The emergence of numerical representations in communicating artificial agents

Anonymous ACL submission

Abstract

Language provides optimized systems for expressing numerosities, but its role in enabling exact numerical representation and reasoning is still debated. We explore the role of communication in learning numerical representations, using a collaborative task involving two neural-network-based artificial agents. The agents communicate using either discrete tokens or continuous drawings, allowing us to study both symbolic and iconic representations. We find that agents can create effective communication systems, shaped by learnability and efficiency constraints, whose properties partially overlap with human numeral systems. Their interactions suggest they develop an abstract concept of numerosity independent of visual features, though their ability to generalise counting to unseen numerosities remains limited. This study provides insights into the development of numerical representation in artificial agents, and more broadly, on the relationship between communication and abstract reasoning abilities.

1 Communicating and representing numerosities

Using number words, such as *two, ten, eighty-one*, humans can communicate precise information about any numerosity of a set of entities, even up to infinity. Although the use of numbers emerged independently across cultures (Coolidge and Overmann, 2012), the organization of number words shows striking cross-linguistic similarities (Comrie, 2013). In natural languages, number words are optimized to encode the unique properties of numerical concepts (Xu et al., 2020; Denić and Szymanik, 2024), and exhibit several key features accordingly. **Precision** means that the relation between a number word and a numerosity is one-to-one. A number word refers to only one numerosity. Synonyms referring to the same numerosity are rare and their meanings can differ, as *twelve* and *dozen*. Due to **compositionality**, small components can be combined to build new meanings. For example, in English, there are nouns for units, tenths, hundreds and bigger orders of magnitude, which can be combined to indicate any numerosity. Compositionality is key for extending words to unfamiliar numerosities and interpreting a word's meaning based on the understanding of its subparts. Moreover, a compositional code is easier to learn, requiring to memorize a limited set of items. **Arbitrariness**: physical properties of number words are mostly unrelated to their encoded meanings. However, the encoding is sensitive the frequency of occurrence of the symbol: frequently occurring numeral words can have a dedicated lexical item, instead of being derived compositionally, e.g. *fifty* vs *forty-nine*. As the numerosities are not distributed equally (Woodin et al., 2024), less frequent number words can be encoded in longer strings.

Thanks to these properties, number words can efficiently express infinite precise meanings, in exchange of relatively low cognitive effort (Xu et al., 2020; Denić and Szymanik, 2024). While the role of language in communicating numerosities is widely accepted (Gordon, 2004), its contribution to enabling numerical representations remains debated (Carey, 2004; Carey and Barner, 2019). Non-symbolic numerical cognition only enables precise recognition of small numerosities and approximate discrimination of large quantities (Cheyette and Piantadosi, 2020; Dehaene, 2003). The processing of precise larger numerosities is enabled by their symbolic representation: number words would serve this purpose, according to studies sustaining the necessity of language to develop numerical representations (Wiese, 2003, 2007). Other studies report precise numerical abilities independently of language use or availability (Varley et al., 2005; Gelman and Butterworth, 2005). While language is naturally acquired by children, however, its presence alone does not suffice to learn numbers and their relations, which requires training

(Núñez, 2017; Guerrero and Park, 2023; Krajcsi and Fintor, 2023; Núñez, 2017). Numerical representation requires abstraction, recognizing that numerosity is independent of the specific entities, as three apples and three oranges share the same numerosity. It is also independent of visual properties like size or perimeter. Additionally, generalising counting involves applying a rule that adds one unit to the series. In what way are number words and their featured properties useful to seem suitable to develop these concepts and sustain these abilities? Can numerical representation and generalisation ability emerge in a communication task using a non-linguistic communication code? Will the properties of code relate to these abilities and to properties of the number domain in the same way they do in human language? Finally, besides number words, various cultures have independently developed numeral writing systems to record exact quantities for practical purposes, like trade and inventories. Due to this origins, early written numerals can reproduce the number of counted units in a non-arbitrary way, for example: a sign with three strokes could encode the number three (Overmann, 2023; Schmandt-Besserat, 2010). This would correspond to a more iconic encoding, in which the relation between a numerosity and its symbol is explicit. Do different encoding modalities modulate abstraction and generalisation abilities?

2 This study

2.1 Aims

In this study, we explore the emergence of symbolic representations of numerosities and of the properties of a dedicated code to communicate about them, using a collaborative task involving two neural network-based agents. In every trial of the task, one of the two agents, the Sender, sees a picture representing a certain 'target' numerosity and sends a message to the other agent, the Receiver. The Receiver is presented with pictures showing different numerosities and has to choose the one representing the target one. The pictures are designed to make the representation of numerosity a necessary cue to solve the task. We explore the emergence of representational abilities, the properties of the discrete and continuous signals used in successful interactions, and how they respond to the specific requirements of the encoded information. We will compare them to properties observed in human number words. Specifically, we aim to assess whether (1) through communication, the agents develop an *abstract* concept of numerosity, independent of visual features; (2) the agents demonstrate *generalisation* of the counting procedure, communicating correctly about unseen numerosities; (3) the developed code is *precise* and encodes numerosities one-to-one; (4) the code follows a *compositional* rule, with clearly interpretable subparts; (5) the code is *arbitrary*, or whether properties of the images or their distribution (*i.e.* image frequency) influence it.

3 Methods

3.1 Communication Task

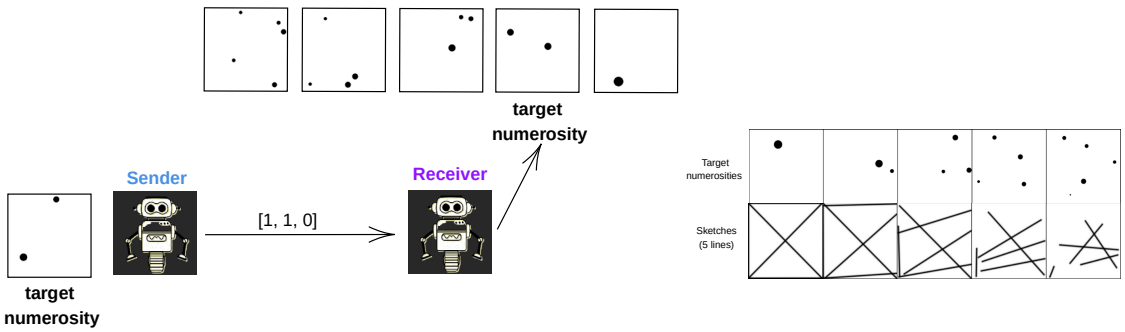
The task is formulated in an emergent communication setup (Boldt and Mortensen, 2024; Lazaridou et al., 2017; Mahaut et al., 2023; Bouchacourt and Baroni, 2018; Kottur et al., 2017), in which a communication protocol is shaped solely through collaborative interaction on a task. The agents' objective is to solve a referential image game being played with images of different numerosities as detailed in Section 3.2.

A referential image game is illustrated in Figure 1a and is constructed as such: one agent, "Sender", has access to a target image, in this case representing a particular numerosity (*e.g.* 2 or 99 as shown in Figure 2) and need to communicate information through either a discrete or a continuous channel. The second agent, "Receiver", parses that information, depending on the format either through an LSTM if the message is a sequence of tokens, or a visual encoding network if the information is encoded through drawing, and makes a decision on which the is the target image the Sender is talking about. The Receiver can choose from a number of possible images consisting of different numerosity images.

To ensure that the agents solve the task by considering numerosity rather than other visual cues, we provide the Receiver with a different target image showing the same numerosity presented to the Sender, as in Figure 1a. The two variants of the communication task can be split based on the nature of the communication channel.

Discrete Communication. "Sender" agent transmits tokens drawn from a predefined vocabulary until it produces a message of a fixed or up to a specified maximum length. Such a discrete communication channel has been studied in several

[1]Robot image sourced from http://surl.li/drdwgh



(a) (b)

Figure 1: (a) Example of a referential game, *different* setting - the Receiver[1] has to guess, based on a discrete message from the Sender, which is the target out of the 5 possible numerosities. Notice that although the target is also **2**, it is not the same instance as seen by the Sender. (b) Examples of target numerosities $1, 2, 3, 4, 5$ and their 5-line corresponding sketches.

other works of emergent communication such as Havrylov and Titov (2017) whose framework we also follow here.

Continuous Communication. "Sender" agent produces a sketch of a predefined number of black, straight lines on a white background. For this scenario, we follow the framework of Mihai and Hare (2021) in which agents learn to communicate about different object categories through drawing.

3.2 Stimuli

Stimuli images are 256 x 256 pixels in size, each representing a different number of black dots on a white background, as illustrated in the examples in Figure 2. In order to make information on numerosity independent of other visual variables such as the area occupied by black pixels, the dots can vary in radius by a margin of 20% but the total black area occupied by pixel is fixed and restricted to 10% of the canvas. The images were created using GeNESIS, a tool designed to create controlled stimuli for numerical cognition research (Zanon et al., 2022). The exact parameters for stimuli generation are detailed in Appendix A. The complete set will be made available upon acceptance. For most experiments, the training set consists in numerosities between 1 and 5, while the extrapolation/interpolation evaluation can include one other numerosity. The exact training and testing numerosities are specified for each experiment.

3.3 Agents

The Sender and Receiver agents are modelled with deep neural network components. Each has a visual encoding system to process the visual input of

(a) (b)

Figure 2: Examples of stimuli: (a) numerosity 2 and (b) numerosity 99.

the numerosities, and a task-specific module. The visual encoder for both agents, for both the discrete and the continuous communication channel, is a pretrained and frozen ViT-B/32 Vision Transformer trained on a contrastive task of pairing images with text captions (Radford et al., 2021).

Discrete communication - Tokens. The sender is modelled following Havrylov and Titov (2017)'s setup. In our case, on top of the pretrained visual encoder, the sender utilises an LSTM to generate a sequence of discrete tokens given the encoded visual information and a start-of-sequence token. To produce discrete tokens but allow learning with gradient backpropagation, the token logits are sampled with a Straight-Through Gumbel Softmax operator. The Receiver also has the visual encoding system which interprets the possible target images (*i.e.* real target + distractors) and encodes each in a reduced dimension vector representation. The Receiver's task-specific module needs to decode the Sender's message. For this, another LSTM processes each token in the message and produces a vector representation of it that can be compared with the image representations through a dot product.

Continuous communication - Sketches. The agents in the continuous communication setup are modelled following Mihai and Hare (2021). Both the Sender and Receiver have the same fixed and pretrained visual encoding network. The Sender uses it to produce a representation of the target image which then is passed on to a drawing module that learns primitive parameters (*i.e.* line coordinates) to then rasterise onto a blank canvas. This will produce the *sketched* message which is passed on to the Receiver (see 1b for examples of target images and corresponding sketches). The sketch is then encoded by the Receiver's visual system, which also encodes the possible target images, and the resulting vectors are compared through an inner product that produces a score for each of the possible target images to indicate the correct one.

Training details. In both communication scenarios, the Receiver outputs the comparison scores between the message (be it in discrete or continuous form) and the batch of numerosity images it can select from. These scores are compared with the real targets and a multi-class hinge loss is utilised to maximise the probability of the correct image being chosen. A Sender-Receiver pair is trained in an end-to-end fashion with gradient backpropagation, for 50 epochs and with a batch size of 50. For each game played, a target instance is sampled randomly from the training set and shown to the Sender and another 5 images for the Receiver (1 target from the same numerosity as the Sender's, but a different instance + 4 images of distinct numerosities). We report results averaged across 3 training seeds.

3.4 Metrics

In each experiment, we will measure task accuracy as the ratio of successful to total interactions, where a successful interaction occurs when the Receiver matches the image with the target numerosity seen by the Sender. We assess **abstraction** abilities considering accuracy on tasks with the same numerosity but different images, and **generalisation** as accuracy on unseen numerosities. Further, we then assess the properties of the emerging communication code.

Precision We measure the precision of a number system by calculating the proportion of one-to-many (i.e. the percentage of messages that map to more than 1 numerosity) and many-to-one mappings (i.e. the percentage of numerosities that are expressed by more than one message).

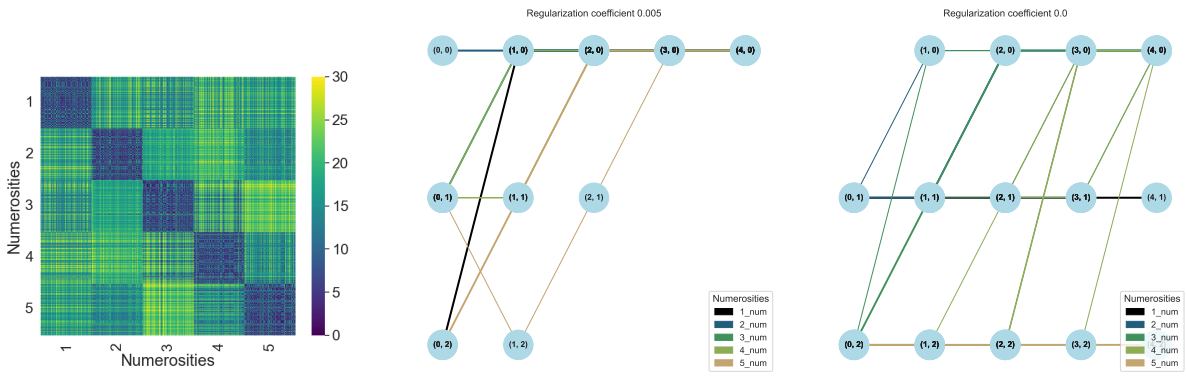
For the discrete case, the mapping is straightforward. For the sketching experiments, messages are continuous and we have to, therefore, bin them into visually similar sketches first. To this end, we project the pixel representation of the sketches into a lower-dimensional space using t-SNE (Van der Maaten and Hinton, 2008, ; see Figure 3a for pairwise sketch dissimilarities in the reduced space) and then apply k-means to cluster visually similar sketches. We can then use these clusters to calculate the one-to-many and many- to-one metrics, similar to the discrete case.

Additionally, we can get a more fine-grained idea of the mappings by calculating the Shannon Entropy H[2] of the message-to-numerosity and the numerosity-to-message-mappings. Low values for H indicate a 'more bijective' mapping, while high values show a more random mapping. Figure 3b and 3c show messages as state-space-diagrams, illustrating the difference between low and high entropy mappings.

Compositionality The experiment is designed such that most semantic features of the input images are fixed (*e.g.* shape, colour and total occupied area) and the only varying feature is the number of elements, *i.e.* the numerosity. The compositionality of emergent numeric symbols is assessed through a series of generalisation experiments. Through extrapolation and interpolation, we assess whether a model which was trained on a specific subset of numerosities can communicate about out-of-distribution numerosities, and if it is successful, whether it produces messages based on previously learned ones (*i.e.* if it learns a rule to produce infinite series of symbols to refer to numerosities).

Arbitrariness Lastly, arbitrariness is assessed through a series of experiments that enforce varying quantities of training samples and compare the emergent symbols with the baseline setup in which all numbers have an equal number of training examples, and hence equal variety. Figure 8c (and Figure 13 in full) shows an illustration of this experiment comparing equal, increasing and decreasing frequency of training samples for each numerosity. The goal is to assess whether frequency affects arbitrariness; if a more varied numerosity results in more complex symbols, or, on the contrary, in a more efficient and simple code.

[2] H for a distribution p over n elements is given by: $H(p) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i)$



(a) (b) (c)

Figure 3: (a) Pair-wise dissimilarity for sketches in the tSNE-reduced pixel-space, also see Figure 14. We can see how agents create more similar sketches for images containing the same numerosity of dots. (b) and (c) show messages as state-space diagrams showing the token sequence from left to right and the different vocabulary items from top to bottom. When message length is penalised (b), message lengths are shortened and produce fewer distinct messages per numerosity compared to the unrestricted case in (c).

4 Results

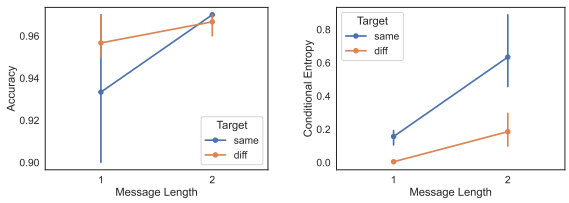
The results are separated based on the two communication modalities: discrete (tokens), and continuous (sketching).

4.1 Discrete Communication

4.1.1 Communication success under various constraints

First, we assess whether a pair of agents can communicate abstract concepts such as numbers without having a predefined symbolic representation in their "vocabulary" (the task is detailed in Section 3.1). The baseline communication accuracy for discrete messages composed of 1 or 2 tokens, in two different setups when the target is the same or a different instance of the same numerosity is shown in Figure 4a. For this experiment, the vocabulary was set to 100 and the numerosities used were between 1 and 5. The results show that agents can successfully come up with a numeric code even when the target differs between sender and receiver and this system becomes more precise as indicated by a lower entropy (see Figure 4b).

The effect of vocabulary size. To assess whether the vocabulary size, *i.e.* the number of possible symbols the agents can utilise is an important factor in the emergence of a concise numeric code, an ablation experiment was run with a fixed message length of 5 and various vocabulary sizes. Agents were trained and tested on numerosities from 1 to 5. The results shown in Table 1 suggest that vocabulary size has no significant effect on communication success, but from the point of view of precision, the entropy is lower with a larger vocabulary size, pointing to more precise code.



(a) (b)

Figure 4: Discrete communication (a) accuracy and (b) message entropy for setups involving identical target pictures (same - blue line) and different pictures showing the same numerosity (diff - orange line). Model trained and tested on numerosities 1 to 5. Results averaged over 3 seeds.

Variable message length and regularisation. Next, the effect of variable-sized messages is explored and the role of an additional regularisation loss term, besides the communication objective, to encourage a more efficient code (*i.e.* shorter messages). In this case, messages of variable length up to maximum **5** tokens are allowed. The numerosities between 1 and 5 were used to train and test the model, and the vocabulary size is 3. The message length regularisation term sums up all the normalised message lengths and is scaled by a coefficient before being added to the communication success objective. The effect of this coefficient ablation on communication success, message entropy and length is shown in Figure 5. The average

Setup	Max Msg Length	Variable Length	Vocabulary Size	Accuracy	Entropy
Diff target	5	No	3	$.95_{\pm.2}$	$1.14_{\pm.34}$
	5	No	5	$.96_{\pm0}$	$1.15_{\pm.31}$
	5	No	10	$.94_{\pm.02}$	$1.13_{\pm.42}$
	5	No	100	$.96_{\pm.008}$	$0.76_{\pm.15}$

Table 1: Model trained on numerosities $1, 2, 3, 4, 5$ with different $vocabulary$ sizes: $3, 5, 10, 100$. Test communication accuracy and entropy is reported across 3 seeds.

message length is halved with a regularisation coefficient of 0.005, without the communication task being highly impacted. Consequently, the code becomes more precise (see Figure 3b versus 3c).

The effect of numerosities' distribution. We vary the training context to assess whether more frequent, and thus varied, numerosities have a more efficient code, and the least varied correspond to a more complex code. The results shown in Appendix C indicate the messages are arbitrary and there is no concrete correlation between message complexity and the number of training samples. The only consistent trend is observed in the accuracy across the different training setups, *Decreasing* being the most difficult to solve with both a discrete and continuous channel (see also Figure 8b)

4.1.2 Generalisation to out-of-distribution numerosities

We further investigate whether the code developed by the agents can generalise to unseen numerosities. For the extrapolation case, results shown in Figure 6, the model was trained on quantities from 1 to 5, and at test time an additional numerosity can be sampled (5 different numerosities are sampled from 6 possible options, 5 already seen numerosities + 1 unseen). Agents communicate with a variable-length message (up to 5 tokens) sampled from a vocabulary of size 3. Figure 6a shows an overall 10% decrease in communication success. For interpolation, we test two intervals as shown in Figure 6b: one narrow (the model being trained on $\{1, 2, 3, 5, 6\}$ and tested also on 4); and one wider (model trained on $\{1, 5, 10, 15, 20\}$ and tested on one additional numerosity from $\{3, 6, 18\}$). Sample messages are shown in Figure 7. Figure 10 in the appendix shows the zero-shot accuracies, which indicate that agents can learn to utilise the message of the closest seen training numerosity, but do not develop a compositional code based on previously learnt numeric representations. The agents were also tested on extrapolating to large numerosities in a close interval (*e.g.* numbers in $[98, 103]$), but in this case, the strategy previously employed fails and their accuracy is at random, *i.e.* 20%, level.

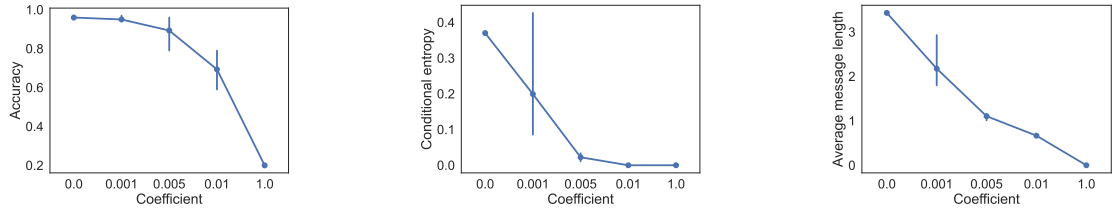
4.2 Continuous Communication

The second communication channel is through sketches and is continuous in the sense that the lines' position, length and orientation are unbounded (in the perimeter of the canvas). The communication task success and properties of the visual code are discussed next.

4.2.1 Communication success and the effect of the training context

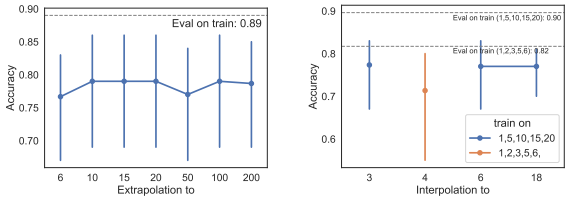
Results of the communication task with a continuous channel (sketches produced with 1 and 5 lines) in 2 different setups: with the same or a different target image from the same numerosity class are shown in Figure 8a. The different setup is only slightly more difficult than the same one resulting in lower accuracy, as is the case for a more restricted communication bottleneck, *i.e.* 1-line sketches. Sample messages of agents conveying information about numerals through 5-lines drawings are shown in Figure 1b and further in Appendix D. The code's precision for the 2 target contexts and sketch complexities are also measured through conditional entropy shown in Figure 9a. Similar to the discrete channel, the 5-line sketches become more precise in the different target contexts and Figure 15a also indicates synonymy decreases.

In Figure 8b, we explore the arbitrariness of the code, more specifically what happens when certain numerosities are more frequent and diverse than others. *Increase* has the least training diversity for numerosity 1 (*e.g.* 100 samples only) and increases to the most examples for numerosity 5 (*e.g.* 700 training samples). *Decrease* is the reverse,



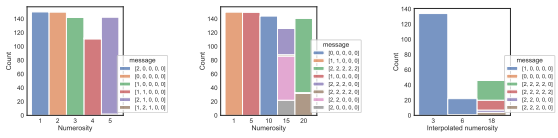
(a) (b) (c)

Figure 5: Agent messaging behaviour when longer messages are penalized by regularising the game loss. Model trained on numerosities $1, 2, 3, 4, 5$, with a variable message length allowed up to **5** tokens, and $vocabulary = 3$. (a) Accuracy (b) conditional entropy of the mapping and (c) average message length for different regularization coefficients.



(a) Extrapolations (b) Interpolations

Figure 6: (a) Model trained on numerosities $1, 2, 3, 4, 5$. Test communication accuracy is reported for each test set that includes, besides trained-on numerosities, one out-of-distribution numerosity. The model usually utilises the same message that was previously chosen for the highest pretrained numerosity (*i.e.* 5) for every new higher extrapolation. (b) Results interpolation.



(a) (b) (c)

Figure 7: (a) Mapping test on 1-5 train numerosities (cond. entropy: 0.015); (b) Mapping test on $[1, 5, 10, 15, 20]$ train numerosities (cond. entropy: 0.262); (c) Mapping interpolated numerosities from (b) (cond. entropy: 0.806). Message indicates overlap.

the most training samples for 1, and the least for 5. The communication game is played in a *different target* setup, and 5-line sketches are allowed. In Figure 8c and Figure 13, sample sketches from agents playing the game with different proportions of training samples are illustrated. The results are further discussed in Appendix D. Overall, the symbol complexity is not influenced by numerosity variety. Instead, the agents seem to focus on the span of the area the elements occupy, or on the number of elements itself, as the sketch for larger numbers also increases in size/span. Further investigation into this relation is needed.

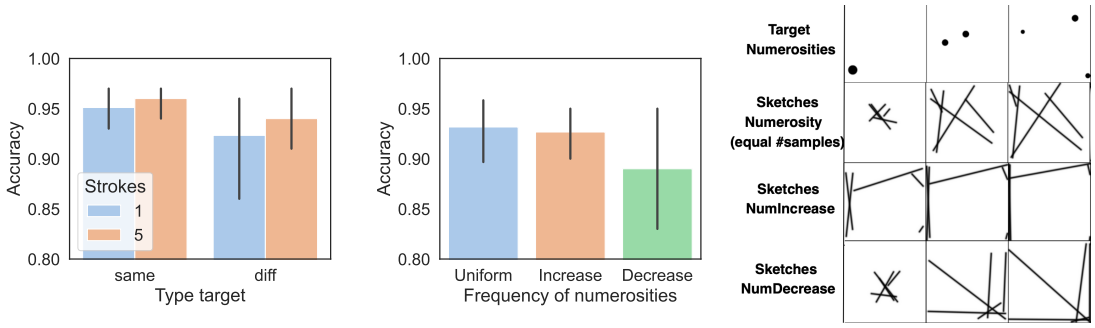
4.3 Generalisation to out-of-distribution numerosities

To assess the ability to generalise to unseen numerosities with a continuous channel, the same procedure described in Section 4.1.2 is followed. The zero-shot accuracy to unseen numerosities (Figure 10c) indicates that on average, extrapolations with a continuous channel are better than chance. That means agents can communicate about quantities, although they do not develop a compositional sketch but rather map everything larger than in training onto a single sketch which usually resembles the highest seen numerosity. Figure 9b shows the mappings of extrapolated numerosities to sketches clustered based on similarity. The numerosity-to-sketch mappings for the train data across different seeds are shown in Appendix E.

5 Discussion

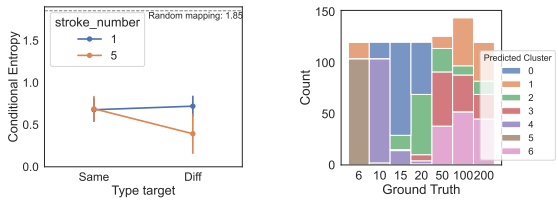
In this experiment series, we showed how two neural network agents learn to communicate about numerosities, and identified the properties of the communication code that emerged from their interactions. The agents develop a code that accurately transmits known numerical information, sharing some properties with human number word organisation:

Precision. In both communication modalities (discrete and sketch), the agents developed high-precision systems, with discrete messages showing minimal overlap in terms of symbol-to-numerosity mapping. The sketch-based communication was more prone to visual clustering but still maintained enough uniqueness to convey distinct numerosities.



(a) (b) (c)

Figure 8: (a) Communication accuracy with a continuous channel in 2 different setups: with the same or a different target image from the same numerosity class. The continuous channel is represented through black 1 or 5-line sketches that can vary in length and position. (b) Communication accuracy on two altered numerosity datasets with unequal number of training samples: increasing from class 1 having the least samples to class 5 having the most, and vice versa. c) Test sketches from training on the 3 different datasets: equal, increasing and decreasing number of class samples.



(a) (b)

Figure 9: (a) Conditional entropy for communication using 1 and 5-line sketches with two target contexts (same or different between the 2 agents). (b) The mapping of sketched messages, after clustering based on similarity to extrapolated numerosities (cond. entropy: 1.348). The model used - different setup and 5-line sketches.

Compositionality. Neither the discrete nor the sketch communication systems exhibited compositionality. The discrete system showed limited compositional properties when tested on out-of-distribution data, while the sketching system tended to rely more on visual heuristics, such as expanding the space used in sketches for larger numerosities. This aspect differs deeply from the highly compositional strategies observed in number word systems in human languages (Xu et al., 2020; Denić and Szymanik, 2024). In the present setting, a non-compositional code is sufficient to communicate, but it would unlikely adapt to expressing infinite numerosities like humans' numerical systems.

Arbitrariness. In the discrete setup, the used code displayed arbitrariness. In sketch-based communication, symbols do not reproduce iconically the numerosity variety; however, they could partially rely on other visual properties like the spread of dots in the space. Moreover, the complexity of the sketches did not appear to strongly correlate with the frequency of numerosities observed in training, contrasting with the tendency to assign shorter or simpler symbols to more frequent meanings observed in human numerical systems, and generally in natural languages (Shannon, 1948; Denić and Szymanik, 2024).

Representing and counting. In line with related research on representation learning, we show that the agents can extract information about the numerosity independent of the visual features of images (Sabathiel et al., 2022; Hein and Diepold, 2024; Testolin et al., 2020), This shows that through the communicated information, the agents can acquire an abstract concept of numerical representation, as revealed in previous studies(Carlsson et al., 2021; Nasr et al., 2019). However, the agents show a limited ability to generalise to unseen numerosities. In human numeral systems, compositionality is an essential property, which permits to encode infinite numerosities starting from finite vocabulary items and is thus related to the generalisation of the counting procedure. We leave the exploration of any causal relation between this inability and the lack of compositionality in the emergent code to future work.

The findings contribute to our understanding of how artificial agents can develop numeric representation systems and provide insights into the potential parallels with human number learning. The study also highlights the limitations in compositionality, suggesting that the agents have not yet developed fully human-like numeral systems.

6 Limitations

The current study focuses on how two artificial agents, modelled through neural networks, learn to develop a communication protocol from scratch to represent and convey numerosities. The study presents the following limitations, which could be addressed in future work.

First, we designed simple images to ensure that numerosity was the only cue available to the agents. Using more naturalistic images could potentially introduce the confounds of additional visual cues, but will extend the applicability of the results to more naturalistic environments where numerosities might be encoded differently (*e.g*. through shape, texture, colour).

Thompson et al. (2024) use more natural images in a zero-shot counting task. Exploring an objective focused solely on counting, potentially as a pertaining step for the communicating agents before interaction, could help improve their abstraction abilities and improve numerosity encodings.

Besides the differences in scene complexity, the agents encountered a limited set of numerosities. This setting differs from the more varied conditions humans experience, limiting the ability to replicate human learning processes and possibly hampered the generalisation of a counting procedure. Further studies with more varied numerosity setting could help answer this limitation.

Next, a considerable limitation is related to the interpretability of the emergent numeric code. Understanding the communication protocols developed by agents is difficult, especially the sketch-based communication. Further analysis and mapping of different parts of a sketch to the corresponding features of the target image is needed.

Lastly, ethical concerns might be raised regarding artificial agents developing their communication protocols and how these could evolve if deployed in an unsupervised environment. This could be addressed by the previous point of enhanced interpretability of the decision-making process and of the emergent code.

References

Brendon Boldt and David R Mortensen. 2024. A review of the applications of deep learning-based emergent communication. *Transactions on Machine Learning Research*.

Diane Bouchacourt and Marco Baroni. 2018. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 981–985.

Susan Carey. 2004. Bootstrapping & the origin of concepts. *Daedalus*, 133(1):59–68.

Susan Carey and David Barner. 2019. Ontogenetic origins of human integer representations. *Trends in cognitive sciences*, 23(10):823–835.

Emil Carlsson, Devdatt P. Dubhashi, and Fredrik D. Johansson. 2021. Learning approximate and exact numeral systems via reinforcement learning. *CoRR*, abs/2105.13857.

Samuel J Cheyette and Steven T Piantadosi. 2020. A unified account of numerosity perception. *Nature Human Behaviour*, 4(12):1265–1272.

Bernard Comrie. 2013. Numeral bases (v2020.3). In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Zenodo.

Frederick L Coolidge and Karenleigh A Overmann. 2012. Numerosity, abstraction, and the emergence of symbolic thinking. *Current Anthropology*, 53(2):204–225.

Stanislas Dehaene. 2003. The neural basis of the weber–fechner law: a logarithmic mental number line. *Trends in cognitive sciences*, 7(4):145–147.

Milica Denić and Jakub Szymanik. 2024. Recursive numeral systems optimize the trade-off between lexicon size and average morphosyntactic complexity. *Cognitive Science*, 48(3):e13424.

Rochel Gelman and Brian Butterworth. 2005. Number and language: how are they related? *Trends in cognitive sciences*, 9(1):6–10.

Peter Gordon. 2004. Numerical cognition without words: Evidence from amazonia. *Science*, 306(5695):496–499.

Diego Guerrero and Joonkoo Park. 2023. Arithmetic thinking as the basis of children's generative number concepts. *Developmental Review*, 67:101062.

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2149–2159. Curran Associates, Inc.

Alice Hein and Klaus Diepold. 2024. Exploring early number abilities with multimodal transformers. *Cognitive Science*, 48(9):e13492.

Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Natural language does not emerge 'naturally' in multi-agent dialog. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2962–2967.

Attila Krajcsi and Edina Fintor. 2023. A refined description of initial symbolic number acquisition. *Cognitive Development*, 65:101288.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2017. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations*.

Matéo Mahaut, Francesca Franzon, Roberto Dessì, and Marco Baroni. 2023. Referential communication in heterogeneous communities of pre-trained visual deep networks. *arXiv preprint arXiv:2302.08913*.

Daniela Mihai and Jonathon Hare. 2021. Learning to draw: Emergent communication through sketching. In *Advances in Neural Information Processing Systems*, volume 34, pages 7153–7166. Curran Associates, Inc.

Khaled Nasr, Pooja Viswanathan, and Andreas Nieder. 2019. Number detectors spontaneously emerge in a deep neural network designed for visual object recognition. *Science advances*, 5(5):eaav7903.

Rafael E Núñez. 2017. Is there really an evolved capacity for number? *Trends in cognitive sciences*, 21(6):409–424.

Karenleigh A Overmann. 2023. *The materiality of numbers: Emergence and elaboration from prehistory to present*. Cambridge University Press.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Silvester Sabathiel, Trygve Solstad, Alberto Testolin, and Flavio Petruzzellis. 2022. Self-communicating deep reinforcement learning agents develop external number representations. In *Proceedings of the Northern Lights Deep Learning Workshop*, volume 3.

Denise Schmandt-Besserat. 2010. The token system of the ancient near east: Its role in counting, writing, the economy and cognition. *The archaeology of measurement: Comprehending heaven, earth and time in ancient societies*, pages 27–34.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Alberto Testolin, Serena Dolfi, Mathijs Rochus, and Marco Zorzi. 2020. Visual sense of number vs. sense of magnitude in humans and machines. *Scientific reports*, 10(1):10045.

Jessica AF Thompson, Hannah Sheahan, Tsvetomira Dumbalska, Julian Sandbrink, Manuela Piazza, and Christopher Summerfield. 2024. Zero-shot counting with a dual-stream neural network model. *arXiv preprint arXiv:2405.09953*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Rosemary A Varley, Nicolai JC Klessinger, Charles AJ Romanowski, and Michael Siegal. 2005. Agrammatic but numerate. *Proceedings of the National Academy of Sciences*, 102(9):3519–3524.

Heike Wiese. 2003. Iconic and non-iconic stages in number development: The role of language. *Trends in cognitive sciences*, 7(9):385–390.

Heike Wiese. 2007. The co-evolution of number concepts and counting words. *Lingua*, 117(5):758–772.

Greg Woodin, Bodo Winter, Jeannette Littlemore, Marcus Perlman, and Jack Grieve. 2024. Large-scale patterns of number use in spoken and written english. *Corpus Linguistics and Linguistic Theory*, 20(1):123–152.

Yang Xu, Emmy Liu, and Terry Regier. 2020. Numeral systems across languages support efficient communication: From approximate numerosity to recursion. *Open Mind*, 4:57–70.

Mirko Zanon, Davide Potrich, Maria Bortot, and Giorgio Vallortigara. 2022. Towards a standardization of non-symbolic numerical experiments: Genesis, a flexible and user-friendly tool to generate controlled stimuli. *Behavior Research Methods*, 54(1):146–157.

A Numerosity stimuli generation

As mentioned in Section 3, the numerical stimuli used for testing whether communication and written language is necessary for the emergence of a number system were produced with the help of GeNEsIS (Zanon et al., 2022), a software tool that allows control over a number of variables.

For the purpose of this study, we generated images with black filled dots on a white background, setting the following parameters and the corresponding values detailed in Table 2:

- **Numerosity** - The number of elements to be generated.
- **Elements' Shape and Colour** - The characteristics of elements that can be created, choosing between circles, squares, squares and triangles and RGB colours.
- **Accepted error** - The tolerance accepted by program when generating stimuli with the specified parameters.

Table 2: Parameters used in GeNEsIS to compute the elements' position and generate a specified number of examples fulfilling these constraints.

Parameter	Value
Numerosity	1-5,100,200
# Generations	1000
Elements' Shape	Circle
Elements' Colour	Black
Arena Shape	Square
Arena Radius	10
Arena Inner and Border Color	White
Accepted Error Rate	0.02%
Elements' Radius Variability	0.2
Total Elements' Area	10

- **Generations** - Number of stimuli to be generated for a specified Numerosity.
- **Radius variability** - The percentage by which the radius can vary between individual elements in a generation. For circle elements, this is the circle radius.
- **Total Area** - The sum of areas of individual elements in a generation cannot exceed this specific parameter. This is done to prevent visual encoding models to learn to communicate about number of black pixels versus white pixels.
- **Arena Radius and Shape** - The dimension and shape of the field in which the elements are created.
- **Arena Inner and Border Color** - The colour of the field and its border.

After generating a stimuli pool with a specific number of elements (*i.e.* a specific numerosity) fulfilling a set of parameters, these are saved as 224 x 224 pixel PNG image files. For each numerosity a pool of 1000 exemplars are generated, out of which 700 are used for training communicating agents, 150 for validation and 150 on test.

For image stimuli showing larger numerosities, such as 100 and 200, it becomes infeasible to use GeNEsIS directly so the approach used instead is to stitch together patches independently generated. Images containing a smaller number of elements, *e.g.* 4 images of 25 elements can be combined and resized to create an image of 100 elements. As this approach is time-consuming, we only evaluate these numerosities with the test set of 150 images for each class.

B Zero-shot accuracy on unseen numerosities

We provide the zero-shot accuracy achieved for both the discrete and the continuous channels in Figure 10. As discussed in the main paper, the agents do not learn a compositional rule that allows them to develop a new code (be it discrete or continuous) for unseen numerosities in a zero-shot manner. It is worth noticing that the variance is high due to agents choosing to utilise the symbol previously developed for the highest numerosity they have seen during training. The accuracy and standard deviation are reported across 3 seeds and while in some, they stick to associating the known symbol with the new numerosity (usually in 2 out of 3 seeds), during other test runs they simply use it for the known numerosity which causes a very a low success rate on the new one and hence the high variance. While this is the result for zero-shot generalisation, it might be worth it in the future to explore if a compositional code could emerge through a few-shot learning approach, such as successive introduction of new numerosities during training.

C Exact mappings of discrete messages to different distributions of training data - Discrete channel

In the main paper, we aim to assess the code's arbitrariness by varying the distribution of input data. Numerosity classes can increase or decrease in number of training samples. The model used for this experiment has variable message length, up to 5 tokens, vocabulary of size 3 and a message regularisation loss with a coefficient of 0.005. Figure 11 shows the test accuracies for different distributions of training data and Figure 12 illustrates the exact mappings for the *Increase* and *Decrease* setups across 3 seeds. An example mapping for a *Uniform* distribution is shown in Figure 7a of the main paper. Contrary to expectations, a significant rule across discrete messages produced with different training data distributions cannot be observed, which indicates that the code is arbitrary and un-

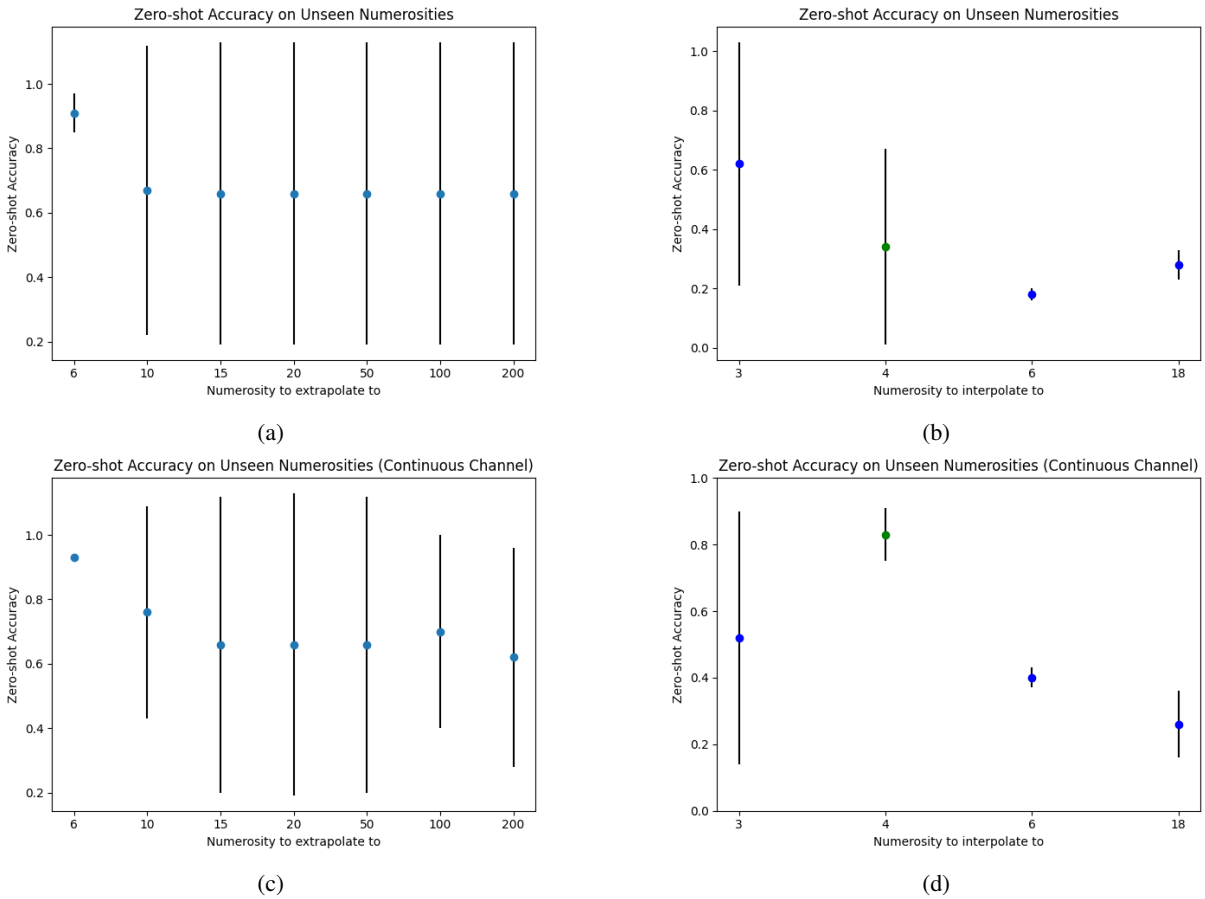


Figure 10: Zero-shot accuracies and standard deviations reported across 3 seeds for communication models tested on out-of-distribution data. (a) and (b) show the extrapolations and interpolation for the discrete channel, while (c) and (d) for the continuous (*i.e.* sketching) channel.

like human numeric systems, it does not follow an obvious compositional rule (*e.g.* the least frequent numerosities to have longer discrete and/or messages derived compositionally). For example, in the case of *Increase* frequency, numerosity 1 (which has the least variety in training) is usually encoded using the least number of tokens compared to 5 which has the most diversity in training.

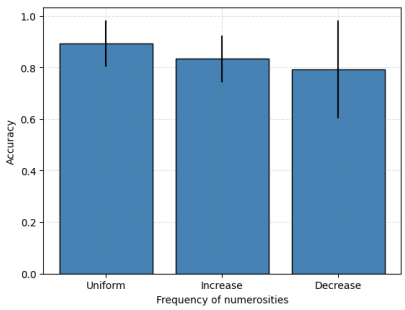
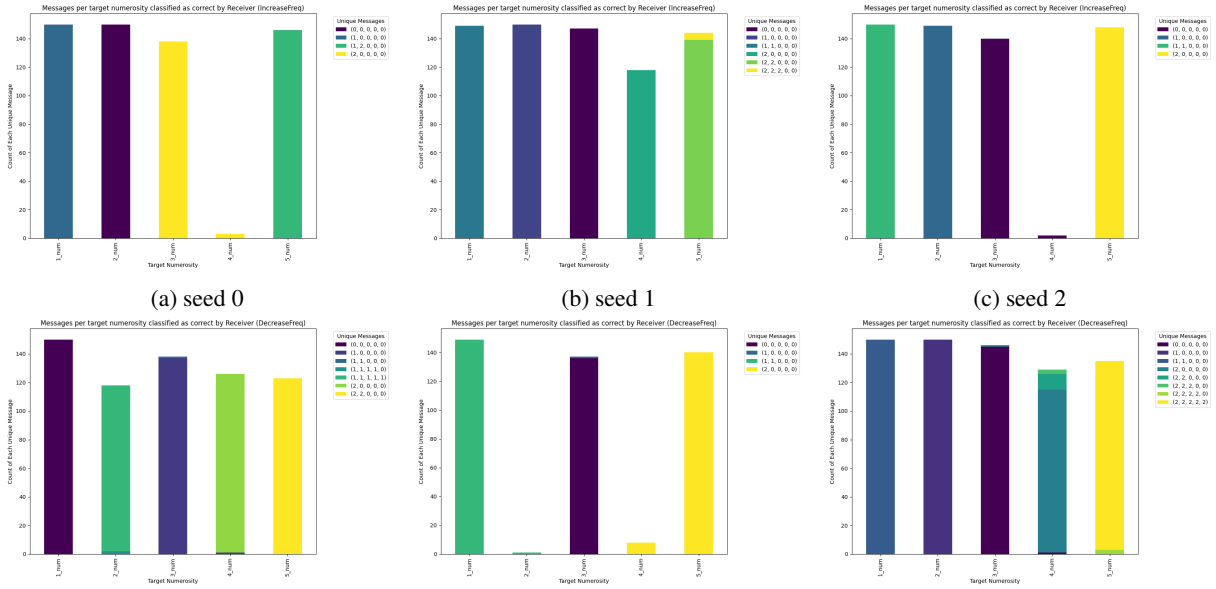


Figure 11: Accuracy and standard deviation across 3 seeds for communication task performed with different frequencies of training numerosities samples: *Uniform*, *Increase* and *Decrease* frequency setups.

D Examples of continuous communication through sketches

Further examples of sketches produced at test time for each of the numerosity classes present in the training set are shown in Figure 13. The different rows indicate results from agents being trained with more or less diverse training samples, some classes having more or fewer instances. Numerosity with a uniform distribution is used in most of our experiments and has equal training samples (700). *Increase* has the following distribution of samples per class:100,200,300,400,700 increasing from 1 to 5 and the reverse holds for *Decrease*. The accuracies across seeds are reported in Figure 8b and are similar in trend to those of the discrete channel (see Figure 11).

An observation from this experiment is that while the symbols do not vary in complexity based on the frequency nor are they compositional (in the sense of combining more frequent numerosities' codes), they all seem to correlate with the area



(a) seed 0 (b) seed 1 (c) seed 2

(d) seed 0 (e) seed 1 (f) seed 2

Figure 12: Exact message to numerosity mappings for correctly classified numerosities, across 3 different seeds, in the *Increase* (a, b and c) and *Decrease* frequency setups (d, e, f).

filled by elements. For example, numerosity 1 is usually represented through a symbol whose lines are concentrated at the centre of the sketch and do not cover much of the available space. As the numerosity increases and the elements start spreading across the target image space, the lines start increasing in length and are positioned in such a way as to cover as much of the available sketching space.

The pairwise dissimilarity of sketches in the three varying frequency setups is shown in Figure 14.

E Exact mapping sketching experiments

See one-to-many and many-to-one mappings for sketches in Figure 15 and the exact mappings in Figure 16.

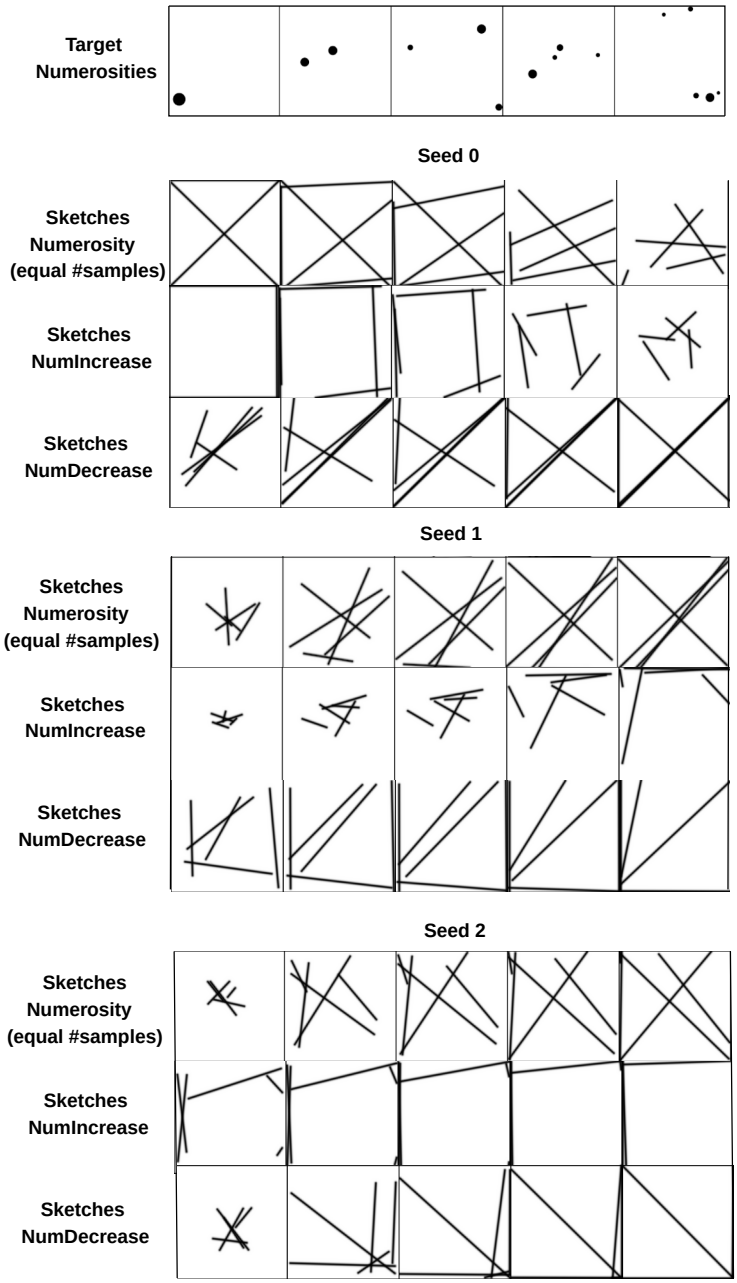
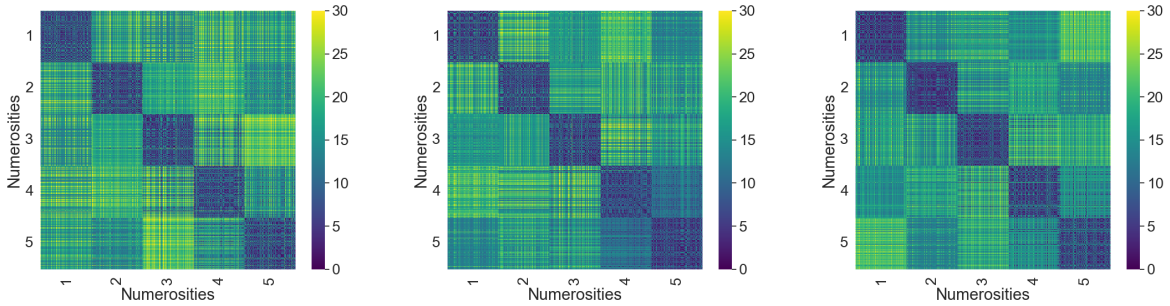


Figure 13: Test target images (each entry in the first row corresponds to a distinct numerosity between 1 and 5) and sketches produced by agents trained on different numerosity datasets, with varying numbers of class samples. Results are reported across 3 training seeds.



(a) Uniform (b) Increase (c) Decrease

Figure 14: Pair-wise dissimilarity for sketches in the tSNE-reduced pixel-space for various frequencies of numerosities: (a) equal, (b) increasing from 1 to 5, (c) decreasing from 1 to 5.

(a) (b)

Figure 15: (a) the proportion of many-to-one and (b) one-to-many relations of the target to sketch mappings across 3 runs (for the exact mappings see Figure 16).

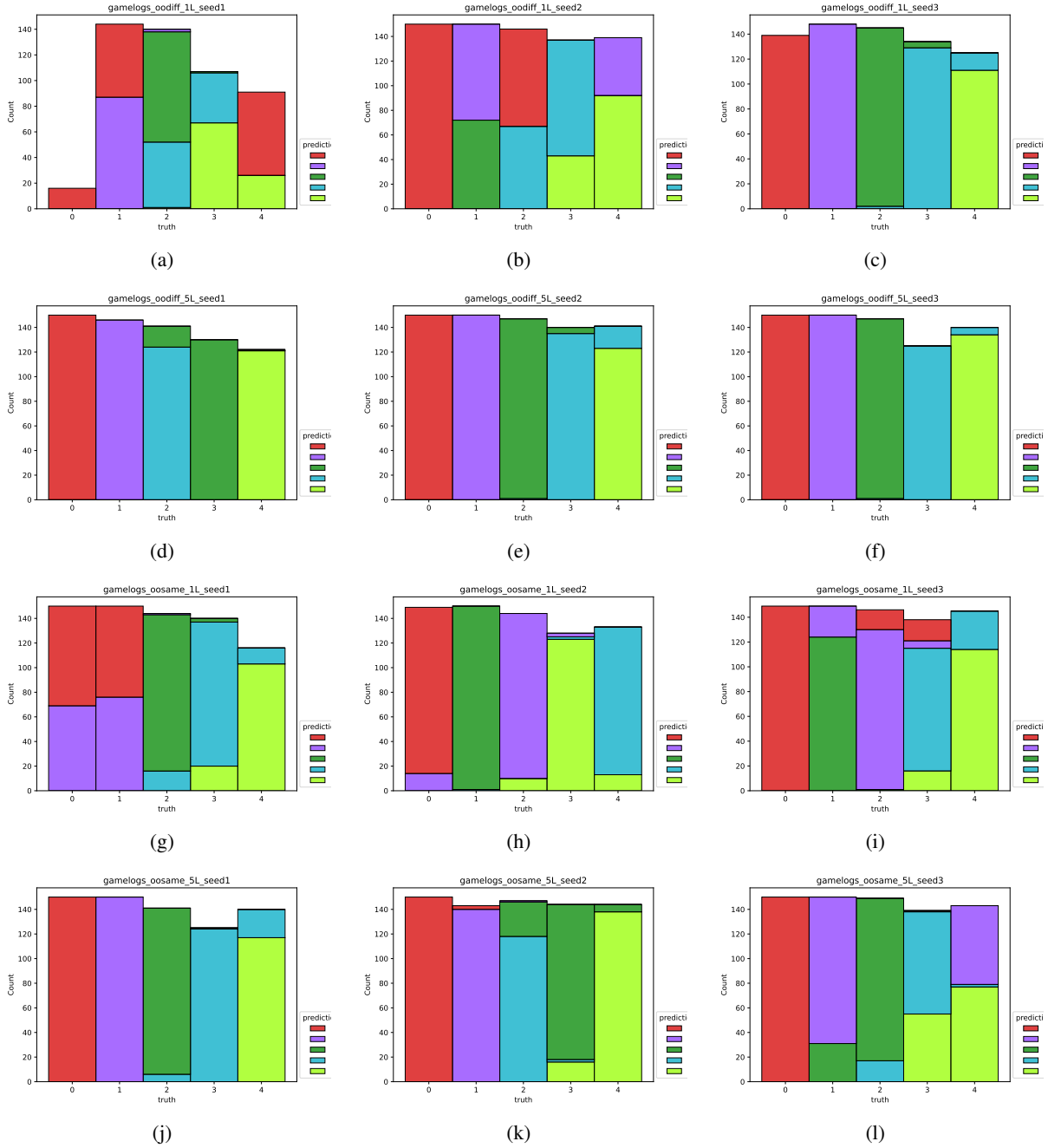


Figure 16: All exact mappings for sketching experiments