Supplementary Information: Computational constructivism: Developmental differences in active inductive inference

Neil R. Bramley (neil.bramley@ed.ac.uk)
Department of Psychology, University of Edinburgh, Scotland

Gwyneth Heuser
Psychology Department, University of California, Berkeley, USA

Fei Xu
Psychology Department, University of California, Berkeley, USA

Supplementary Information: Computational constructivism: Developmental differences in active inductive inference

This document contains supplementary details about the analyses in *Computational constructivism: Developmental differences in active inductive inference.*

Hypothesis generators

The paper considers two constructivist learning algorithms. We provide the full details for each of these here.

Generating context free (PCFG) model predictions

Here, we created a grammar (specifically a *probabilistic context free grammar* or PCFG; Ginsburg, 1966) that can be used to produce any rule that can be expressed with first-order logic and lambda abstraction. The grammatical primitives are detailed in Table 1.

Table 1
A Concept Grammar for the Task

Meaning	Expression
There exists an x_i such that...	$\exists(\lambda x_i :, \mathcal{X})$
For all x_i ...	$\forall(\lambda x_i : ., \mathcal{X})$
There exists {at least, at most, exactly} N objects in x_i such that...	$N_{\{<,>,=\}}(\lambda x_i : ., N, \mathcal{X})$
Feature f of x_i has value {larger, smaller, (or) equal} to v	$\{<, >, \leq, \geq, =\}(x_i, v, f)$
Feature f of x_i is {larger, smaller, (or) equal} to feature f of x_j	$\{<, >, \leq, \geq, =\}(x_i, x_j, f)$
Relation r between x_i and x_j holds	$\Gamma(x_i, x_j, r)$
Booleans {and,or,not}	$\{\wedge, \vee, \neq\}(x)$
Object feature	**Levels**
Color	{red, green,blue}
Size	{1:small, 2:medium, 3:large}
x-position	(0,8)
y-position	(0,8)
Orientation	{Upright, left hand side, right hand side, strange}
Grounded	true if touching the ground
Pairwise feature	**Condition**
Contact	true if x_1 touches x_2
Stacked	true if x_1 is above and touching x_2 and x_2 is grounded
Pointing	true if x_1 is orientated {left/right} and x_2 is to x_1s {left/right}
Inside	true if x_1 is smaller than x_2 + has same x and y position ($\pm$0.3), false

Note that $\{<, >, \geq, \leq\}$ comparisons only apply to numeric features (e.g., size).

There are multiple ways to implement a PCFG. Here we adopt a common approach to set up a set of string-rewrite rules (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Thus, each hypothesis begins life as a string containing a single *non-terminal symbol* (here, S) that is replaced using rewrite rules, or *productions*. These productions are repeatedly

applied to the string, replacing non-terminal symbols with a mixture of other non-terminal symbols and terminal fragments of first order logic, until no non-terminal symbols remain. The productions are so designed that the resulting string is guaranteed to be a valid grammatical expression and all grammatical expressions have a nonzero chance of being produced. In addition, by having the productions tie the expression to bound variables and truth statements, our PCFG serves as an automatic concept generator. Table 2 details the PCFG we used in the paper.

We use capital letters as non-terminal symbols and each rewrite is sampled from the available productions for a given symbol.[1] Because some of the productions involve branching (e.g., $B \to H(B, B)$), the resultant string can become arbitrarily long and complex, involving multiple boolean functions and complex relationships between bound variables.

We include a variant that samples uniformly from the set of possible replacements in each case, but we also reverse engineer a set of productions that produce exactly the statistics the empirical samples, as described in the main text.

Table 2
Prior Production Process

Production	Symbol	Replacements→		
Start	$S \to$	$\exists(\lambda x_i{:}\ A, \mathcal{X})$	$\forall(\lambda x_i{:}\ A, \mathcal{X})$	$N_I(\lambda x_i{:}\ A, K, \mathcal{X})$
Bind additional	$A \to$	B	S	
Expand	$B \to$	C	$J(B, B)$	$\neg(B)$
Function	$C \to$	$=(x_i, D1)$	$I(x_i, D2)$	$=(x_i, x_j, E1)^{\mathbf{a}}$
		$I(x_i, x_j, E2)^{\mathbf{a}}$	$\Gamma(x_i, x_j, E3)^{\mathbf{a}}$	
Feature/value	$D1 \to$	value,	feature	
(numeric only)	$D2 \to$	value,	feature	
Feature	$E1 \to$	feature		
(numeric only)	$E2 \to$	feature		
(relational)	$E3 \to$	feature		
Boolean	$J \to$	$\wedge$	$\vee$	...
Inequality	$I \to$	$\leq$	$\geq$	$>$
		$<$		
Number	$K \to$	$n \in \{1, 2, 3, 4, 5, 6\}$		

Note: Context-sensitive aspects of the grammar: [a]Bound variable(s) sampled uniformly without replacement from set; expressions requiring multiple variables censored if only one.

[1] The grammar is not strictly context free because the bound variables (x_1, x_2, etc.) are automatically shared across contexts (e.g. x_1 is evoked twice in both expressions generated in Figure 2a). We also draw feature value pairs together and conditional on the type of function they inhabit, to make our process more concise, however the same sampling is achievable in a context free way by having a separate function for every feature value, i.e. "'isRed()" and sampling these directly (c.f. **?**).

Generating instance driven (IDG) model predictions

We used the algorithm proposed in Bramley, Rothe, Tenenbaum, Xu, and Gureckis (2018) to produce a sample of 10,000 "grounded hypotheses" for each participant and trial, splitting these evenly across the 8 learning scenes that participant produced and tested.

To generate hypotheses as candidates for the hidden rule, the model uses the following procedure with probabilities either set to uniform or drawn from the PCFG-fitted productions for adults or for children (Figure 3gh in main text) and denoted with square brackets:

1. **Observe.** either:

   (a) With probability $[A \to B]$: Sample a cone from the observation, then sample one of its features f with probability $[G \to f]$ — e.g., {#1}:[2] "medium, size" or {#3}: "red, color".

   (b) With probability $[A \to \text{Start}]$: Sample two cones uniformly without replacement from the observation, and sample any shared or pairwise feature — e.g., {#1,#2}: "size", or "contact"

2. **Functionize.** Bind a variable for each sampled cone in Step 1 and sample a true (in)equality statement relating the variable(s) and feature:

   (a) For a statement involving an unordered feature there is only one possibility — e.g, {#3}: "$= (x_1, \text{red}, \text{color})$", or for {#1,#2}: "$=(x_1, x_2, \text{color})$"

   (b) For a single cone and an ordered feature, this could also be a nonstrict inequality ($\geq$ or $\leq$). We assume a learner only samples an inequality if it expands the number of cones picked out from the scene relative to an equality — e.g., in Figure 2b in the main text, there is also a large cone {#1} so either $\geq(x_1, \text{medium}, \text{size})$ or $=(x_1, \text{medium}, \text{size})$ might be selected with uniform probability.

   (c) For two cones and an ordered feature, either strict or non-strict inequalities could be sampled if the cones differ on the sampled feature, equivalently either equality or non-strict inequality could be selected if the cones do not differ on that dimension — e.g., {#1,#2} $>(x_1, x_2, \text{size})$, or {#3,#4} $\geq(x_1, x_2, \text{size})$. In each case, the production weights from Figure **??** for the relevant completions are normalized and used to select the option.

[2] Numbers prepended with # refer to the labels on the cones in the example observation in Figure **??**b.

3. **Extend.** With probability $\frac{[B\to D]}{[B\to D]+[B\to C(B,B)]}$ go to Step 4, otherwise sample a conjunction with probability $[C(B,B)\to \text{And}]$ or a disjunction with probability $[C(B,B)\to \text{Or}]$ and repeat. For statements with two bound variables, Step 3 is performed for x_1, then again for x_2:

   (a) **Conjunction.** A cone is sampled from the subset picked out by the statement thus far and one of its features sampled with probability $[G\to f]$ — e.g., $\{\#1\}$ $\wedge(=(x_1,\text{green},\text{color}),\geq(x_1,\text{medium},\text{size}))$. Again, inequalities are sample-able only if they increase the true set size relative to equality — e.g., "$\wedge(\leq(x_1,3,\text{xposition}),\geq(x_1,\text{medium},\text{size}))$", which picks out more objects than "$\wedge(=(x_1,3,\text{xposition}),\geq(x_1,\text{medium},\text{size}))$".

   (b) **Disjunction.** An additional feature-value pair is selected uniformly from *either* unselected values of the current feature, *or* from a different feature — e.g., $\vee(=(x_1,\text{color},\text{red}),=(x_1,\text{color},\text{blue}))$ or $\vee(=(x_1,\text{color},\text{blue}),\geq(x_1,\text{size},2))$. This step is skipped if the statement is already true of all the cones in the scene.[3]

4. **Flip.** If the inspiration scene is not rule following wrap the expression in a $\neg()$.

5. **Quantify.** Given the contained statement, select true quantifier(s):

   (a) For statements involving a single bound variable (i.e., those inspired by a single cone in Step 1) the possible quantifiers simply depend on the number of the cones in the scene for which the statement holds. If the statement is true of all cones in the scene Quantifier is selected using probabilities [Start$\to$] combined with $[L\to]$ where appropriate. If it is true of only a subset of the cones then $\forall(\lambda x_i:A,\mathcal{X})$ is censored and the probabilities re-normalized. K is set to match number of cones for which the statement is true.

   (b) Statements involving two bound variables in lambda calculus have two nested quantifier statements each selected as in (a). The inner statement quantifying x_2 is selected first based on truth value of the expression while taking x_1 to refer to the cone observed in '1.'. The truth of the selected inner quantified statement is then assessed for all cones to select the outer quantifier — e.g., $\{\#3,\#4\}$ "$\wedge(=(x_2,\text{green},\text{color}),\leq(x_1,x_2,\text{size}))$" might become "$\forall(\lambda x_1:\ \exists(\lambda x_2:\ \ \wedge(=(x_2,\text{green},\text{color}),\leq(x_1,x_2,\text{size})),\mathcal{X}),\mathcal{X})$". The inner quantifier $\exists$ is selected (three of the four cones are green $\{\#1, \#2, \#4\}$), and

[3] We rounded positional features to one decimal place in evaluating rules to allow for perceptual uncertainty.

the outer quantifier ∀ is selected (all cones are less than or equal in size to a green cone).

Note that a procedure like the one laid out above is, in principle, capable of generating any rule generated by the PCFG in Figure

One way to think of the IDG procedure is as a partial inversion of a PCFG. As illustrated by the blue text in the examples in Figure 2b in the main text. While the PCFG starts at the outside and works inward, the IDG starts from the central content and works outward out to a quantified statement, ensuring at each step that this final statement is true of the scene.

Free response coding

Coding the free responses

To analyze the free responses, we first had two coders go through all responses and categorize them as either:

1. Correct: The subject gives exactly the correct rule or something logically equivalent

2. Overcomplicated: The subject gives a rule that over-specifies the criteria needed to produce stars relative to the ground truth. This means the rule they give is logically sufficient but not necessary. For example, stipulating that "there must be a small red" is overcomplicated if the true rule is "there must be a red" because a scene could contain a medium or large red and emit stars.

3. Overliberal: The opposite of overcomplicated. The subject gives a rule that under-specifies what must happen for the scene to produce stars. For example, stipulating that "there must be a blue" if the true rule is that "exactly one is blue". This is logically necessary but not sufficient because a scene could contain blue objects but not produce stars because there is not exactly one of them.

4. Different: The subject gives a rule that is intelligible but different from the ground truth in that it is neither necessary or sufficient for determining whether a scene will produce stars.

5. Vague or multiple. Nuisance category.

6. No rule. The subject says they cannot think of a rule.

We were able to encode 205/238 (86%) of the children's responses and (219/250) 87% for adults as correct, overcomplicated, overliberal or different. Table 3 shows the complete confusion matrix. The two coders agreed 85% of the time, resulting in a Cohen's Kappa of .77 indicating a good level of agreement (Krippendorff, 2012).

Table 3
Agreement Matrix for Independent Coders' Free Response Classifications

	correct	overliberal	overspecific	different	vague	no rule	multiple
correct	**93**	1	5	0	0	0	0
overliberal	5	**13**	1	8	0	1	0
overspecific	1	2	**42**	12	0	0	0
different	0	5	3	**224**	15	3	0
vague	0	1	2	3	**11**	6	0
no rule	0	0	0	0	0	**31**	0
multiple	0	1	0	2	0	0	**0**

We then had one coder familiar with the grammar go through each free response that was not assigned vague or no rule, and encode it as a function in our grammar. The second coder then blind spot checked 15% of these rules (64) and agreed in 95% of cases 61/64. The 6 cases of disagreement were discussed and resolved. In 5/6 cases, this was in favor of the primary coder. The full set of free text responses along with the requisite classification, encoded rules are available in the Online repository.

Scene similarity measurement

To establish the overall similarity between two scenes, we need to map the objects in a given scene to the objects in another scene (for example between the scenes in Figure1 a and b) and establish a reasonable cost for the differences between objects across dimensions. We also need a procedure for cases where there are objects in one scene that have no analogue in the other. We approach the calculation of similarity via the principle of minimum edit distance (Levenshtein, 1966). This means summing up the elementary operations required to convert scene (a) into scene (b) or visa versa. We assume objects can be adjusted in one dimension at a time (i.e. moving them on the x axis, rotating them, or changing their color, and so on.

Before focusing on how to map the objects between the scenes we must decide how to measure the adjustment distance for a particular object in scene a to its supposed analogue in scene b. As a simple way to combine the edit costs across dimensions we first Z-score each dimension, such that the average distance between any two values across all objects and all scenes and dimensions is 1. We then take the L1-norm (or city block

distance) as the cost for converting an object in scene (a) to an object in scene (b), or visa versa. Note this is sensitive the size of the adjustment, penalizing larger changes in position, orientation or size more severely than smaller changes, while changes in color are all considered equally large since color is taken as categorical. Note also that for orientation differences we also always assume the shortest distance around the circle.

If scene (a) has an object that does not exist in scene (b) we assume a default adjustment penalty equal to the average divergence between two objects across all comparisons (3.57 in the current dataset). We do the same for any object that exists in (a) but not (b).

Calculating the overall similarity between two scenes involves solving a mapping problem of identifying which objects in scene (a) are "the same" as those in scene (b). We resolve this "charitably", by searching exhaustively for the mapping of objects in scene (a) to scene (b) that minimizes the total edit distance. Having selected this mapping, and computed the final edit distance including any costs for additional or removed objects, we divide by the number shared cones, so as to avoid the dissimilarities increasing with the number of objects involved.

Figure 2 computes the inter-scene similarity components that go into Figure 6c in the main text. Summing up the edit distances across all objects, children's scenes seem much more diverse than adults (Figure 2a). However this is primarily due to their containing a greater average number of objects. Scaling the edit distance by the number of objects in the target scene gives a more balanced perspective (Figure 2b) but does not account for the fact that the compared scene may contain more or fewer objects in total. Figure 2c visualizes just the object difference showing that children's scenes contain roughly as many objects on average as the initial example while adults' scenes contain around 0.75 fewer objects than are present in the initial example (dark shading in top row). Thus, we opted to combine b and c by weighting the unsigned cone difference by the mean inter-object distance across all comparisons to give our combined distance measure (Figure 2d and Figure 6c in the main text).

Information gain analysis of active learning data

Children's and adults' scene generation patterns manifest in small differences in the quality of the total evidence generated according to an information gain analysis. For example, using the unweighted PCFG sample, prior entropy is 13.31 bits and children's evidence produces an information gain (reduction in uncertainty) of 6.86±0.55 bits while adults data allows for marginally higher information 7.04±0.44 bits $t(102) = -1.8, p = 0.068$. Relative to the fitted PCFG priors, the difference in information

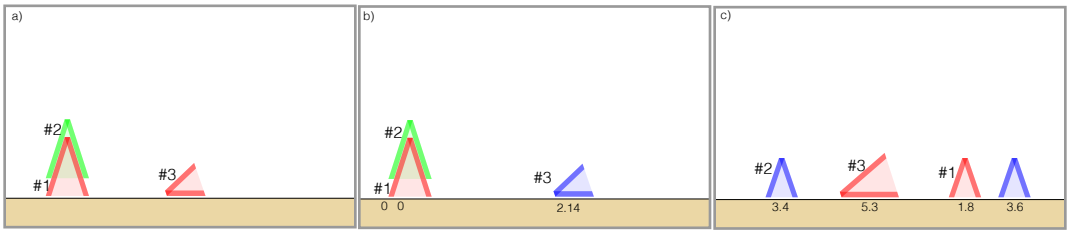

Figure 1

Three example scenes. Objects indices link the most similar set of objects in b to those in a. Numbers below indicate the edit distance for each object (i.e. the sum of scaled dimension adjustments). Intuitively scene a) is more similar to scene b) than to scene c) and this is reflected in the similarity scores.

gains is rather larger, with children's scenes leading to information gain at 6.92±0.70 bits (prior entropy 12.94), and adults' at 7.50±0.66 (prior entropy 12.65) $t(102) = 4.4, p < .0001$. Under the mismatched priors — that is, taking the adultlike PCFG prior for children and childlike PCFG prior for adults — children's tests look slightly more informative than under their own prior, generating 7.14±0.72bits, and adults' tests slightly less informative than under their own prior 7.21±0.61 bits, eliminating the statistical difference $t(102) = 0.5, p = 0.62$. On the face of it, this is evidence against the idea that children's more elaborate hypothesis generation and concomitantly flatter latent prior is driving them rationally toward more elaborate testing patterns. However, we see this information-theoretic analyses as limited in what reveals. This is because is predicated on an implausibly complete representation of uncertainty, e.g. using a large sample of prior hypotheses, while we might expect constructivist search behavior to be driven by more focal testing of a smaller number of possibilities. Nevertheless, we present these information scores as norms for completeness and comparison with other active learning tasks.

Generalization models

As described in main text, we fit 18 model variants to participant's data. All models have between 0 and 2 parameters. For each model, we fit the parameter(s) by maximizing the model's likelihood of producing the participant data, using R's `optim` function. We compare models using the Bayesian Information Criterion (Schwarz, 1978) to accommodate their different numbers of fitted parameters.[4]

[4] On one perspective, our derivation of the child-like and adult-like productions constitutes fitting an additional 39 parameters ($m - 1$ for each production step), so evoking an additional BIC parameter penalty of $39 \times \log(3940) = 323$ for PCFG over PCFG Uniform and similarly for the IDG. If we were to apply this penalty, the uniform weighted variants would be clearly preferred under the BIC criterion at the aggregate level. It is less clear how to apply this penalty at the individual level. We chose to include the

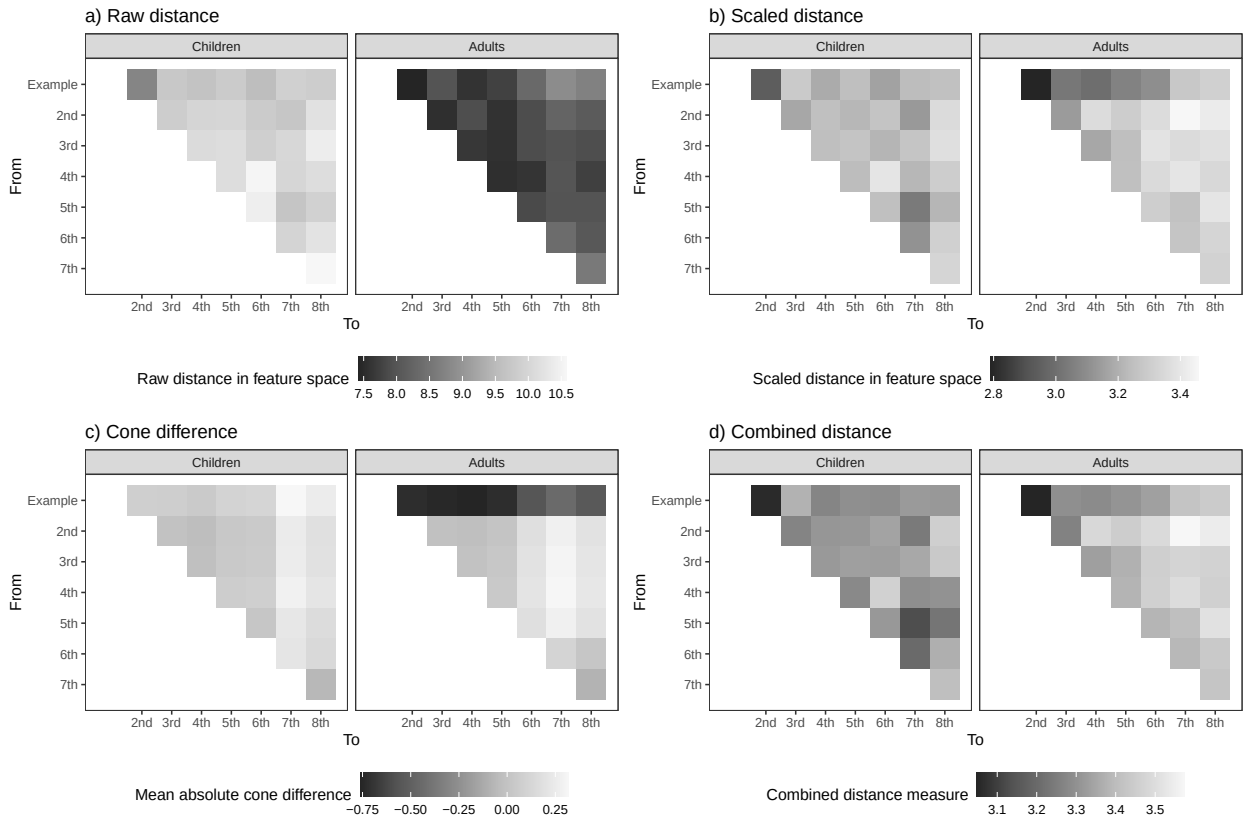

Figure 2

a) The average minimum edit distance summed up across shared objects. b) Rescaling a by dividing by the number of objects. c) The penalty for additional or omitted objects. d)Combined distance as in main text.

Comparison with Bramley et al (2018) dataset

Finally, for interest and to demonstrate replication of our core results. We provide a direct comparison between the generalization accuracies in the current sample of children and adults and those in the sample of 30 adults modelled in (Bramley et al., 2018). Bramley et al (2018) included 10 ground truth concepts, and the current paper uses just the first five of these. Figure 3 shows these accuracy patterns side by side revealing the adults in the current experiment performed approximately as well as those in the original conference paper.

---

fitted versions alongside the uniform versions here without penalty as demonstrations of the differences that arise from different generation probabilities.

Table 4
Models of Participants' Generalizations

	Model	Group	log(Likelihood)	BIC	λ	τ	N	Accuracy
1.	Baseline	children	-1319.75	2639.50			7	50%
2.	Encoded Guess	children	-1143.69	2294.92		0.98	15	62%
3.	Similarity	children	-1316.44	2640.42		-0.50	0	41%
4.	PCFG Uniform	children	-1319.75	2647.05		-0.01	0	60%
5.	PCFG Off	children	-1318.85	2645.26		0.09	0	65%
6.	PCFG	children	-1319.57	2646.69		0.04	1	63%
7.	IDG Uniform	children	-1299.72	2607.00		0.55	2	66%
8.	IDG Off	children	-1304.92	2617.39		0.45	1	**70%**
9.	IDG	children	-1308.52	2624.59		0.39	2	68%
10.	Intercept	children	-1218.96	2445.47	0.32		**16**	50%
11.	**Encoded Guess + Intercept**	**children**	**-1067.18**	**2149.47**	0.26	1.24	9	
12.	Similarity + Intercept	children	-1214.71	2444.52	0.32	-0.77	1	
13.	PCFG Uniform + Intercept	children	-1210.30	2435.70	0.35	0.43	0	
14.	PCFG Off + Intercept	children	-1207.64	2430.39	0.34	0.48	0	
15.	PCFG + Intercept	children	-1208.74	2432.59	0.35	0.46	0	
16.	IDG Uniform + Intercept	children	-1195.19	2405.48	0.32	0.83	0	
17.	IDG Off + Intercept	children	-1193.01	2401.12	0.34	0.83	0	
18.	IDG + Intercept	children	-1194.19	2403.49	0.34	0.82	0	
1.	Baseline	adults	-1386.29	2772.59			2	50%
2.	Encoded Guess	adults	-893.49	1794.58		1.78	**32**	70%
3.	Similarity	adults	-1359.05	2725.70		-1.38	0	36%
4.	PCFG Uniform	adults	-1333.95	2675.50		0.69	0	62%
5.	PCFG Off	adults	-1293.60	2594.79		0.94	1	66%
6.	PCFG	adults	-1267.89	2543.38		1.06	2	69%
7.	IDG Uniform	adults	-1229.69	2466.97		1.50	2	69%
8.	IDG Off	adults	-1208.11	2423.83		1.52	0	73%
9.	IDG	adults	-1185.64	2378.89		1.62	1	**74%**
10.	Intercept	adults	-1364.90	2737.40	0.15		6	50%
11.	**Encoded Guess + Intercept**	**adults**	**-880.59**	**1776.38**	0.08	2.01	4	
12.	Similarity + Intercept	adults	-1337.55	2690.30	0.14	-1.63	0	
13.	PCFG Uniform + Intercept	adults	-1268.87	2552.93	0.26	1.35	0	
14.	PCFG Off + Intercept	adults	-1226.61	2468.42	0.25	1.60	0	
15.	PCFG + Intercept	adults	-1203.66	2422.53	0.24	1.69	0	
16.	IDG Uniform + Intercept	adults	-1179.02	2373.24	0.20	2.13	0	
17.	IDG Off + Intercept	adults	-1147.46	2310.13	0.22	2.26	0	
18.	IDG + Intercept	adults	-1131.92	2279.04	0.20	2.28	0	

NB: Accuracy column shows performance of the requisite model across 100 simulated runs through the task using participants active learning data and τ set to 100 (essentially hard maximizing over the model's predictions). The +Intercept models perform strictly worse due to their bias so are not included in this column.

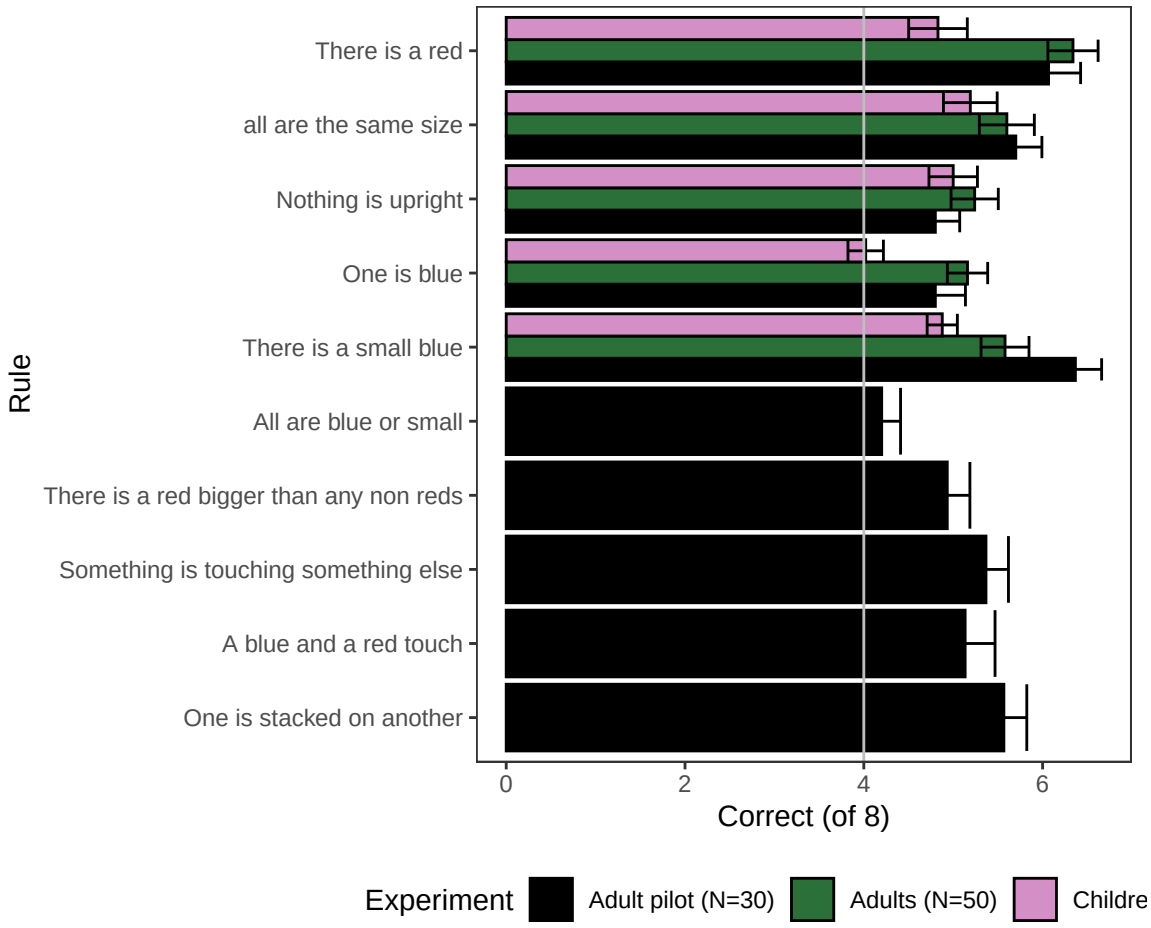

Figure 3

Generalization accuracy by number of objects per test scene comparing with 10 rule adult pilot from Bramley et al. (2018).

References

Bramley, N. R., Rothe, A., Tenenbaum, J. B., Xu, F., & Gureckis, T. M. (2018). Grounding compositional hypothesis generation in specific instances. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society.* Austin, TX: Cognitive Science Society.

Ginsburg, S. (1966). *The mathematical theory of context free languages.* McGraw-Hill Book Company.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology.* Sage.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.