# How Tall Is *Tall*? Compositionality, Statistics, and Gradable Adjectives

**Lauren A. Schmidt[1] (lschmidt@mit.edu), Noah D. Goodman[1] (ndg@mit.edu), David Barner[2] (barner@ucsd.edu), and Joshua B. Tenenbaum[1] (jbt@mit.edu)**
[1]Department of Brain and Cognitive Sciences, MIT;
[2]Department of Psychology, University of California, San Diego

**Abstract**

What is *tall*? Like many words, *tall* is fundamentally compositional in meaning — whether an item is tall depends on the statistical properties of the set of items it is being compared to. Despite preliminary evidence to this effect, no mathematical models of how *tall* operates in a given context have ever been empirically evaluated. We compare a number of statistical models to adults and children in judging which items are tall in various contexts, including both threshold-based and categorization-based models. We find that non-parametric statistical models of gradable adjectives best describes the judgments of people across a wide variety of contexts.

**Keywords:** psychology; language understanding; Bayesian modeling

## Introduction

Skyscrapers, giraffes, and basketball players all belong to a single category, despite having almost nothing in common. In fact, the one property that they do share — being *tall* — is not true of all three when considered together. Basketball players are short relative to skyscrapers, and even compared to giraffes. So how do we – or better yet, 3-year-old children – ever figure out what these things have in common? For that matter, what do tall things have in common?

The answer to this problem has two components, the first of which is old, and was most famously defended by Frege (1892). This is the observation that noun phrases – and expressions of language more generally – are subject to compositionality: the meanings of complex expressions are a function of their syntax and the meanings of their constituent parts (see Fodor & Lepore, 2002, for discussion). In the case of an expression like *tall boy*, the interpretation of *tall* is determined in part by the meaning of the noun it modifies – i.e., *boy*. *Tall*, a subsective adjective, picks out a subset of things referred to by the noun it modifies ($\|tall\| \subseteq \|boy\|$) unlike intersective adjectives, like *Californian*, which pick out the intersection between two sets ($\|Californian\| \cap \|boy\|$). As a result, it makes sense to ask whether someone is tall for a boy, but not whether he is Californian for a boy. *Tall* is interpreted relative to the class of things denoted by the noun it modifies, or the *comparison set*.

The compositional nature of *tall* explains why it can be true of both humans and towers. However, it does not explain how we know which humans and which towers are tall for their respective kinds. What, precisely, does it mean to be tall for a building?

Past work from both psychology and linguistics falls short of answering this question. Whereas theories of concepts have traditionally been concerned with determining how mental representations encode properties of things in the world (e.g., via definitions, prototypes, inferential roles, or causal histories), previous studies have restricted these investigations almost entirely to nominal categories (e.g., natural kinds, artifact kinds), or to stable perceptual properties (e.g., color, shape, texture). Although previous studies suggest that young children are sensitive to the statistics of sets when using gradable adjectives like *tall* and *high* (e.g., Smith, Cooney, & McCord, 1986; Barner & Snedeker, 2008) none of these studies has characterized what type of statistical function is used by children in understanding these terms. Considerable attention has been paid to the formal semantics of gradable adjectives (Cresswell, 1976; Kennedy, 1999; Klein, 1991), but these accounts do not specify how speakers identify things in the world as *tall, short, big,* or *small*.

Although no previous study has investigated how classes of things are divided into gradable categories like *big* and *tall*, some suggestions have been made. For example, things may count as *tall* if they are taller than average, or taller than most things in a class (Barner & Snedeker, 2008).

In this paper we compare the performance of a number of possible models of *tall* to the tallness judgments of people. These models include both models with an absolute standard of comparison (included as a baseline for model performance), and those which perform simple statistical functions to determine the standard of comparison for a given set of items. We additionally explore an alternative model of *tall* that combines statistics and categorization, using Bayesian methods to probabilistically cluster items (e.g., buildings) into subcategories (e.g., tall buildings) based on their heights. The standard of comparison for such a model is the boundary between the tallest subcategory of items and all shorter subcategories. Experiments 1 and 2 explore the performance of all of these models in a number of different contexts.

## Models of *tall*

We considered several possible models of gradable adjectives based on the context. We describe the models with regards to the meaning of *tall*, but each of the models applies directly to any gradeable adjective.

For all model definitions, let $C$ be the set of objects in the context, and $h : C \rightarrow \mathbb{R}$ be the height function from objects to their height[1].

Each model below defines a probability $P(x \text{ is tall}|C)$. That is, each gives an alternative way to answer the question "is object $x$ tall, given the context?"

[1]To apply to another gradeable adjective, replace the height function $h$ with an appropriate function mapping from items to the scale of the adjective.

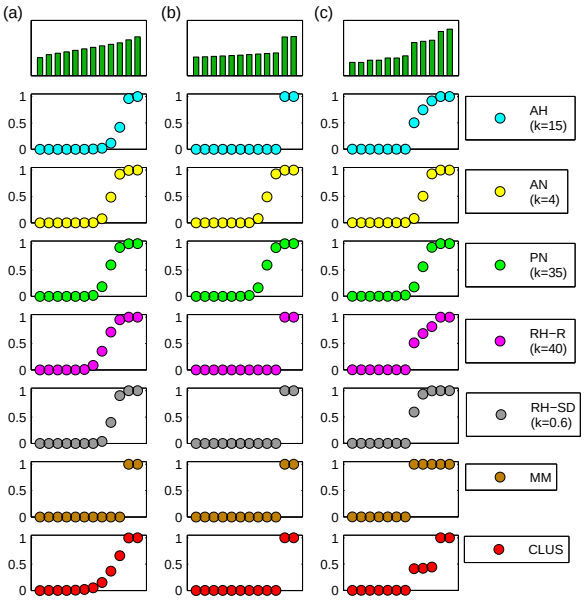


Figure 1: A distribution of rectangular items of different heights is shown at the top. The graphs below show the probability that each item is tall, according to each model. Model name and parameter setting specified to right of each graph. (Cluster model shown uses typical hyperparameter settings.)

We considered two families of models. The first family is a wide class in which a height threshold is computed from the context, and tallness decisions are made based on this threshold. To the best of our knowledge, all previous proposals for the meaning of gradeable adjectives fall into this family.

The second class treats tallness judgement as a categorization problem and invokes the machinery of probabilistic clustering.

## Threshold-based models

Threshold-based models compute a *threshold statistic* from the heights of the objects in the context, and make tallness judgements by comparing to this threshold.

We include normally-distrubuted noise on the threshold[2] in order to permit vague use of "tall." Let $T$ be a threshold function mapping from context sets $C$ to a positive real number. The probability of item $x$ being tall is the cumulative probability that a normal random variate with mean $T(C)$ is less than $h(x)$:

$$P(x \text{ is tall}|C) = \frac{1}{2}\left[1+\text{erf}\left(\frac{h(x)-T(C)}{\varepsilon\sqrt{2}}\right)\right],$$

where erf is the error function, and $\varepsilon$ a noise-width parameter.

We consider a number of possible threshold functions:

- *Absolute height (AH)*: all items taller than a fixed reference height are tall: $T(C) = k$, with $k$ a fixed parameter.
- *Absolute number (AN)*: the tallest $k$ items are tall. Let $x_1,...,x_N$ be an ordering of $C$ by height (i.e. $h(x_i) \leq h(x_j)$ if $i < j$). Then: $T(C) = h(x_k)$.
- *Unique percent number (UAN)*: as in AN, but computed based on only one object of each height.
- *Percent number (PN)*: the tallest $k\%$ of the items are tall: $T(C) = h(x_{\lfloor k \cdot N \rfloor})$.
- *Unique percent number (UPN)*: as in PN, but computed based on only one object of each height.
- *Relative height by range (RH-R)*: any item within the top $k\%$ of the range of heights is tall. If we write $Mx = max_{x\in C}h(x)$ and $Mn = min_{x\in C}h(x)$, then: $T(C) = Mx - k \cdot (Mx - Mn)$.
- *Relative height by standard deviation (RH-SD)*: any item with a height greater than $k$ standard deviations above the mean is tall. If we write $\bar{x}$ and $\sigma$ for the mean and standard deviation of heights of object in $C$, then: $T(C) = \bar{x} + k \cdot \sigma$.
- *Maximum margin (MM)*[3]: items taller than the biggest gap in heights are tall: $T(C) = h(argmax_{x\in C} min_{y\in C s.t.\ h(y)<h(x)} |h(x) - h(y)|)$.

## Category-based models

Category-based models classify an element as tall if it is in the "tall category." The same psychological mechanism used for other categorization tasks can be used to determine which subset of the items are in the tall category, but subject to the constraint that the *tallest* item (i.e. $argmax_{x\in C}h(x)$) is in the tall category. For example, the max-margin threshold model, described above, can be seen as a primitive category-based model: it first separates the items into clusters (determined by the largest gap in heights), then sets the threshold in order to divide these clusters.

Next consider a more sophisticated category-based model, related to the infinite Gaussian mixture model (Rasmussen, 2000), a modern version of Anderson's rational model of categorization (Anderson, 1991); we'll refer to this as the *Cluster model (CLUS)*. Let $Q$ be a partition of $C$ into clusters of contiguous items based on height, and write $q_x$ for the partition containing $x$. First, an item is tall if it is in the same cluster as the tallest item (call this cluster $q_{tall}$). The shortest item is required to be in a separate category from the tallest item, based on the idea that people expect the items in any given context to range from "short" to "tall" within that context.

Thus the probability that a particular item is tall is given by:

$$P(x \text{ is tall}|C) = \sum_Q P(x \in q_{\text{tall}}|Q)P(Q|C)$$

The posterior probability of a particular partition is:

$$\begin{aligned} P(Q|C) &\propto P(C|Q)P(Q) \\ &= P(Q)\left(\prod_{x\in C} P(x|\mu_{q_x},\sigma_{q_x})\right)\left(\prod_{q\in Q} P(\mu_q)P(\sigma_q)\right), \end{aligned}$$

with the remaining probabilities determined by the model setup (see Rasmussen, 2000, for details): $P(x|\mu_{q_x},\sigma_{q_x})$ is

[2] This noise may result from within- or between-subjects factors.

[3] The noise parameter for the MM model is set to 0 to serve as a simple heuristic clustering model.

Gaussian, $P(\mu_q)$ is Gaussian, $P(\sigma_q)$ is Gamma, and $P(Q)$ is given by the Chinese restaurant process.

Intuitively, Bayesian inference using this model looks for clusters of items based on height. The model parameters set prior expectations about the mean and variance of these clusters, as well as the number of items in the clusters, but strong evidence can overcome these prior expectations.[4]

For each possible partition of items into clusters, only the tallest cluster of items is considered the "tall category"; all other items are not tall. To find the overall probability that an item is tall, we weight whether it is tall in each possible partition of items by the posterior probability of items being partitioned in that manner. The posterior probability is based on both prior expectations and the heights of the items.

### Model predictions

Figure 1 shows several different sample contexts $C$, and examples of model probability that each item is tall, for all of the models described above.

In the case of an approximately uniform distribution of items, Figure 1(a), all models show a sigmoid-like drop-off in probability in approximately the same place (with the exception of the MM model, which is a step function for reasons described above). Though all the items apparently fall into a single cluster here, the prior expectations of the CLUS model about the size of clusters, together with the requirement that the shortest item not be tall, lead the Cluster-based model to also show a sigmoid-like drop-off in determining which items are tall.

Figure 1(b) and (c) show two more distributions, with 2 or 3 apparent clusters of items respectively. Some of the threshold models separate the two clusters and make the same predictions in Figure 1(b) as the CLUS model, but others show the same sigmoid-like drop-off as before. In Figure 1(c), the CLUS model treats the items in each of the three clusters uniformly and distinctly from each of the other clusters, while all the other models show very different predictions. Note that varying either the parameter settings (which changes the location of the drop-off) or the noise parameter setting (which changes how gradual or sudden the drop-off is) for the threshold-based models can cause their predictions to change greatly for individual items, though they will always have a sigmoidal drop-off.

The major differences in threshold models are also apparent across the distributions, including some limitations: AH is entirely insensitive to context; AN is sensitive only to the height-ordering, not the actual heights; MM is determined solely by the largest gap in heights. Meanwhile the relative height models are more flexible, extracting useful statistics of the height distribution. In Experiment 1 we will use these differences to seek double-dissociations ruling out the less flexible models as an explanation of human judgements.

[4]Four model hyperparameters were varied in the following studies, one controlling the CRP prior, and three controlling the Normal-Gamma conjugate prior distribution for the cluster mean and variance.

## Experiment 1

In Experiment 1, we compared the performance of the models and people in judging which items were tall given distributions of items of different heights.

Adult subjects judged which items were tall in a wide variety of distributions of items. We generated distributions by sampling randomly or at regular intervals from one or two Gaussian, uniform, or other statistical distributions (such as an exponential distribution). The means, variances, and number of items within each of the resulting clusters of items varied. Representative distributions are shown in Figure 2(a).

181 adults participated. Each adult saw one distribution of items, all presented at the same time. Items were shown in a frame, like one of those in Figure 2(a), but they were shuffled into a pseudo-random order instead of being sorted by height. We labeled each set of items with a novel name like "pimwits," and then asked adults to specify which ones were the tall pimwits.

All models ran on the distributions with a wide variety of parameter settings.

**Results** For each distribution, we compared a given model's probability that each item was tall to the percent of adults who labeled that item tall using a measure called mean difference, which measured the average error per item. The method of calculating the mean difference is illustrated in Figure 2(c).

We selected the best parameter setting for each model by averaging the mean differences between the adult judgments and the model probabilities across all distributions. Figure 2 shows a histogram of the MD across all distributions for each of these best models. The models are ranked in order of average performance.

The Relative Height by Standard Deviation, Relative Height by Range, and Cluster models all perform approximately equally well. The Unique Percent Number and Percent Number models perform only slightly worse overall; however, there are some distributions where both of these models perform very poorly, as shown in Figure 3(a). For these two distributions we can see a strong *double dissociation* in performance; on the left, the PN model predicts far too many items are tall, and on the right, far too few (UPN has the same problem in many cases, since it reduces to the same model as PN whenever the distribution values are unique). Because of this double dissociation, the PN threshold parameter cannot be adjusted in either direction to better predict all the adult results, and it systematically fails to capture human judgments about what is tall.

Figure 3(b)-(d) also shows similar double dissociations for the AN, MM, and AH models. Figure 2 demonstrates these models' increasingly poor performance on the overall distribution set. Because all of these models fail in both directions – sometimes calling too few items tall and sometimes too many – these models also cannot be account for human judgments of tallness. This failure of the two absolute standard of comparison models (AN and AH) is consistent with

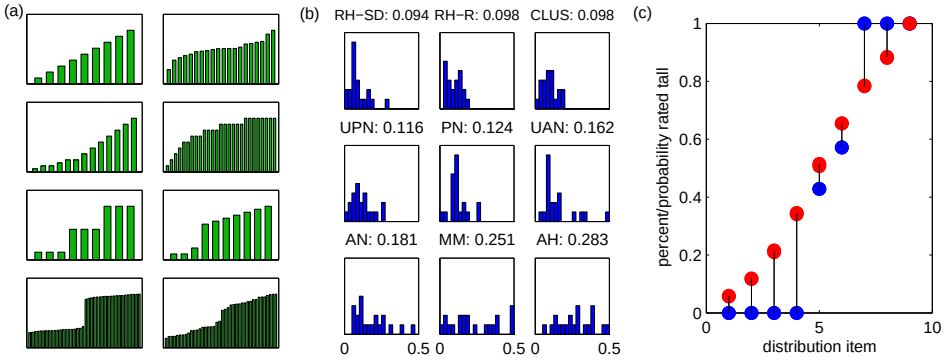


Figure 2: (a) A number of representative distributions used as stimuli in Experiment 1 (subjects saw these and other sets of objects in randomized order instead of sorted by height, as shown here). (b) Histograms demonstrating the performance of each model in Experiment 1, where success is measured along the x-axis in terms of mean difference (see (c) for explanation of measure), with a performance of 0 being best and 1 being worst (for these histograms, all results worse than 0.5 are binned at 0.5). Each histogram represents the mean difference measure of one model across all distributions, with the number of distributions receiving each score measured on the y-axis. The model name and the average mean difference score for that model is shown at the top of each histogram. The models are sorted from best to worst performance overall. (c) An illustration of how mean difference is calculated for one distribution. For each item in the distribution, the distance between people's judgments (blue) and the model's predictions (red) is calculated (black lines) and the mean difference is the average of all these distances.

previous empirical evidence that people are sensitive to context in making judgments of tallness. The failure of the Maximum Margin model indicates that people are not performing this simple clustering heuristic in judging the tall items.

Also of interest is the one distribution where the RH-SD model did significantly more poorly than its general performance (see Figure 3(e)). Here, there are many items that are taller than 0.4 standard deviations above the mean height (the optimal RH-SD threshold overall), and the RH-SD model labels them all tall. Adult judgments fall off far more quickly, and there are many items that very few people think are tall and the RH-SD model predicts are tall with 100% probability. Despite this large divergence between model and human performance, however, this model does not show a similar but opposite pattern of picking far too few items on any of the distributions in Experiment 1. Experiment 2 further investigates this model's performance given similar distributions in Experiment 2.

Overall, in Experiment 1, many of the models failed to match human judgments of tallness. However, all of the models that depend on the height of objects in a given distribution (RH-R, RH-SD, and CLUS) performed well overall. This suggests that, in assessing which items in a given context are tall, people are juding tallness in a way that can be modeled by a statistical function based on the heights of the objects in the category. In Experiment 2, we further explore which of these three models best describes human judgments.

## Experiment 2

We attempted to find distributions where the three best models from Experiment 1 made significantly different predictions.

We designed a set of distributions with a large number of items taller than the best previous RH-SD threshold (0.4 standard deviations above the mean), as in the distribution in Figure 3(e), such that the best previous RH-SD model would make different predictions from the best RH-R and CLUS models. Additionally, this set has several distributions with very distinct clusters of objects, causing the CLUS model to make somewhat different predictions from the more gradual sigmoidal fall-off of the threshold models.

We created six distributions, three of which were designed with a mid-height cluster of objects that people found to be ambiguous in terms of tallness, and three which were not. Two of the distributions used are pictured in Figure 4. The distributions were each divided two "clumps" of items – the clearly short items, divided into 1–3 clusters, and the taller items, divided into either one broad cluster or two very distinct clusters (the shorter of which was the ambiguously tall cluster). All the clusters were sampled from a Gaussian distribution. Across all six distributions, the mean and variance of the short clump remained the same, and likewise for the tall clump. All the items in the tall clump were greater than 0.4 standard deviations above the mean of the overall distribution.

A total of 107 adults judged the tallness of the items in these distributions, each adult viewing only 1 distribution. They viewed the distributions on a computer screen and were asked to identify each of the tall pimwits.

**Results** Figure 4(a) compares the performance of the three models to human judgments for two of the distributions. The performance of both the best parameter fit for just the six Experiment 2 disttributions is shown, as well as the performance of the best models overall (for the combination of all Experiment 1 and Experiment 2 distributions).

The best parameters for the RH-R and CLUS models are similar for the Experiment 2 distributions and for all the distributions together, and the performances of these mod-

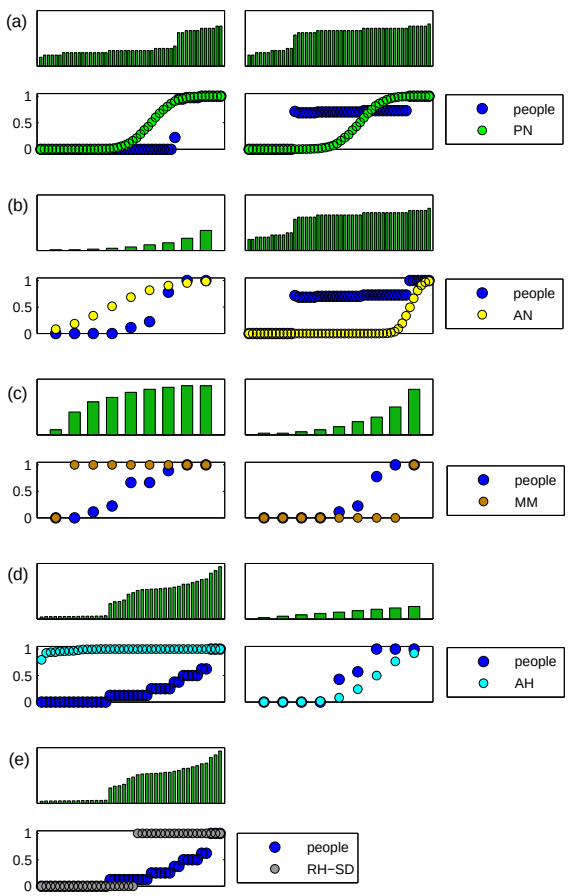


Figure 3: Several models failed to predict the tallness judgments of people in Experiment 1. (a)-(d) show four models exhibiting a *double dissociation* – they predict too many items being tall given the distribution on the left, and too few items being tall given the distribution on the right. In (e) we see that RH-SD also performed poorly in one of these directions for one distribution.

els are very similar in both cases; the best RH-R threshold is $k = 27\%$ for Experiment 2 (with a noise parameter of $\varepsilon = 0.01$) and $k = 29\%$ overall (with a noise parameter of $\varepsilon = 0.05$). The CLUS hyperparameters were also similar for the best Experiment 2 and best overall performance.

In strong contrast, the parameter values and the performance for the RH-SD model are far different for the distributions in Experiment 2 than the combined distribution set. The best overall model ($k = 0.4, \varepsilon = 2.5$) predicts with certainty that all the items in the tall clump are tall, whereas people are uncertain about the middle cluster of items. The RH-SD parameters can be adjusted to better match the human judgments for Experiment 2 ($k = 1.0, \varepsilon = 75$), but these parameter results do poorly on the overall set of distributions, as shown in the histograms in Figure 4(b). These histograms also show the good performance overall of the best RH-R and CLUS models from Experiment 2.

Both the RH-R and the CLUS models fit the human judgments reasonably closely, and it is not clear which model provides the best fit. There are, however, differences in their predictions. The CLUS model does show an almost equal set of probabilities for the mid-height cluster items being tall; the RH-R model shows a more sigmoidal fall-off for these items. Additionally, the RH-R model predicts a sharper fall-off in tallness judgments for the tallest cluster of items than the CLUS model does. And, with the best overall parameters, the RH-R model predicts a greater-than-zero probability of some of the short clump of items being labeled *tall*, though no humans rated any of these items tall. If the RH-R noise parameter is adjusted such that the predictions for the mid-cluster items are even flatter, then the sigmoidal curve becomes more apparent in both the shorter and taller items. The CLUS model, by contrast, does not predict any of the short clump items as tall without drastic parameter variation. Though these results are suggestive about differences in the models, we do not have a definitive answer as to which of the two models best fit human judgments, which, given the current amount of data, are well approximated by both.

Overall, the results of Experiment 2 show clearly that the RH-SD model, an intuitive parametric model based on the mean and standard deviation of a distribution, cannot account for how people make judgments about gradable adjectives. The qualitatively different predictions of the two non-parametric models, RH-R and CLUS, suggest that future work will help us understand which of these two models is the strategy used by people, though on average both models predict the Experiment 2 data well.

## Conclusion

The success of the RH-R model and the CLUS model in Experiments 1 and 2 suggests that people do perform a statistical computation upon the comparison set when using the word tall. However, this computation is not based on a simple parametric statistical criterion as in the RH-SD model, but some more nonparametric function that is meaningful for a wider range of distributions.

Though a parametric model such as RH-SD may seem intuitively appealing, people regularly encounter sets of items that do not follow a parametric distribution. Such contexts include cases where the objects under consideration do not belong to one natural type, e.g., “the things on the table”. In these situations, the distribution of object heights may be highly nonuniform or multimodal, and non-parametric models can accomodate this in ways that parametric models cannot.

The RH-R model is simpler than the CLUS model, in the sense that it has two free parameters where the CLUS model has four. However, there are two reasons why the CLUS model is a compelling model from a cognitive perspective. First, it seems to capture a crucial idea of linguistic compositionality: we use noun phrases to pick out complex subcategories that are not lexicalized as single nouns but that nevertheless correspond to interesting chunks of the world. Second, this model relies on well-established, domain-general

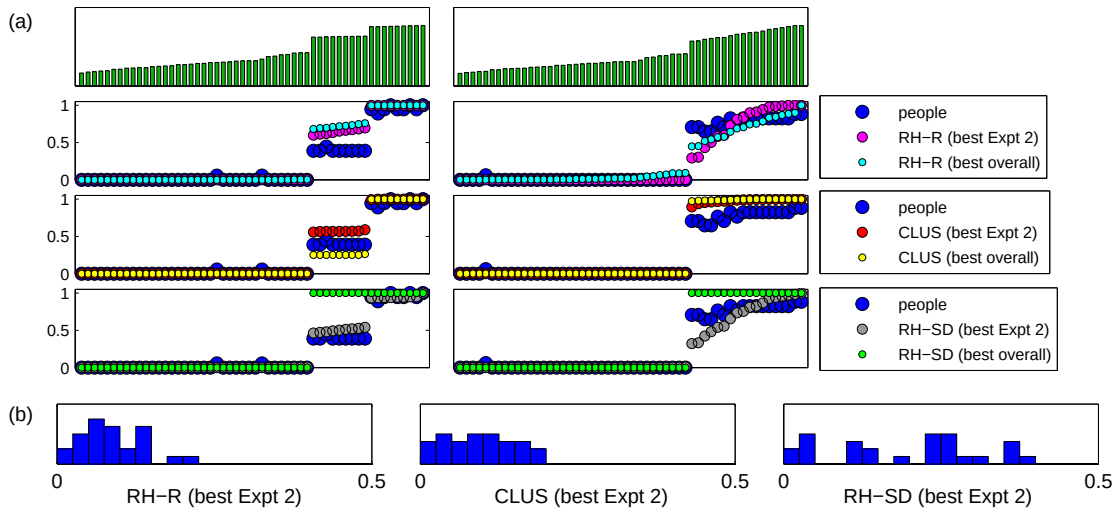


Figure 4: (a) Two sample distributions from Experiment 2 are shown at the top, with the performance of the three models shown below. For each model, the best parameter settings for just Experiment 2 and the best parameter settings for all the distributions of both experiments are compared. The best RH-R and CLUS models are very similar for overall results vs. for just the Experiment 2 distributions. The best RH-SD model overall, however, is very different from the best RH-SD model for just the Expt 2 distributions – so much so that the best overall model does not capture adult judgments for the Experiment 2 distributions well at all. (b) The best Experiment 2 RH-R and CLUS models also perform very well overall, but that is not the case for the best Experiment 2 RH-SD model. The histograms show the mean differences for the best Experiment 2 models across all distributions from Experiments 1 and 2.

mechanisms for categorization. If these strategies also apply to understanding gradable adjectives, this would not entail positing any truly new cognitive complexity.

Future work will further explore the different predictions of the RH-R model and the CLUS model to answer the question of which non-parametric method people are employing when they use gradable adjectives. Once the best model has been found for *tall*, this work can be extended to related questions of language compositionality — does the best model also apply to how people use other gradable adjectives, such as *big, small, short*, and *loud*, and are the best parameters similar? What about intensifiers like *very* — can the *very tall boys* be identified by applying either the Cluster model or range-based thresholding to the class of *tall boys*? Additionally, there are quantifiers such as *most* or *several* that apply to an ordered scale of set-sizes, and like gradable adjectives, can have flexible, context-sensitive meanings that seem to draw on both compositionality and statistics (Halberda, Taing, & Lidz, 2008). Can a model of gradable adjectives give insight into how those words are used? Further work remains in order to determine the extent to which the model applies to other words with compositional meaning besides *tall*.

Also of interest is the developmental progression of gradable adjective understanding. While work by Barner and Snedeker (2008) suggests that children use *tall* as a statistical categorization function from an early age, it remains to be seen whether they start out with the same model of gradable adjectives that they eventually use as adults, or whether they learn the structure of the meaning of *tall* based on experience.

While much work remains to be done, the knowledge that a statistical, non-parametric model predicts human usage of *tall* marks a step forward in our understanding of how people use gradable adjectives and, more broadly, how word meanings can compose to form phrases.

## References


Aldous, D. (1985). Exchangeability and related topics. In *Ecole d'Ete de Probabilites de Saint-Flour, XIII-1983*. Berlin: Springer.

Anderson, J.R. (1991). The adaptive nature of human cat- egorization. *Psychological Review, 98(3)*, 409–429.

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child Development, 79*, 594–608.

Cresswell, M.J. (1976). The Semantics of Degree. In: B.H. Partee (Ed.), *Montague Grammar*. New York: Academic Press.

Fodor, J.A., & Lepore, E. (2002). *The Compositionality Papers.* Oxford: Clarendon Press, 2002.

Frege, G. (1892). On concept and object. Reprinted in B. McGuiness (Ed.), *Collected Papers on Mathematics, Logic, and Philosophy*. Oxford: Blackwell, 1984.

Halberda, J., Taing, L., & Lidz, J. (2008). The development of "most" comprehension and its potential dependence on counting-ability in preschoolers. *Language Learning and Development, 4(2)*, 99-121.

Kennedy, C. (1999). *Projecting the adjective: The syntax and semantics of gradability and comparison*. New York: Garland.

Klein, E. (1991). Comparatives. In A. von Stechow & D. Wunderlich (Eds.), *Semantics: An International Handbook of Contemporary Research*. Berlin: Walter de Gruyter.

Rasmussen, C.E. (2000). The Infinite Gaussian Mixture Model. In S.A. Solla, T.K. Leen & K.R. Müller (Eds.), *Advances in Neural Information Processing Systems, 12* (pp. 554-560). Cambridge, MA: MIT Press.

Smith, L.B., Cooney, N. & McCord, C. (1986). What is "High"? The development of reference points for "High" and "Low." *Child Development, 57*, 583-602.