Comparing Pluralities

Gregory Scontras (scontras@fas.harvard.edu)

Department of Linguistics, Harvard University

Peter Graff (graff@mit.edu)

Department of Linguistics and Philosophy, Massachusetts Institute of Technology

Noah D. Goodman (ngoodman@stanford.edu)

Department of Psychology, Stanford University

Abstract

What does it mean to compare sets of objects along a scale, for example by saying "the men are taller than the women"? We explore comparison of pluralities in two experiments, eliciting comparison judgments while varying the properties of the members of each set. We find that a plurality is judged as "bigger" when the mean size of its members is larger than the mean size of the competing plurality. These results are incompatible with previous accounts, in which plural comparison is inferred from many instances of singular comparison between the members of the sets (Matushansky and Ruys, 2006). Our results suggest the need for a type of predication that ascribes properties to plural entities, not just individuals, based on aggregate statistics of their members. More generally, these results support the idea that sets and their properties are actively represented as single units.

Keywords: Comparatives; plurality; set-based properties; natural language semantics; mental representations

Word count: 3801

1. Introduction

When we think and talk about groups of individuals—pluralities—do we represent the collection as a single entity with its own properties? For example, when we say "the red dots are big" is there an aggregate size for the group of red dots to which we refer? In this paper we investigate this question by studying plural comparison—e.g. "the red dots are bigger than the blue dots". Given two individuals, there is little question of how to proceed with comparison of, e.g., size: find the degree of size for the first and compare it to the degree of size for the second. In situations where pluralities of entities are compared, this is less straightforward: each member of the plurality has a degree associated with it and it is not clear whether these degrees should be used together for comparison.

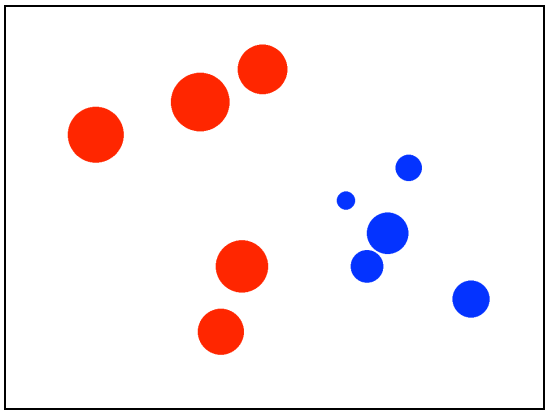


Figure 1: Every red dot is bigger than every blue dot.

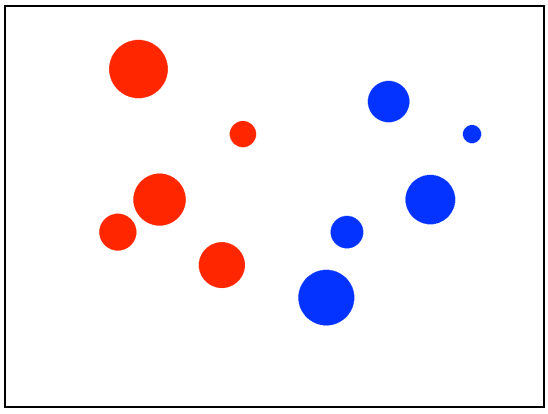


Figure 2: Every red dot is bigger than some blue dot and every blue dot is smaller than some

red dot.

Some instances of plural comparison can be relatively unambiguously translated into single-entity comparisons. In Figure 1, every member of the red dots is bigger than every member of the blue dots, and one concludes that the red dots are bigger. However, Matushansky and Ruys (2006) point to cases like Figure 2, where they claim that "the red dots are bigger than the blue dots", yet it is not the case that every single red dot is bigger than every single blue dot. To account for this intuition, they propose a semantics for plural comparison that imposes a categorical condition stating that for the sentence "the red dots are bigger than the blue dots" to be true, the biggest red dot must be larger than the biggest blue dot, the second-biggest red dot larger than the second-biggest blue dot, and so on. This reduces plural comparison to a condition on singular comparisons between the two pluralities.

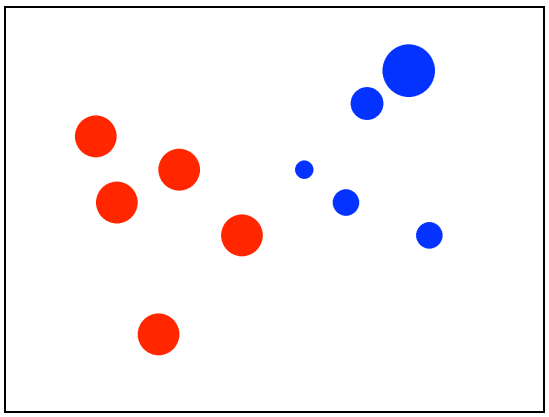


Figure 3: one blue dot is bigger than every red dot.

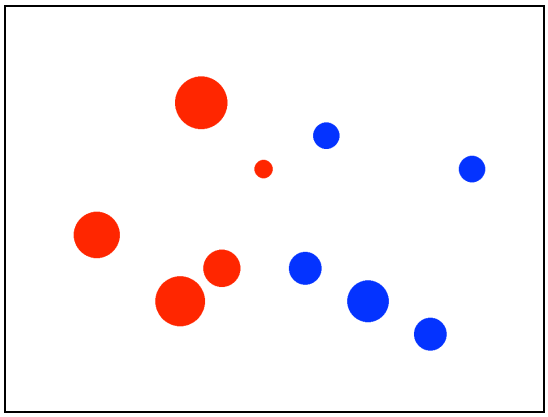


Figure 4: one red dot is smaller than every blue dot.

Scontras (2008) claims that there are cases where we may judge the red dots as bigger, even though single-entity comparison fails. For example, there could be one blue

dot that is larger than all red dots, or one red dot that is smaller than all blue dots (Figures 3 and 4), and still intuitions, which we explore experimentally below, suggest that the red dots are bigger. Based on this observation, Scontras (2008) claims that the semantics of plural comparison should not be viewed as many instances of singular comparison, but rather as a single comparison of collective properties of each plurality.

These conflicting theories for plural comparison expose a deeper theoretical issue. Under a point-wise approach (Matushansky and Ruys, 2006), only the properties of the members of the competing pluralities are relevant to the outcome of plural comparison—the plurality need not be represented as an entity with its own properties. Collective approaches (Scontras, 2008) treat the pluralities as complex individuals with properties derived from the members of their corresponding pluralities, and crucially these derived properties need not be true of any individual member (for instance, no single red dot in Figure 4 has the mean size). Linguistic theories of plurality differ on whether a plural definite description such as *the red dots* refers to a single, complex individual, which may possess properties independently of its members, or whether it must refer to the many individuals that make up the plurality (see, e.g., Landman, 1996; Schwarzschild, 1996 for discussion). The debate centers on the ontological commitments of the two approaches, with the complex individual approach requiring an augmented ontology containing complex individuals corresponding to every plurality (Landman, 1989). Critics of a complex individual approach argue that such an expansion of the ontology is not empirically justified, and that we can make do simply with pluralities in our logical representations of language (Schwarzschild, 1996).

In the experiments presented below, we use plural comparison as a window into this ontological question, and thus inform the linguistic debate on complex individuals while initiating a psychological study of the representation of pluralities. If we find that plural comparison proceeds in terms of collective properties associated with the pluralities being compared, we may conclude that plural entities are represented as individuals with their own properties (derived from their members). Such a finding would suggest that our ontology of the plural domain should include structure rich enough to derive these complex individuals from their corresponding plural sets.

A further question surrounds ways that aggregate degrees can be constructed: Scontras's (2008) account necessitates representation of an *average* degree of a plurality. This claim is unexpected given traditional views of collective predication (e.g., Link, 1983; Scha, 1984), which have been proposed to account for sentences such as "the boxes are heavy" describing situations in which the weight of many light boxes sums to a large total weight. The traditional view of collective predication thus predicts that the aggregate *sum* of individual sizes is the collective property predicated.

In this paper, we experimentally investigate plural comparison among concrete sets of objects for the first time. Experiment 1 shows that people compare inferred collective properties of pluralities. Experiment 2 shows that the best predictor of the outcome of plural comparison in our setting is the difference in mean degree (and not the sum). These results provide evidence both for representations of definite descriptions in terms of complex individuals and for collective predication in terms of averages.

2. Models of Plural Comparison

In this section we elaborate testable mathematical models of plural comparison based on the linguistic theories presented above. The models come in two families: (a) point-wise models, which perform singular comparisons between the individual members within the pluralities; and (b) collective models, which infer collective properties from the relevant pluralities and compare those inferred properties. Since both linguistic theories compared are formulated to make categorical predictions about judgments, we include a probabilistic version of each model to allow for better approximation of judgments averaged across subjects and items.

2.1. Point-wise Models

The first model we consider is the categorical version of the point-wise model. This model is based on the semantics given in Matushansky and Ruys (2006), which states that a plurality X is bigger than a plurality Y just in case there exists a bijective function *f* from the members of X to the members of Y such that each member x of X wins its comparison with *f*(x).

The probabilistic version of the point-wise model computes the expectation of a given point-wise comparison coming out in favor of plurality X: that is the probability that a random element of X is bigger than a random element of Y. The model is formulated as follows:

$$P(X > Y) = \sum_{i,j} \frac{x_i > y_j}{|X| * |Y|}$$

One should note that the probabilistic point-wise model is not a direct translation from Matushansky and Ruys's (2006) semantics. Given the overall poor performance of the categorical point-wise model (see below), we relaxed its conditions on comparison to get a more naturalistic probabilistic model, while remaining consistent with the theoretical motivation for point-wise comparisons.

2.2. Collective Models

The categorical collective mean model is based on Scontras's (2008) semantics: a plurality X is bigger than a plurality Y if the average size of X is larger than the average size of Y.

In the probabilistic version, we account for potential response variation by assuming normally distributed noise around the estimate for the mean size of each plurality. Thus, X is bigger than Y if the estimated mean of X is bigger than that of Y. The probability of this event is described by the cumulative distribution function (CDF)[1] of a normal distribution, or error function `erf` (Temme, 2010) on the difference in means. That is:

$$P(X > Y) = \frac{1}{2}\left[1 + erf\left(\mu(X) - \mu(Y)\right)\right]$$

In addition to collective mean models, in Experiment 2 we also consider a collective model based on a comparison of total size. Like the mean-based models, this

[1] For a value x of a random variable X distributed according to some probability distribution, the CDF returns the probability of X assuming a value smaller than or equal to x.

sum-based model first infers an aggregate statistic (here sum of sizes), and then compares this statistic assuming noisy measurements.

3. Experiment 1

In order to test the predictions of point-wise vs. collective models, we asked subjects to provide judgments on comparison for 32 images with 5 red and 5 blue dots each. Because difference-in-mean and difference-in-sum strategies give the same predictions when the cardinalities of the two pluralities are equal, we only consider the mean models for these data. That is, we consider the predictions of the categorical and probabilistic mean models, as well as the categorical and probabilistic point-wise models.

3.1. Participants

We recruited 50 subjects through Amazon.com's Mechanical Turk Crowdsourcing Service, a marketplace tool for gathering behavioral data. Subjects were compensated for their participation.

3.2. Stimuli

We selected 32 instances of plural comparison with five red dots and five blue dots for which the predictions of the models differed maximally. Dot size varied between 1 and

7.[2] Instances of comparison varied continuously for the predicted outcome of comparison according to the four models and were composed in such a way as to minimize the correlation between the models' predictions. The average absolute difference in mean sizes was 2.5. Placement of the dots on the screen was randomly chosen with the constraint that no red dot ever appeared to the right of any blue dot. This constraint was used to ensure that the compared pluralities were visually separable to further facilitate the identification of the groupings according to color. An example stimulus is given in Figure 5.

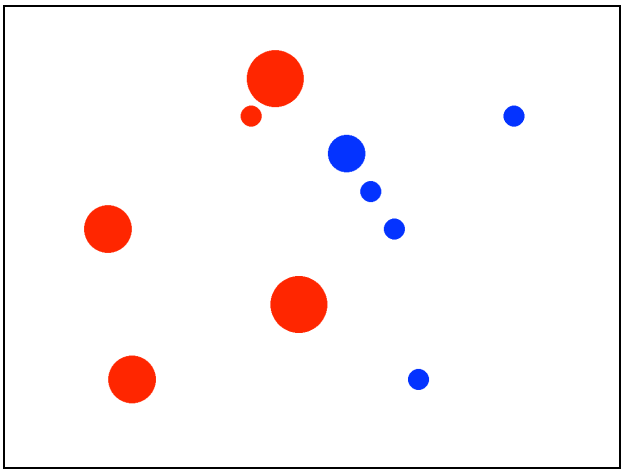


Figure 5: Sample stimulus; 100% of subjects consider the red dots bigger.

3.3. Design

We generated 4 different random orders for the 32 instances of comparison. Each list was posted as a single Human Intelligence Task to Amazon's Mechanical Turk

[2] Actual area on the screen as measured in pixels is computed as follows: 2/3•((400(dot-size))^2).

crowdsourcing platform. Ten subjects were assigned to each list, except for the fourth list, which was run twice due to technical error. Whether or not we include the extra subjects run for the fourth list does not affect the results we report.

For each image, subjects responded to the question, "Are the red dots bigger than the blue dots?" They chose between the answers YES and NO. Before seeing the images, subjects were asked to indicate their native language. Seven subjects were excluded because they indicated a native language other than English and/or failed to answer all questions. Our results are not affected by their exclusion. Data from 43 subjects was analyzed.

3.4. Results

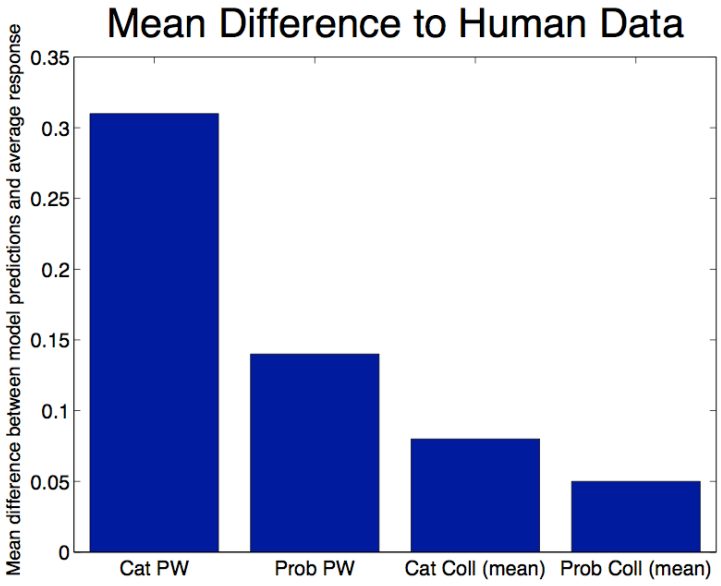


Figure 6: Average deviations from human data, calculated by comparing the models' predictions for each stimulus to the average judgment for that stimulus. Models

considered: categorical point-wise (Cat PW), probabilistic point-wise (Prob PW), categorical collective mean (Cat Coll), and probabilistic collective mean (Prob Coll).

The results are summarized in Figure 6, where we report the mean difference of each model from the human data. We find that the collective models outperform the point-wise models, and that the probabilistic models outperform the categorical models; the model with the best fit is the probabilistic collective mean model.

To understand the success of the collective models, Figure 7 plots the difference in mean size between red and blue dots against the percent of YES answers to the question "Are the red dots bigger than the blue dots?". Note the clear trend.

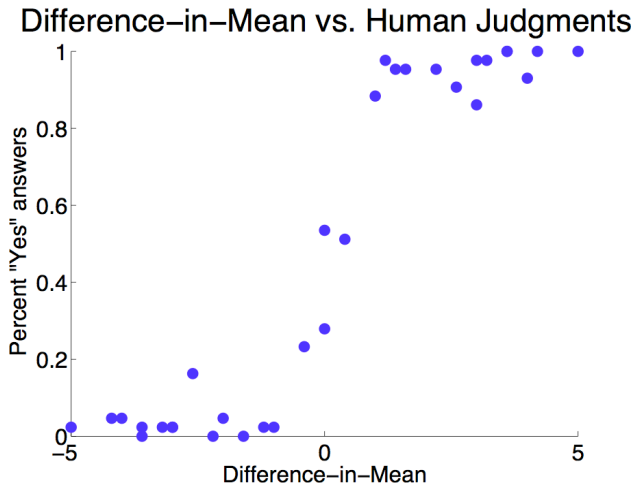
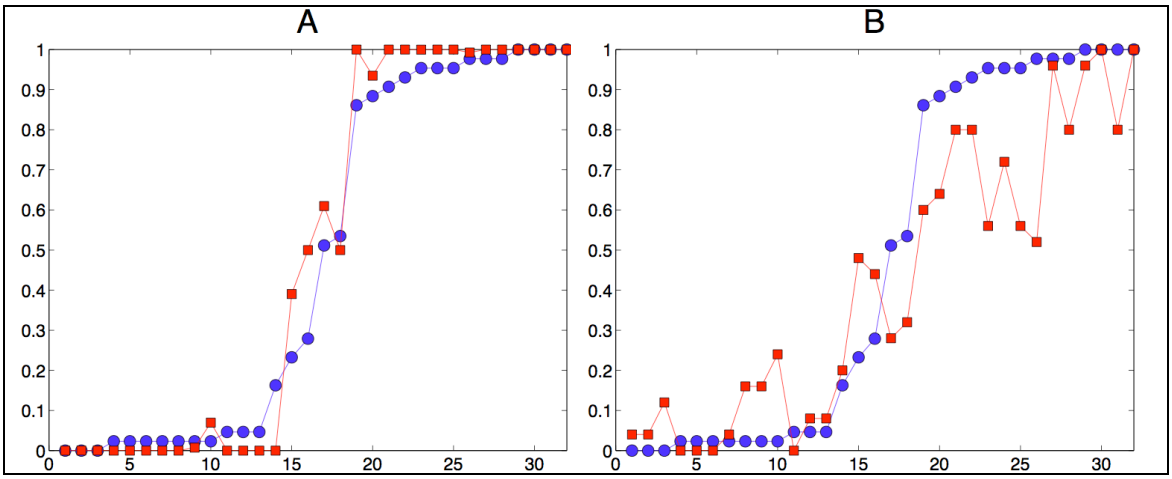


Figure 7: Difference in mean size plotted against average responses.

The patterns in Figure 6 are substantiated by mixed logit model analysis, predicting human responses, including random intercepts for subject and item and random slopes for the predictions of the probabilistic and categorical collective models grouped by subject (i.e., the maximal random effects structure justified by the data). Neither categorical

model significantly improves data-likelihood ($\chi^2(1)=.08$, p=.78, collective; $\chi^2(1)=.22$, p=.64, point-wise). The probabilistic collective model explains significant variance ($\chi^2(1)=7$, p<.01).[3] Residual variance is accounted for by the probabilistic point-wise model ($\chi^2(1)=5.27$, p<.05). A look at the predictions of the four models compared with the judgment data (Figure 8) confirms that the probabilistic collective model outperforms the other models. As expected, the predictions of the probabilistic mean model significantly improve data-likelihood compared to the null-model ($\chi^2(1)= 87.57$,p<.001), which guesses the majority outcome for all comparisons.



[3] Given that the erf function approximates a normal CDF, which has a free parameter for standard deviation, it is possible that our probabilistic collective model could achieve a better fit once this parameter is optimized. Simulated annealing reveals that the best value of this parameter for our data is 0.47, as opposed to 0.50 (the default value used by erf). However, the optimized value does not affect the results we report, and so we continue to use 0.50 as the value for standard deviation in our probabilistic collective models.

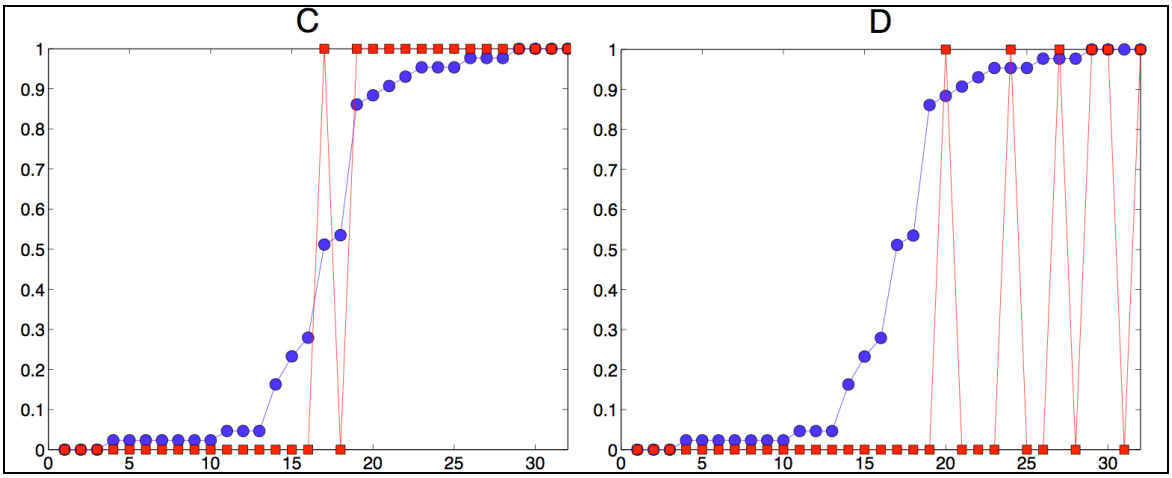


Figure 8: Average judgments (in blue) plotted against models' predictions (in red); X axis: instances of plural comparison in order of percent YES answers; Y axis: percent YES answers to the question "Are the red dots bigger than the blue dots?" (A) probabilistic collective (mean) model; (B) probabilistic point-wise model; (C) categorical collective (mean) model; (D) categorical point-wise model.

3.6. Discussion

We have shown evidence that collective models outperform point-wise models, and that probabilistic models are better than categorical ones. The probabilistic collective model performs the best; the probabilistic point-wise model does less well, but still accounts for residual variance. When we compare the model predictions to the data (Figure 8), we see that the probabilistic collective model makes most of its errors when the means of the two pluralities are very close. This raises the possibility that a point-wise strategy arises in cases where collective comparison is too difficult. The next experiment follows up on this result by testing scenarios where we expect more graded responses—ones in which the means of the compared pluralities differ by a smaller margin; if the probabilistic point-wise model outperforms the collective model in these cases of graded judgments, we

expect it to account for more variance in Experiment 2 where the comparisons are more difficult in terms of a mean strategy.

Another issue raised by our results is whether the collective interpretation people demonstrate is driven by the fact that the dots are filled in with color, potentially allowing people to compute the amount of blue versus red on the screen, and thus facilitating a collective comparison strategy. Our intuitions suggest that such a strategy is less available when the dots appear as unfilled circles, and so in the next experiment we vary between subjects whether dots are filled with color or unfilled. Assuming filled dots privilege a sum-based collective strategy (cf. Scha, 1984), we would expect to find an interaction such that difference-in-sum is a better predictor of judgments when the dots are filled.

Finally, having found that collective strategies are at play in plural comparison, we must next determine which collective strategy is relevant: difference-in-mean or difference-in-sum.

4. Experiment 2

In this experiment, we compare the predictions made by the probabilistic mean model and the probabilistic sum model. Additionally, we follow up on the results of the previous experiment by testing the predictions of the probabilistic point-wise model with novel stimuli. We used similar methods from Experiment 1 to test situations of plural comparison in which the two pluralities being compared differ in cardinality. We also decrease the difference in means between the competing pluralities. Lastly, we vary

whether dots are filled with color or unfilled between subjects to assess the likelihood of the sum interpretation.

4.1. Participants

We recruited 35 subjects through Amazon.com's Mechanical Turk. Subjects were compensated for their participation.

4.2. Stimuli

Stimuli consisted of ten instances of plural comparison, seven with differing cardinalities between the compared pluralities (between 3 and 5 dots) and three filler scenarios with equal cardinality (5 dots). The seven differing cardinality instances were chosen so that the values of difference-in-mean and difference-in-sum were always of different signs. The average value of difference-in-mean was smaller than in Experiment 1; before, stimuli had a mean value of 2.5 for absolute difference-in-mean; here 0.8. As before, the predictions of our models varied continuously across target items.

All ten scenarios were presented in two configurations, reversing the sizes of the red and the blue dots. Both comparison configurations were displayed in two random spatial organizations. Stimulus composition was identical for both filled and unfilled dots between subjects.

4.3. Design

Images were presented in a different random order for each subject. As before, subjects responded to the question, "Are the red dots bigger than the blue dots?" and chose between the answers YES and NO. The filled and unfilled conditions were posted separately to Mechanical Turk, with 20 subjects recruited for each condition. Only native English speakers who responded to all questions and who did not take part in Experiment 1 were included in the analysis. Six subjects were excluded because they completed both the filled and the unfilled version of the experiment; no other subjects were excluded. Data from 34 subjects (17 from each of the fill conditions) was analyzed.

4.4. Results

Because the predictions of the difference-in-sum and the difference-in-mean strategies differ only when the cardinalities of the pluralities differ, we analyze the responses to these scenarios only.

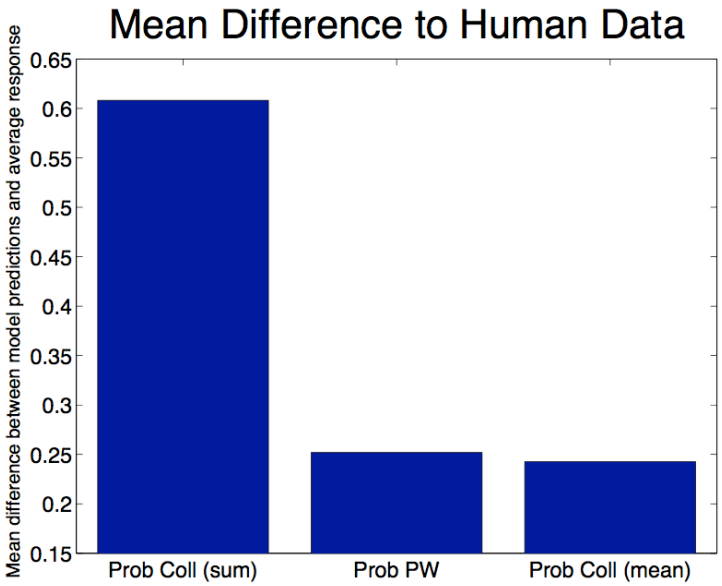


Figure 10: Average deviations from human data, calculated by comparing the models' predictions for each stimulus to the average judgment for that stimulus. Models considered: probabilistic collective sum (Prob Coll (sum)), probabilistic point-wise (Prob PW), and probabilistic collective mean (Prob Coll (mean)).

Figure 10 reports the mean difference of each model from the human data. Based on this measure, we find that the collective mean model performs the best, and the collective sum model performs the worst. This measure also shows that the probabilistic point-wise model performs slightly worse than the probabilistic mean model. Figures 11 and 12 plot the difference in mean size and difference in sum size against the human responses; we see a clear trend in the predicted direction between difference-in-mean and average response, and no such trend with difference-in-sum.

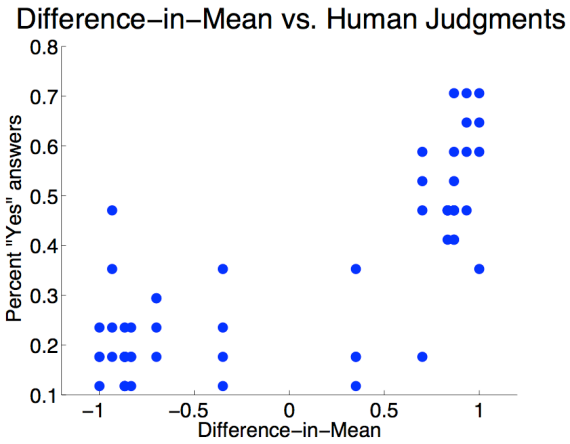


Figure 11: Difference in mean size plotted against average response.

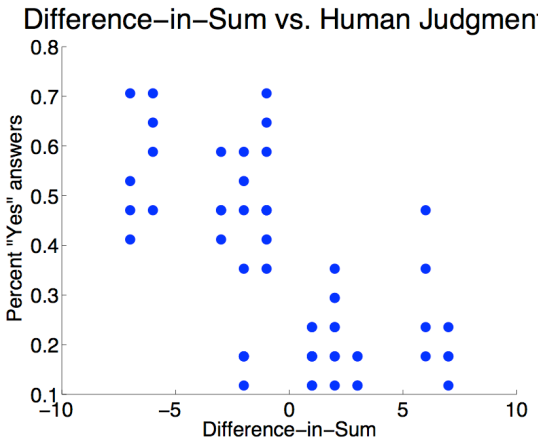


Figure 12: Difference in sum size plotted against average response.

We fit a mixed logit model including the predictions of the probabilistic point-wise model, the difference in the mean sizes, the difference in the sum of sizes, the

difference in cardinality, as well as dot fill and interactions with dot fill as fixed effects. The model also included random intercepts for subjects and items, as well as a random slope for difference in mean grouped by subject and a random for the predictions of the probabilistic point-wise model grouped by item (i.e., the maximum random effects structure justified by the data). The results of the respective χ^2 likelihood tests are given Table 1.

Predictor	d.f.	χ^2	*p*
Mean + interaction with Fill	**2**	**6.09**	**0.048 ***
Point-wise + interaction with Fill	2	2.55	0.279
Sum + interaction with Fill	2	0.83	0.660
Cardinality + interaction with Fill	2	0.87	0.649
Fill + all interactions	5	2.54	0.770

Table 1: Predictors and their contribution to data likelihood.

The difference in mean size is the only significant predictor in terms of improvement of data-likelihood ($\chi^2(2)=6.09, p<.05$) in addition to the other models. As expected, the difference in mean size also significantly improves data-likelihood compared to the null-model ($\chi^2(1)= 6.45, p<.05$), which guesses the majority outcome for all comparisons.

4.5. Discussion

As subjects perform plural comparison, the difference in sums is not relevant; the deciding factor is the difference in means. Of particular interest is the lack of interaction of dot fill with difference-in-sum. Filling in the dots makes the total amount of red vs. blue on the screen more salient, and so we might expect a bias in favor of a difference-in-sum strategy. The lack of interaction provides evidence against the possibility of difference-in-sum being a viable strategy for plural comparison in our setting. Lastly, the probabilistic point-wise model does not account for any residual variance; here we perform a more detailed test of the collective (mean) vs. probabilistic point-wise strategies by decreasing the difference in mean size between the compared pluralities. Thus, we provide further evidence against the role of point-wise strategies in plural comparison.

5. General Discussion

We have presented evidence that the comparison of pluralities along the size dimension proceeds in terms of a comparison of the means of the relevant sizes inferred from each plurality. We have shown that models of plural comparison that operate on individual comparisons perform less well. Even in Experiment 1, where the probabilistic point-wise model captures some variance, the collective model based on difference-in-mean performs better in terms of data likelihood. These results suggest the need for a richer kind of plural predication that treats pluralities as complex individuals with properties inferred from the members of the plurality; here, the relevant property is a degree inferred by averaging over individual degrees.

While these results suggest the need for a plural ontology that includes complex individuals, additional work will be needed to determine when such complex individual interpretations are available. For example, if we assume that complex individual interpretations are always available for definite descriptions, then a term like *the red dots and the blue dots* should be multiply ambiguous such that it may behave as a single individual, as two, or as many. The potential availability of these readings requires empirical study.

We have shown that average is one function used to aggregate individual properties into a collective property, but there remains the possibility that the relevant aggregation function may itself be contextually determined. For instance, the assertion that plurality X beats plurality Y in a given comparison could be true when a contextually specified aggregation function *f* (determined by property, discourse goals, etc) is such that *f*(X) is greater than *f*(Y). We hope future research will shed light on this possibility.

Finally, the results of the current experiments invite follow-ups to tease apart the nature of the psychological processes and representations involved in this judgment task. In the tasks discussed here, subjects were given concrete, perceptually available sets (grouped by color), and asked about perceptually salient properties of these sets. Previous work on the extraction of set properties (e.g., Halberda, 2007) has shown that different grouping characteristics affect the ease with which subjects infer properties of collections of objects. Future work will explore whether people can infer appropriate aggregate values without the availability of salient sets by manipulating the grouping characteristics of the sets, and similarly, whether the concreteness of the property being considered affects the way in which plural comparison can proceed.

Acknowledgements

The authors would like to thank Irene Heim, Josh Tennenbaum, and audiences at the 46th Annual Meeting of the Chicago Linguistic Society, the 85th Annual Meeting of the Linguistic Society of America, and the MIT Workshop on Comparatives.

Appendix A: Experiment 1 comparison scenarios

Red dot sizes	Blue dot sizes
1,1,1,1,3	1,1,2,2,7
1,1,1,1,3	1,1,2,7,7
1,1,1,1,3	1,5,5,5,7
1,1,1,1,3	1,5,5,7,7
1,1,1,1,3	3,6,6,6,7
1,1,1,1,3	4,4,4,6,7
1,1,1,1,3	5,6,7,7,7
1,1,2,2,4	3,3,4,4,6
1,1,2,2,7	1,1,1,1,3
1,1,2,2,7	1,2,2,3,3
1,1,2,2,7	5,5,5,5,6
1,1,2,2,7	6,6,7,7,7
1,1,2,7,7	1,1,1,1,3
1,1,2,7,7	6,6,7,7,7
1,2,2,3,3	1,1,2,2,7
1,2,2,5,6	2,3,3,6,7
1,2,3,3,4	4,5,6,6,7
1,2,3,4,6	2,2,2,3,7
1,5,5,5,7	1,1,1,1,3
1,5,5,7,7	1,1,1,1,3
2,3,4,7,7	1,3,4,4,4
2,4,5,7,7	1,3,4,6,6
3,5,5,5,6	5,6,7,7,7
3,6,6,6,7	1,1,1,1,3
4,4,4,6,7	1,1,1,1,3
4,5,5,7,7	1,3,4,6,6
4,5,5,7,7	4,5,5,7,7
4,5,6,6,7	1,2,3,3,4
5,5,5,5,6	1,1,2,2,7
5,6,7,7,7	1,1,1,1,3
6,6,7,7,7	1,1,2,2,7
6,6,7,7,7	1,1,2,7,7

Appendix B: Experiment 2 comparison scenarios

Red dot sizes	Blue dot sizes
1,2,2,3,6	2,3,4,4,4
1,2,3,4,5	2,3,4,5,6
1,3,4,5,6	2,3,3,4,4
2,2,3,3,4	3,4,4
2,3,3,4,4	1,3,4,5,6
2,3,4,4,4	1,2,2,3,6
2,3,4,5,6	1,2,3,4,5
3,3,3,3,5	3,3,4,5
3,3,3,5	4,4,5
3,3,4,5	3,3,3,3,5
3,4,4	2,2,3,3,4
3,4,4,5	5,5,5
3,4,5,5,7	4,5,6,7
4,4,4,4,6	5,5,6
4,4,5	3,3,3,5
4,4,5,5,6	5,6,6
4,5,6,7	3,4,5,5,7
5,5,5	3,4,4,5
5,5,6	4,4,4,4,6
5,6,6	4,4,5,5,6

References

Halberda, J. (2007). Subitizing sets and set-based selection: Early visual features determine what counts as an individual for visual processing. Poster presented at VSS, the Vision Sciences Society, Sarasota, FL.

Landman, F. (1989). Groups I. *Linguistics and Philosophy* 12: 559–605.

Landman, F. (1996). Plurality. In *The Handbook of Contemporary Semantic Theory* (pp. 425–457). Oxford: Blackwell.

Link, G. (1983). The logical analysis of plurals and mass terms. In *Meaning, Use, and Interpretation of Language* (pp. 302–323). Berlin: Walter de Gruyter.

Matushansky, O., and Ruys, E. G. (2006). Meilleurs voeux: quelques notes sur la comparaison plurielle. *Empirical Issues in Formal Syntax and Semantics* 6: 309–330.

Scha, R. (1984). Distributive, collective and cumulative quantification. In *Truth, Interpretation, and Information* (pp. 131–158). Dordrecht: Foris.

Schwarzschild, R. (1996). *Pluralities*. Dordrectht: Kluwer.

Scontras, G. (2008). The Semantics of Plural Superlatives. Bachelor's Thesis, Boston University, 2008.

Temme, N. M. (2010). Error Functions, Dawson's and Fresnel Integrals. In *NIST Handbook of Mathematical Functions* (pp. 159–171). Cambridge University Press.