/S1364-6613(03)00055-X.

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, 1217–1220. doi:10.1126/science.1156540.

Deutscher, G., 2011. Through the Language Glass: Why the World Looks Different in Other Languages. Arrow Books.

Everett, C., 2013a. Independent cross-cultural data reveal linguistic effects on basic numerical cognition. Language and Cognition 5, 99–104. doi:10.1515/langcog-2013-0005.

Everett, C., 2013b. Without language, no distinctly human numerosity: A Reply to Coolidge and Overmann. Current Anthropology 54, 81–82. doi:10.1086/668795.

Everett, C., Madora, K., 2012a. Quantity recognition among speakers of an anumeric language. Cognitive Science 36, 130–141.

Everett, C., Madora, K., 2012b. Quantity recognition among speakers of an anumeric language. Cognitive Science 36, 130–141. doi:10.1111/j.1551-6709.2011.01209.x.

Everett, D., 2005. Cultural constraints on grammar and cognition in pirahã: Another look at the design features of human language. Current anthropology 46, 621–646.

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.

Feigenson, L., Dehaene, S., Spelke, E., 2004. Core systems of number. Trends in Cognitive Sciences 8, 307–314. doi:https://doi.org/10.1016/j.tics.2004.05.002.

Fodor, J.A., Pylyshyn, Z.W., 1988. Connectionism and cognitive architecture: A critical analysis. Cognition 28, 3–71. URL: https://www.sciencedirect.com/science/article/pii/0010027788900315, doi:https://doi.org/10.1016/0010-0277(88)90031-5.

Frank, M.C., 2023. Bridging the data gap between children and large language models. URL: psyarxiv.com/qzbgx, doi:10.31234/osf.io/qzbgx.

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

Frank, M.C., Fedorenko, E., Gibson, E., 2008b. 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.

Gelman, R., Gallistel, C., 1978. The Child's Understanding of Number. Harvard University Press. URL: https://books.google.com/books?id=EeslAQAAIAAJ.

Gerstner, W., Kreiter, A.K., Markram, H., Herz, A.V., 1997. Neural codes: Firing rates and beyond. Proceedings of the National Academy of Sciences of the United States of America 94, 12740–12741. doi:10.1073/pnas.94.24.12740.

Gleitman, L., Papafragou, A., 2005. Language and Thought. Cambridge University Press.

Gopnik, A., Wellman, H.M., 2012. Reconstructing constructivism: causal models, bayesian learning mechanisms, and the theory theory. Psychological bulletin 138 6, 1085–108. URL: https://api.semanticscholar.org/CorpusID:2496804.

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

Halberda, J., Feigenson, L., 2008. Developmental Change in the Acuity of the "Number Sense": The Approximate Number System in 3-, 4-, 5-, and 6-Year-Olds and Adults. Developmental Psychology 44, 1457–1465. doi:10.1037/a0012682.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. CoRR abs/1502.01852. URL: http://arxiv.org/abs/1502.01852, arXiv:1502.01852.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735, doi:10.1162/neco.1997.9.8.1735.

James M. McClelland, T.T.R., 2003. The parallel distributed processing approach to semantic cognition. Nat Rev Neurosci 4, 310–322. doi:https://doi.org/10.1038/nrn1076.

Jara-Ettinger, J., Piantadosi, S., Spelke, E.S., Levy, R., Gibson, E., 2017. Mastery of the logic of natural numbers is not the result of mastery of counting: Evidence from late counters. Developmental science 20, e12459.

Kim, N., Jang, S., Cho, S., 2018. Testing the efficacy of training basic numerical cognition and transfer effects to improvement in children's math ability. Frontiers in Psychology 9. doi:10.3389/fpsyg.2018.01775.

Lampinen, A.K., Dasgupta, I., Chan, S.C.Y., Matthewson, K., Tessler, M.H., Creswell, A., McClelland, J.L., Wang, J.X., Hill, F., 2022. Can language models learn from explanations in context? arXiv:2204.02329.

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, 395–438. doi:10.1016/j.cognition.2006.10.005.

Levinson, S., 1997. 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.

Libertus, M.E., Brannon, E.M., 2010. Stable individual differences in number discrimination in infancy. Developmental Science 13, 900–906. doi:10.1111/j.1467-7687.2009.00948.x.

Luo, X., Sexton, N.J., Love, B.C., 2021. A deep learning account of how language affects thought. Language, Cognition and Neuroscience 0, 1–10. URL: https://doi.org/10.1080/23273798.2021.2001023, doi:10.1080/23273798.2021.2001023.

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, 108–114. doi:10.1016/j.tics.2004.01.003.

McKinney, W., 2010. Data structures for statistical computing in python, in: van der Walt, S., Millman, J. (Eds.), Proceedings of the 9th Python in Science Conference, pp. 51 – 56.

Mu, J., Liang, P., Goodman, N.D., 2019. Shaping visual representations with language for few-shot classification. CoRR abs/1911.02683. URL: http://arxiv.org/abs/1911.02683, arXiv:1911.02683.

Park, J., Brannon, E.M., 2013. Training the approximate number system improves math proficiency. Psychological Science 24, 2013–2019. doi:10.1177/0956797613482944.

Pecyna, L., Cangelosi, A., Nuovo, A.G.D., 2019. A deep neural network for finger counting and numerosity estimation. CoRR abs/1907.05270. URL: http://arxiv.org/abs/1907.05270, arXiv:1907.05270.

Peters, E., 2020. The Approximate Number System (ANS) and Discriminating Magnitudes. Innumeracy in the Wild , 127–139doi:10.1093/oso/9780190861094.003.0011.

Pica, P., Lemer, C., Izard, V., & Dehaene, S., 2004. Exact and approximate arithmetic in an amazonian indigene group 306, 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. Exact Number Concepts Are Limited to the Verbal Count Range. Psychological Science , 095679762110345doi:10.1177/09567976211034502.

Prystawski, B., Goodman, N.D., 2023. Why think step-by-step? reasoning emerges from the locality of experience arXiv:2304.03843.

Revkin, S.K., Piazza, M., Izard, V., Cohen, L., Dehaene, S., 2008. Does subitizing reflect numerical estimation? Psychological Science 19, 607–614. doi:10.1111/j.1467-9280.2008.02130.x. pMID: 18578852.

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

Spelke, E.S., Kinzler, K.D., 2007. Core knowledge. Developmental Science 10, 89–96. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-7687.2007.00569.x, doi:https://doi.org/10.1111/j.1467-7687.2007.00569.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-7687.2007.00569.x.

Syms, C., 2008. Principal components analysis, in: Jørgensen, S.E., Fath, B.D. (Eds.), Encyclopedia of Ecology. Academic Press, Oxford, pp. 2940–2949. doi:https://doi.org/10.1016/B978-008045405-4.00538-3.

Tenenbaum, J.B., Kemp, C., Griffiths, T.L., Goodman, N.D., 2011. How to grow a mind: Statistics, structure, and abstraction. Science 331, 1279–1285. URL: https://www.science.org/doi/abs/10.1126/science.1192788, doi:10.1126/science.1192788, arXiv:https://www.science.org/doi/pdf/10.1126/science.11

Testolin, A., Zou, W.Y., McClelland, J.L., 2020. Numerosity discrimination in deep neural networks: Initial competence, developmental refinement and experience statistics. Developmental Science 23, 1–13. doi:10.1111/desc.12940.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Le, Q., Zhou, D., 2022. Chain of thought prompting elicits reasoning in large language models. CoRR abs/2201.11903. URL: https://arxiv.org/abs/2201.11903, arXiv:2201.11903.

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, 7780–7785. doi:10.1073/pnas.0701644104, arXiv:https://www.pnas.org/doi/pdf/10.1073/pnas.0701644104.

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

Wynn, K., 1998. Psychological foundations of number: numerical competence in human infants. Trends in Cognitive Sciences 2, 296–303. doi:https://doi.org/10.1016/S1364-6613(98)01203-0.

Xu, F., Spelke, E.S., 2000. Large number discrimination in 6-month-old infants. Cognition 74, B1–B11. doi:https://doi.org/10.1016/S0010-0277(99)00066-9.

Zhang, H., 2010. Partially observable markov decision processes: A geometric technique and analysis. Operations Research 58, 214–228. URL: http://www.jstor.org/stable/40605971.

Supplemental Figures

Variant	Zipf Exponent	Cumulative Error	Final Accuracy
English	0	5.0 ± 0.3	0.9990 ± 0.0004
English	1	3.1 ± 0.2	0.9996 ± 0.0002
English	2	7.0 ± 0.6	0.97 ± 0.01
No Lang	0	6.3 ± 0.5	0.995 ± 0.001
No Lang	1	6.5 ± 0.4	0.9977 ± 0.0007
No Lang	2	8.0 ± 0.7	0.971 ± 0.006
Pirahã	0	6.3 ± 0.4	0.984 ± 0.004
Pirahã	1	5.1 ± 0.1	0.991 ± 0.002
Pirahã	2	8.4 ± 0.5	0.90 ± 0.02
Pre Random	1	3.7 ± 0.2	0.9980 ± 0.0007
Pre English	1	0.66 ± 0.06	0.9988 ± 0.0004
Auxiliary English	1	4.7 ± 0.2	0.9971 ± 0.0007
Correct Eng	1	0.3 ± 0.1	1.0 ± 0.0
Separated English	1	2.14 ± 0.08	0.9976 ± 0.0007

Table 1: Cumulative Error Rate is the sum of the error rate calculated on testing data over the course of the entire training period. This value is averaged over 3 model seeds using SEM for confidence intervals. Zipf refers to the Zipfian sampling exponent.

Variant	Zipf	Slope	Intercept	p	R^2
English	0	-0.0088	0.22	0.0	0.652
English	1	-0.0024	0.09	0.0001	0.385
English	2	0.0089	0.14	0.0023	0.242
English	human	0.0021	0.12	0.6287	0.035
No Lang	0	-0.0045	0.17	0.0004	0.308
No Lang	1	-0.0046	0.14	0.0	0.67
No Lang	2	0.0072	0.15	0.0	0.544
Pirahã	0	-0.0057	0.2	0.0126	0.169
Pirahã	1	-0.0005	0.09	0.6331	0.007
Pirahã	2	0.0107	0.11	0.0	0.543
Pirahã	human	0.0135	0.12	0.3308	0.188

Table 2: Linear fits to the Coefficient of Variation demonstrating a relatively flat trend in all model variants. The Zipf column denotes the Zipfian sampling exponent or whether the data came from a human population (for which we do not know the true sampling distribution). Variant denotes the language type used by the models and the humans, although the human English speakers were verbally inhibited during testing. The p-value is from a Wald Test using a t-distribution against the null hypothesis of a slope of 0. And R^2 is the coefficient of determination. We used SciPy to perform the linear regression (McKinney, 2010). These results demonstrate that the CV potentially has an increasingly downward slope with smaller Zipfian sampling exponents.

Comparison	Statistic	p-value	Lower CI	Upper CI
Auxiliary English - Correct Eng	4.469	0.0	0.0	0.0
Auxiliary English - English	1.664	0.0006	0.0006	0.0006
Auxiliary English - No Lang	-1.768	0.0003	0.0003	0.0003
Auxiliary English - Piraha	-0.341	0.9	0.9	0.9
Auxiliary English - Pre English	4.076	0.0	0.0	0.0
Auxiliary English - Pre Random	1.025	0.04	0.04	0.04
Auxiliary English - Separated English	2.589	0.0	0.0	0.0
Correct Eng - English	-2.806	0.0	0.0	0.0
Correct Eng - No Lang	-6.237	0.0	0.0	0.0
Correct Eng - Piraha	-4.81	0.0	0.0	0.0
Correct Eng - Pre English	-0.393	0.9	0.9	0.9
Correct Eng - Pre Random	-3.444	0.0	0.0	0.0
Correct Eng - Separated English	-1.881	0.0001	0.0001	0.0001
English - No Lang	-3.431	0.0	0.0	0.0
English - Piraha	-2.004	0.0	0.0	0.0
English - Pre English	2.412	0.0	0.0	0.0
English - Pre Random	-0.639	0.4	0.4	0.4
English - Separated English	0.925	0.08	0.08	0.08
No Lang - Piraha	1.427	0.003	0.003	0.003
No Lang - Pre English	5.844	0.0	0.0	0.0
No Lang - Pre Random	2.793	0.0	0.0	0.0
No Lang - Separated English	4.356	0.0	0.0	0.0
Piraha - Pre English	4.417	0.0	0.0	0.0
Piraha - Pre Random	1.366	0.004	0.004	0.004
Piraha - Separated English	2.929	0.0	0.0	0.0
Pre English - Pre Random	-3.051	0.0	0.0	0.0
Pre English - Separated English	-1.487	0.002	0.002	0.002
Pre Random - Separated English	1.564	0.001	0.001	0.001

Table 3: Tukey's HSD Pairwise Group Comparisons (95.0% Confidence Interval) comparing mean cumulative error rates over model seeds between different model variants.

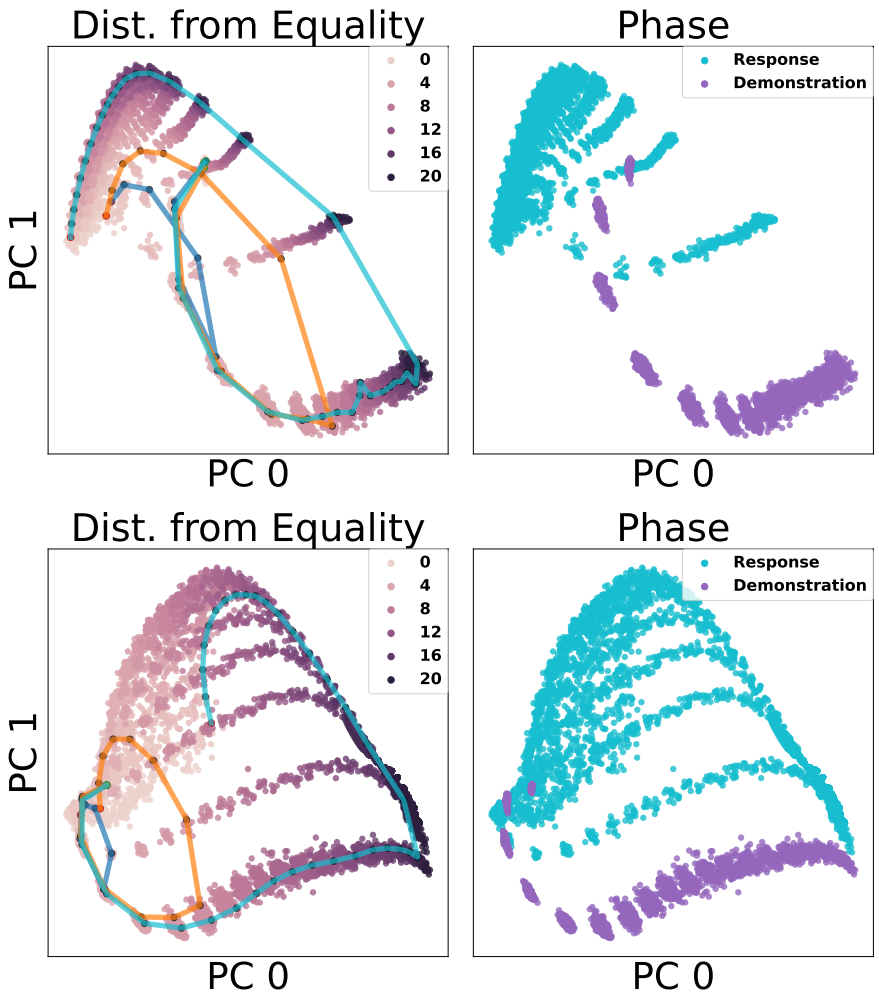

Figure 12: Latent state vectors projected into the first two Principal Components for models trained on a target sampling distribution with a Uniform distribution. The color of each dot in the left panel represents the difference between the number of target items displayed up until time t minus the number of response items. Darker means a greater difference.

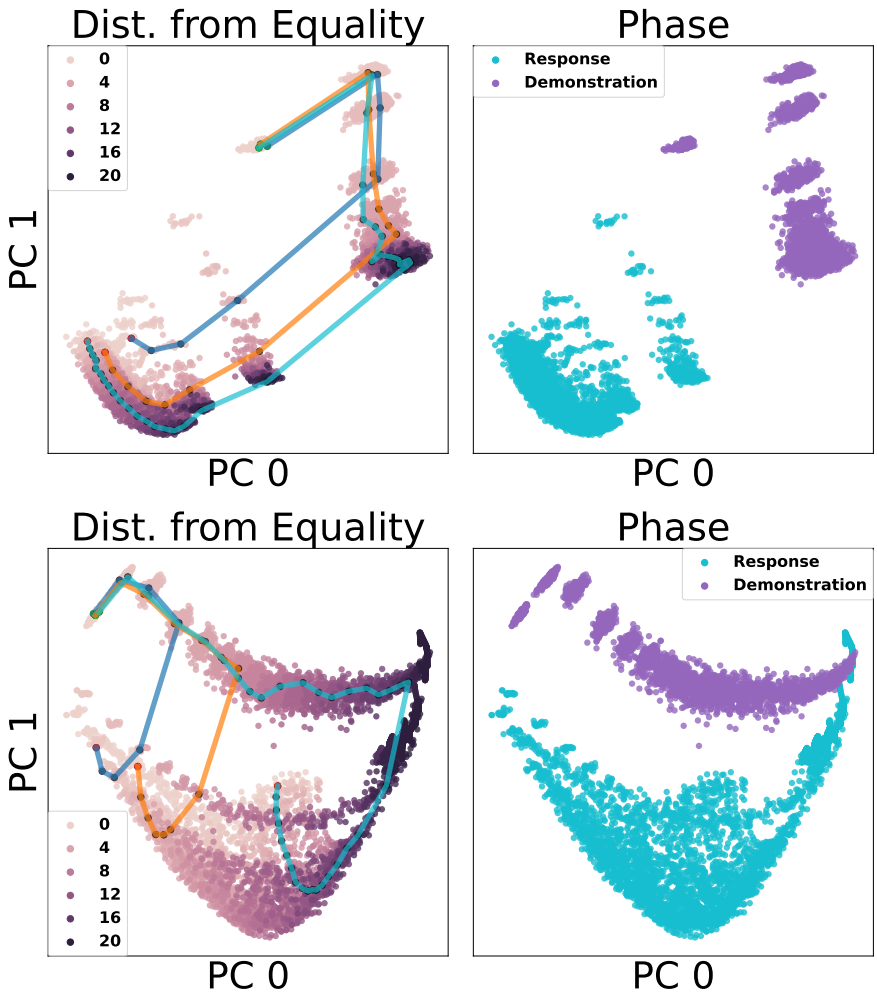

Figure 13: Latent state vectors projected into the first two Principal Components for models trained on a target sampling distribution with a Zipfian distribution using an exponent of 2. The color of each dot in the left panel represents the difference between the number of target items displayed up until time t minus the number of response items. Darker means a greater difference.

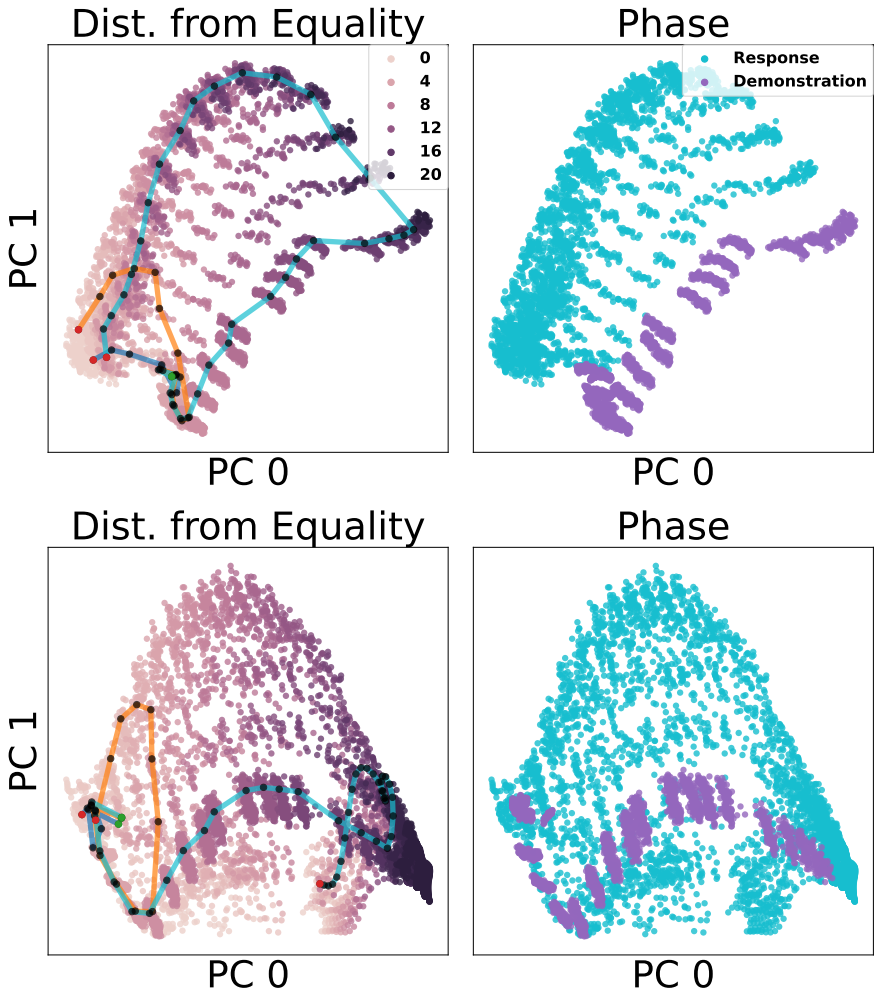

Figure 14: Latent state vectors projected into the first two Principal Components for a single English model. The color of each dot in the left panel represents the difference between the number of target items displayed up until time t minus the number of response items. Darker means a greater difference.

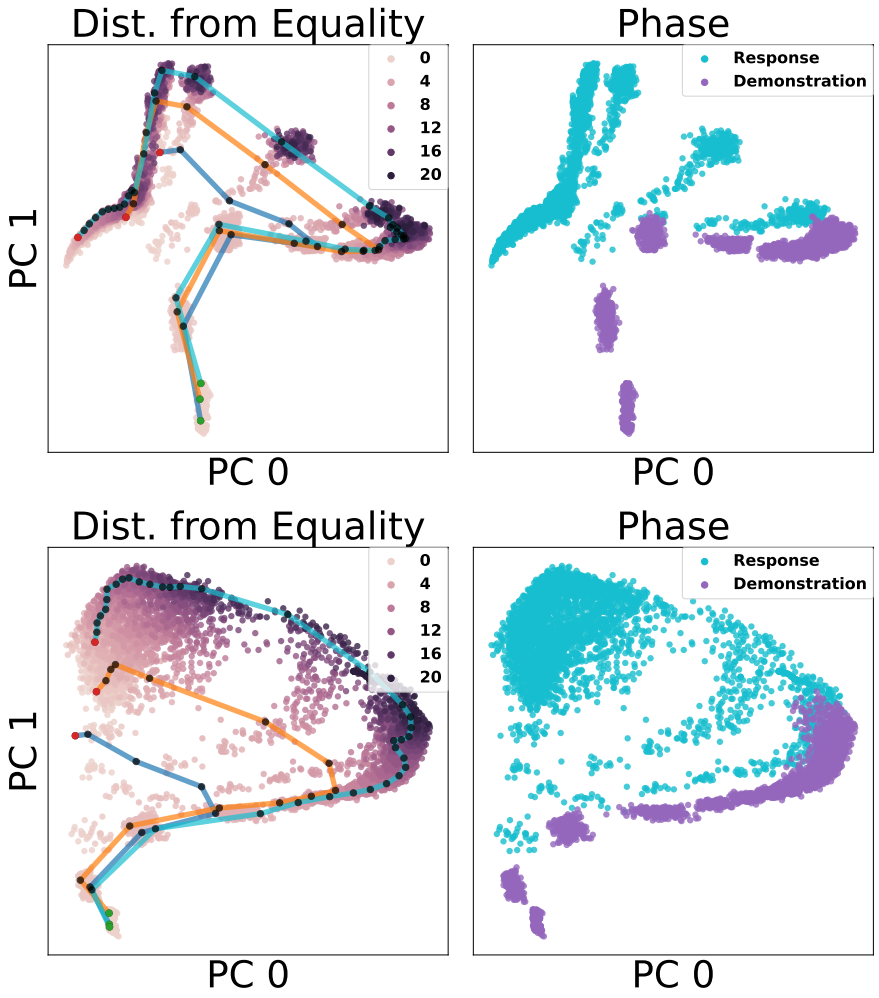

Figure 15: Latent state vectors projected into the first two Principal Components for a single Pirahã model. The color of each dot in the left panel represents the difference between the number of target items displayed up until time t minus the number of response items. Darker means a greater difference.

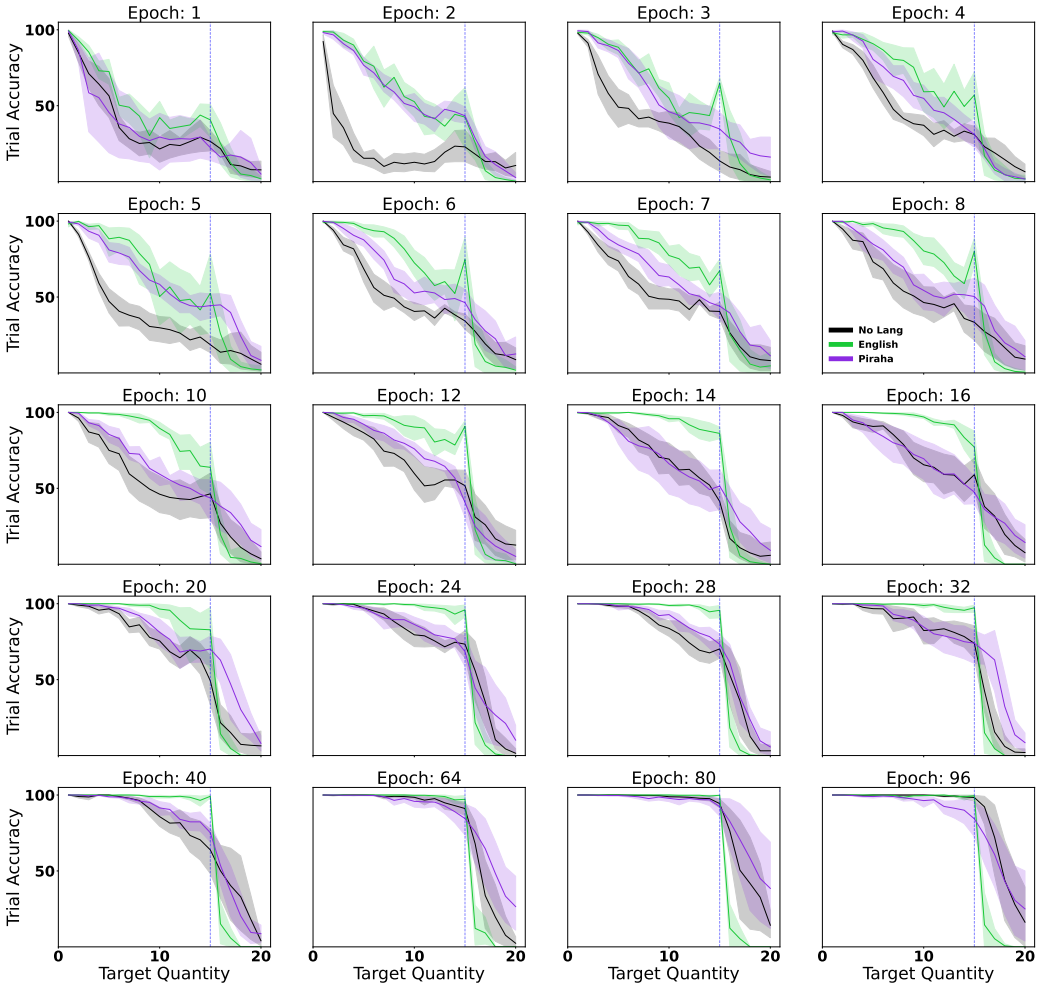

Figure 16: Button press accuracy at different epochs over the course of training.

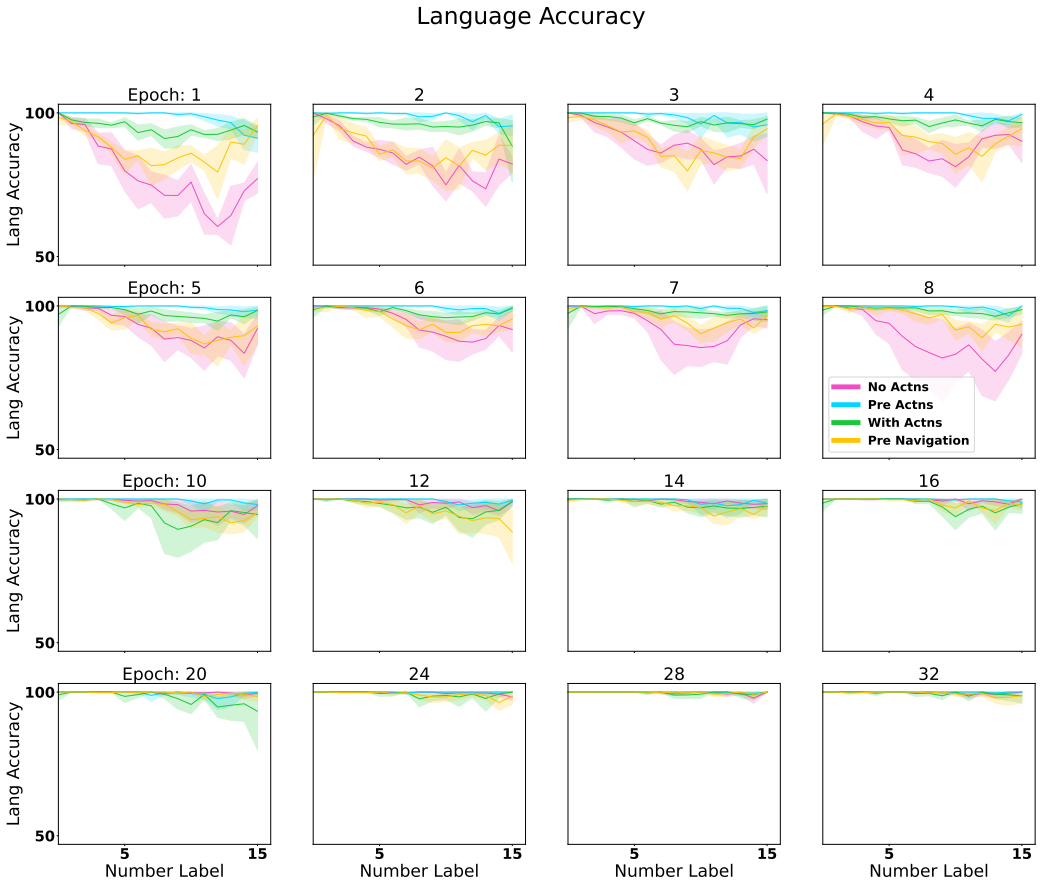

Figure 17: Language accuracy at different epochs over the course of training.

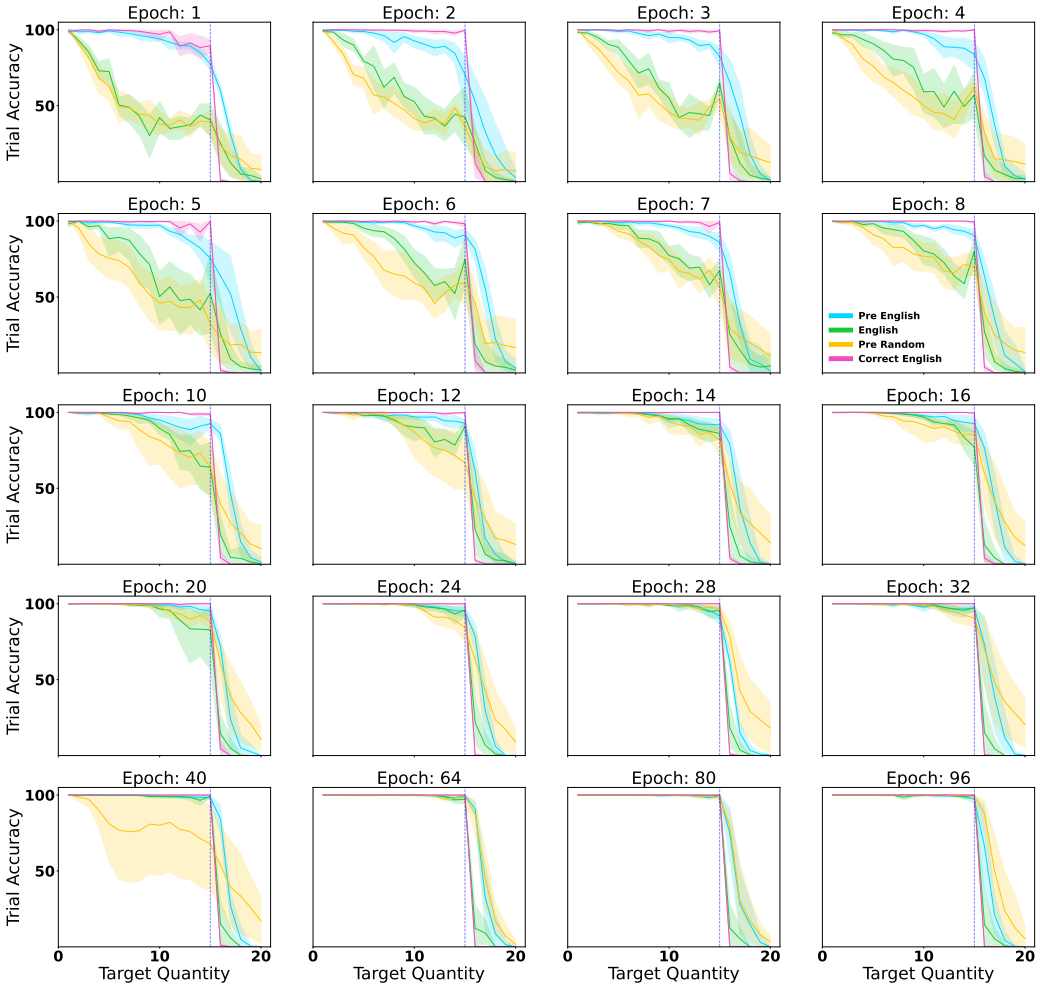

Figure 18: Button press accuracy at different epochs over the course of training.

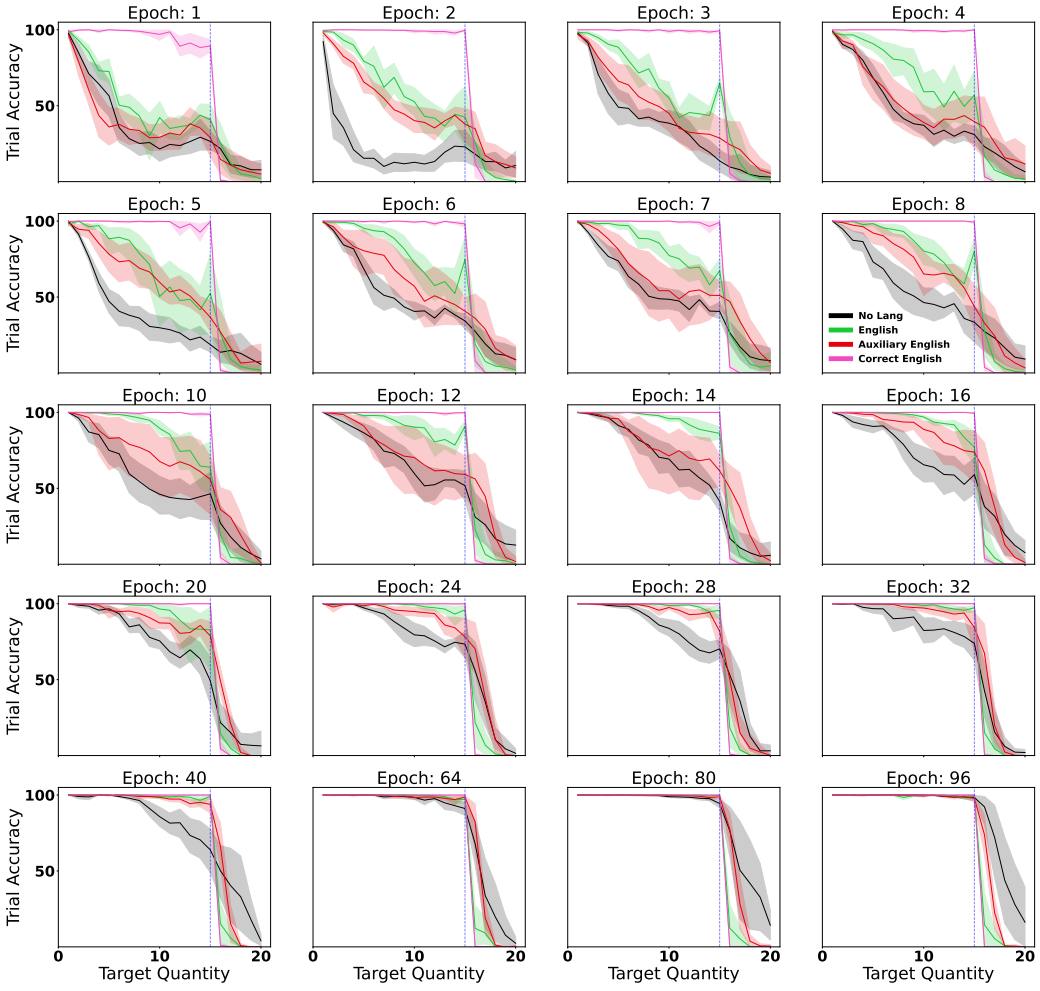

Figure 19: Button press accuracy at different epochs over the course of training.

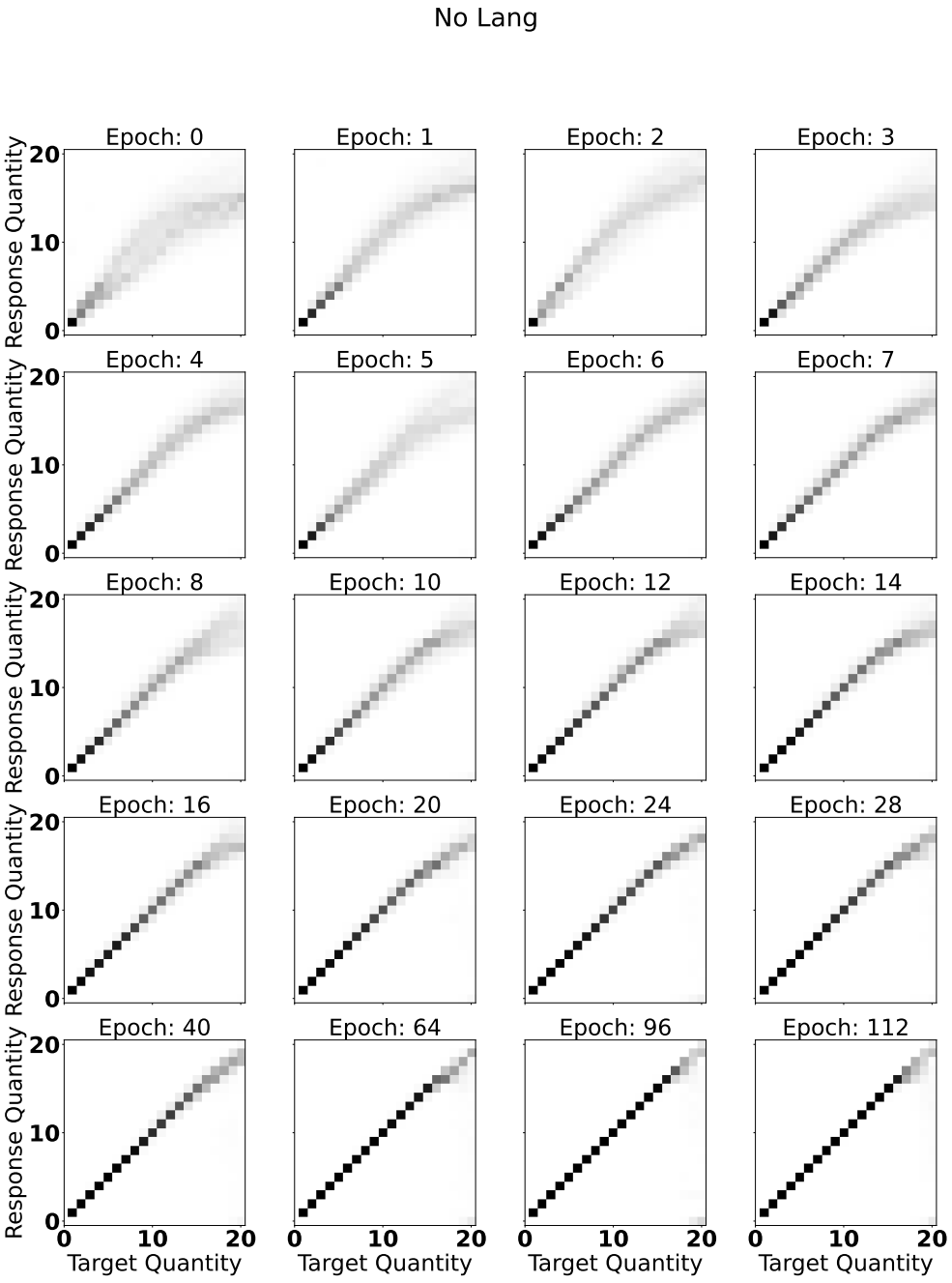

Figure 20: 2D response histograms for No Language model variants.

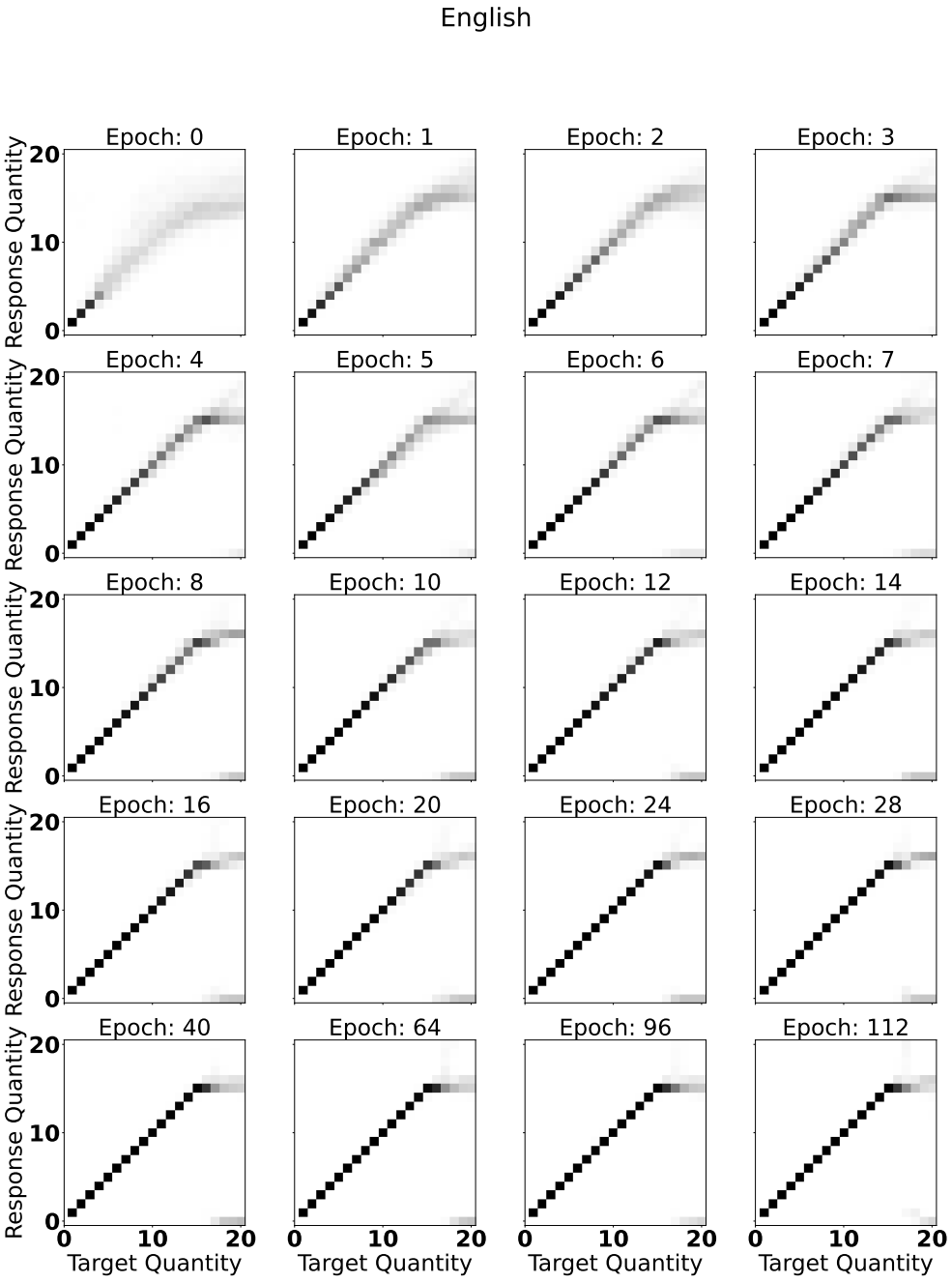

Figure 21: 2D response histograms for English model variants.

Piraha

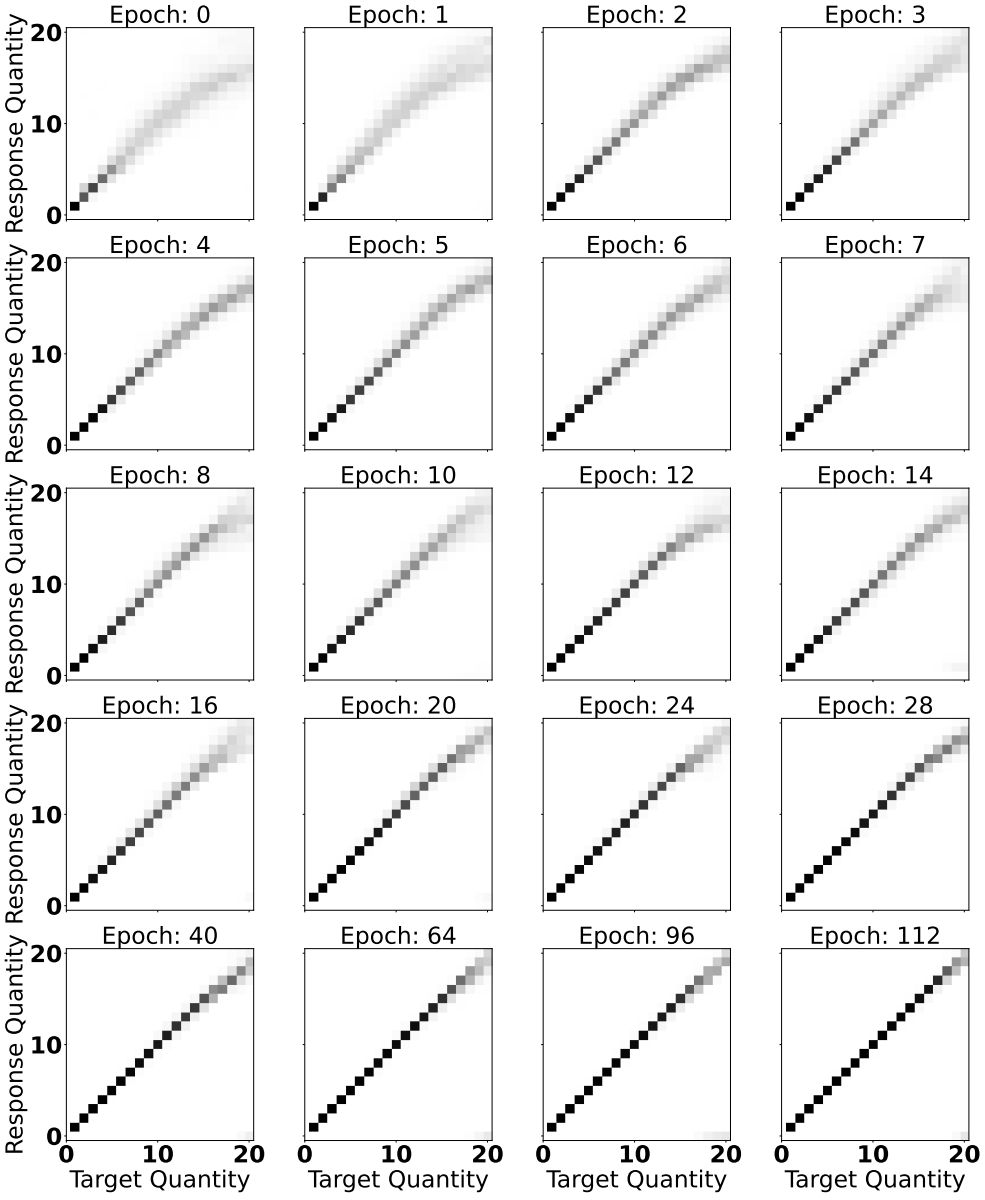

Figure 22: 2D response histograms for Pirahã model variants.