Supplementary Materials: Improving measurements of similarity judgments with machine-learning algorithms

Jeffrey R. Stevens[1], Alexis Polzkill Saltzman[1], Tanner Rasmussen[1], & Leen-Kiat Soh[1]

[1] University of Nebraska-Lincoln

Table S1
Predictors

Predictor name	Value/function
Small value	S
Large value	L
Difference	$L - S$
Ratio	$\frac{S}{L}$
Mean ratio	$\frac{S}{\frac{S+L}{2}}$
Log ratio	$\log(\frac{S}{L})$
Relative difference	$\frac{L-S}{L}$
Disparity ratio	$\frac{L-S}{\frac{S+L}{2}}$
Salience	$\frac{L-S}{S+L}$
Discriminability	$\log(\frac{L}{L-S})$
Logistic	$\frac{1}{1+e^{L-S}}$

Note. Table from Stevens & Soh (2018).

Table S2
Confusion matrix and classification performance metrics

Data split	Algorithm	True positives	True negatives	False positives	False negatives	Accuracy	True positive rate	True negative rate	Positive predictive value	Negative predictive value
Training	C5.0	39.6	57.2	1.4	1.8	96.8	95.2	97.3	96.4	96.7
Training	CART	36.4	54.7	3.9	5.0	91.1	86.2	92.8	88.4	90.9
Training	kNN	33.6	53.9	4.7	7.9	87.4	75.0	90.0	87.3	87.2
Training	Naive Bayes	38.0	54.9	3.7	3.4	92.9	92.5	93.4	90.5	93.3
Training	Neural Network	39.2	56.7	1.9	2.2	95.9	94.2	96.2	95.5	95.8
Training	Random Forest	41.4	58.5	0.1	0.0	99.9	99.9	99.9	99.9	99.9
Training	Regression	40.0	57.3	1.3	1.4	97.3	96.6	97.3	97.0	97.2
Training	SVM	38.3	56.0	2.6	3.1	94.3	91.7	94.7	93.8	94.2
Testing	C5.0	34.8	53.9	5.6	5.7	88.7	85.2	89.3	84.3	89.0
Testing	CART	33.0	53.5	6.0	7.4	86.6	79.2	88.9	81.7	86.4
Testing	kNN	28.7	51.7	7.8	11.7	80.4	62.4	83.6	72.6	80.1
Testing	Naive Bayes	35.1	53.5	6.0	5.4	88.6	86.6	88.8	83.9	89.4
Testing	Neural Network	35.2	54.7	4.8	5.3	89.8	85.4	90.6	86.5	89.8
Testing	Random Forest	35.2	54.6	4.9	5.3	89.8	85.9	90.5	86.6	89.9
Testing	Regression	33.6	52.6	6.9	6.9	86.3	81.7	87.1	80.8	86.8
Testing	SVM	34.7	55.2	4.3	5.7	90.0	84.1	91.4	88.0	89.4

Note. Rates based on random ordering of a sample size of 30 instances. True positive rate = recall; Positive predictive value = precision.

Table S3
Predictor importance

Algorithm	Predictor importance calculation
C5.0	Percentage of training set samples that fall into all the terminal nodes after the split
CART	Reduction in the loss function (e.g., mean squared error) attributed to each variable at each split
kNN	Area under the ROC curve
Naive Bayes	Area under the ROC curve
Neural network	Absolute values of node weights
Random forest	Difference between prediction accuracy on the out-of-bag portion of the data and after permuting each predictor variable, averaged over all trees and normalized by the standard error
Regression	Absolute value of the t–statistic for each model parameter

Note. Drawn from *caret* package documentation (Kuhn, 2020).

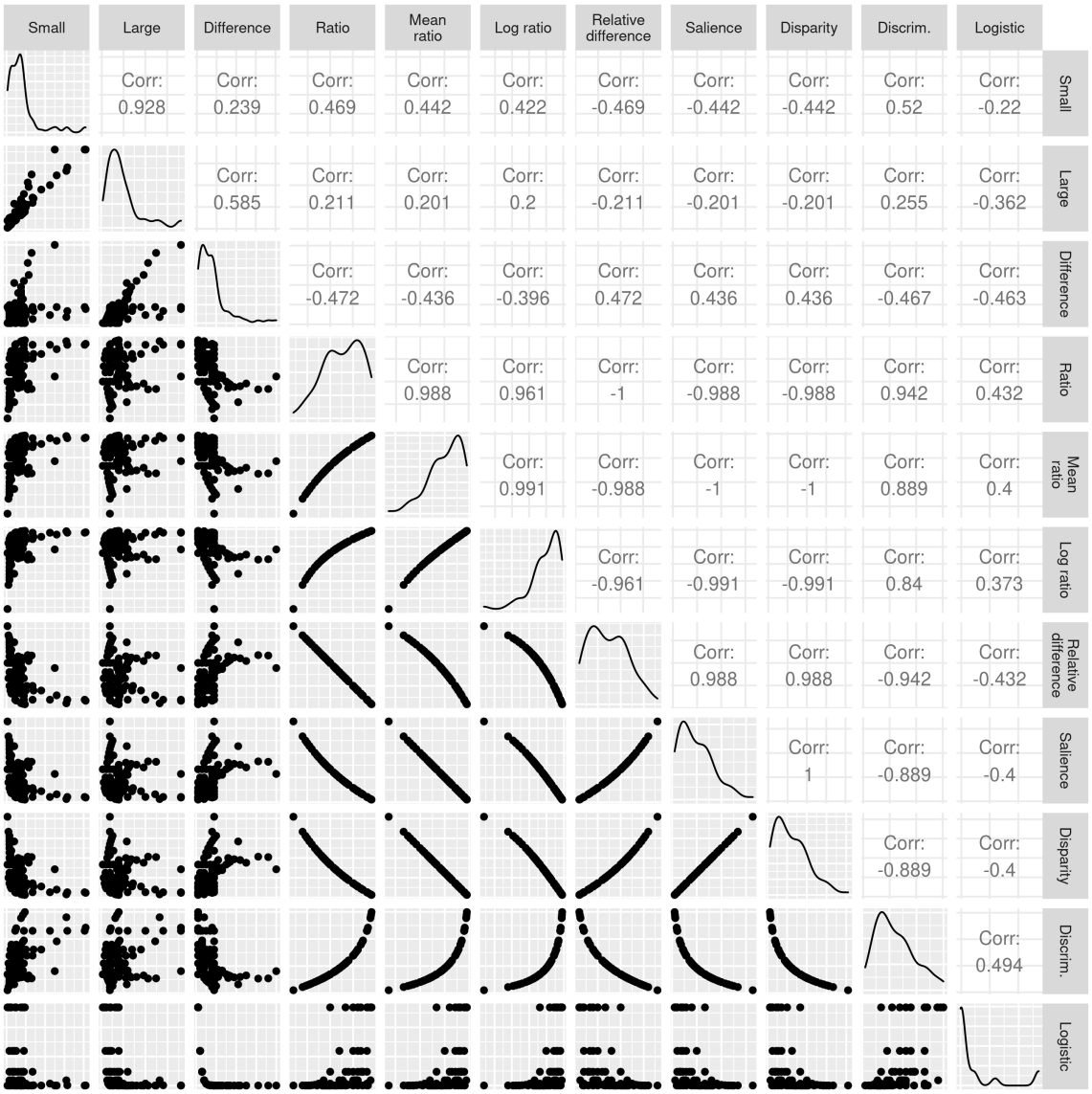

Figure S1. Pairwise correlations for *amount* similarity judgment predictors. Diagonal shows histogram, below diagonal shows correlation plots, above diagonal shows correlation coefficients.

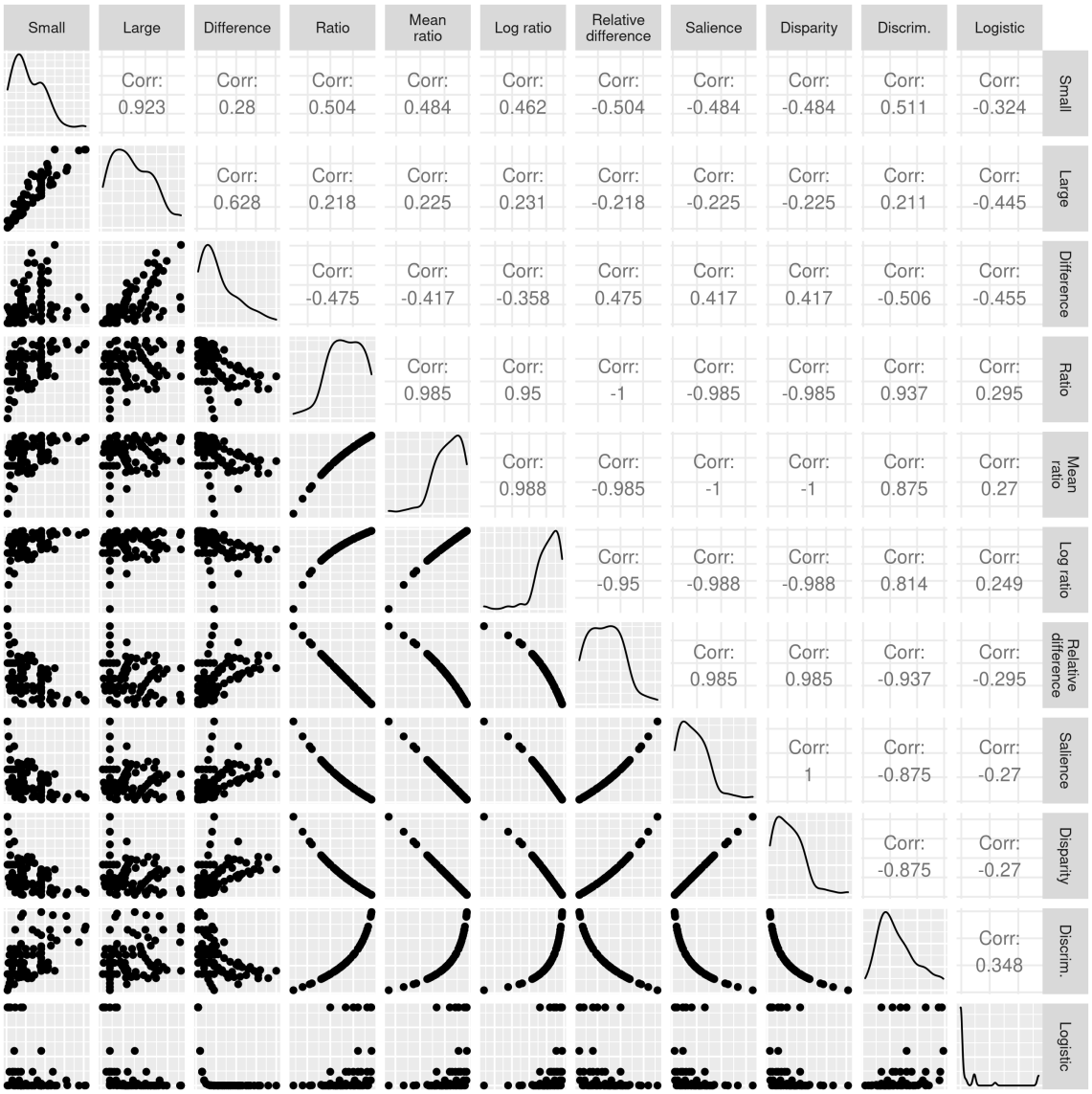

Figure S2. Pairwise correlations for *delay* similarity judgment predictors. Diagonal shows histogram, below diagonal shows correlation plots, above diagonal shows correlation coefficients.

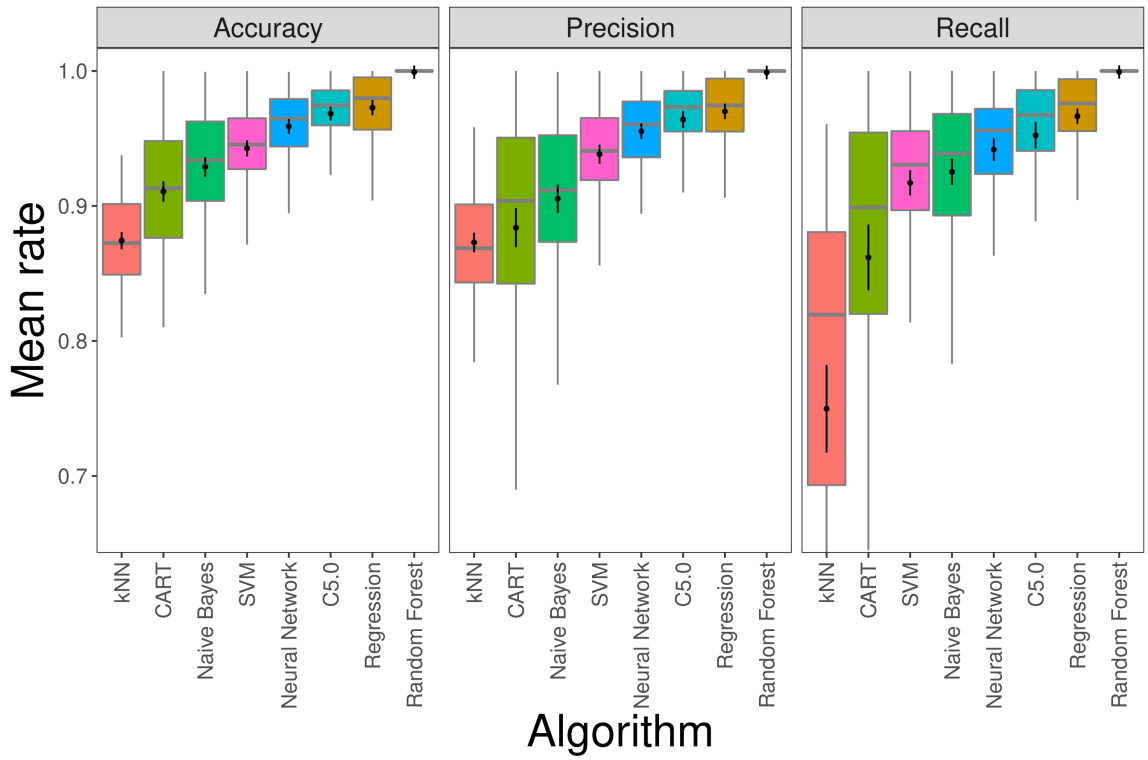

Figure S3. Training set accuracy, precision, and recall for each algorithm based on random ordering of a sample size of 30 instances. For each performance measure, algorithms are ordered by mean score. Dots represent means, error bars represent within-subjects 95% confidence intervals, boxplot horizontal lines represent medians, boxes represent interquartile range (25-75th percentile), whiskers represent 1.5 $\times$ interquartile range. Outliers are not shown. Note the y-axis is truncated at 0.65 to enlarge the presentation of the means and confidence intervals.

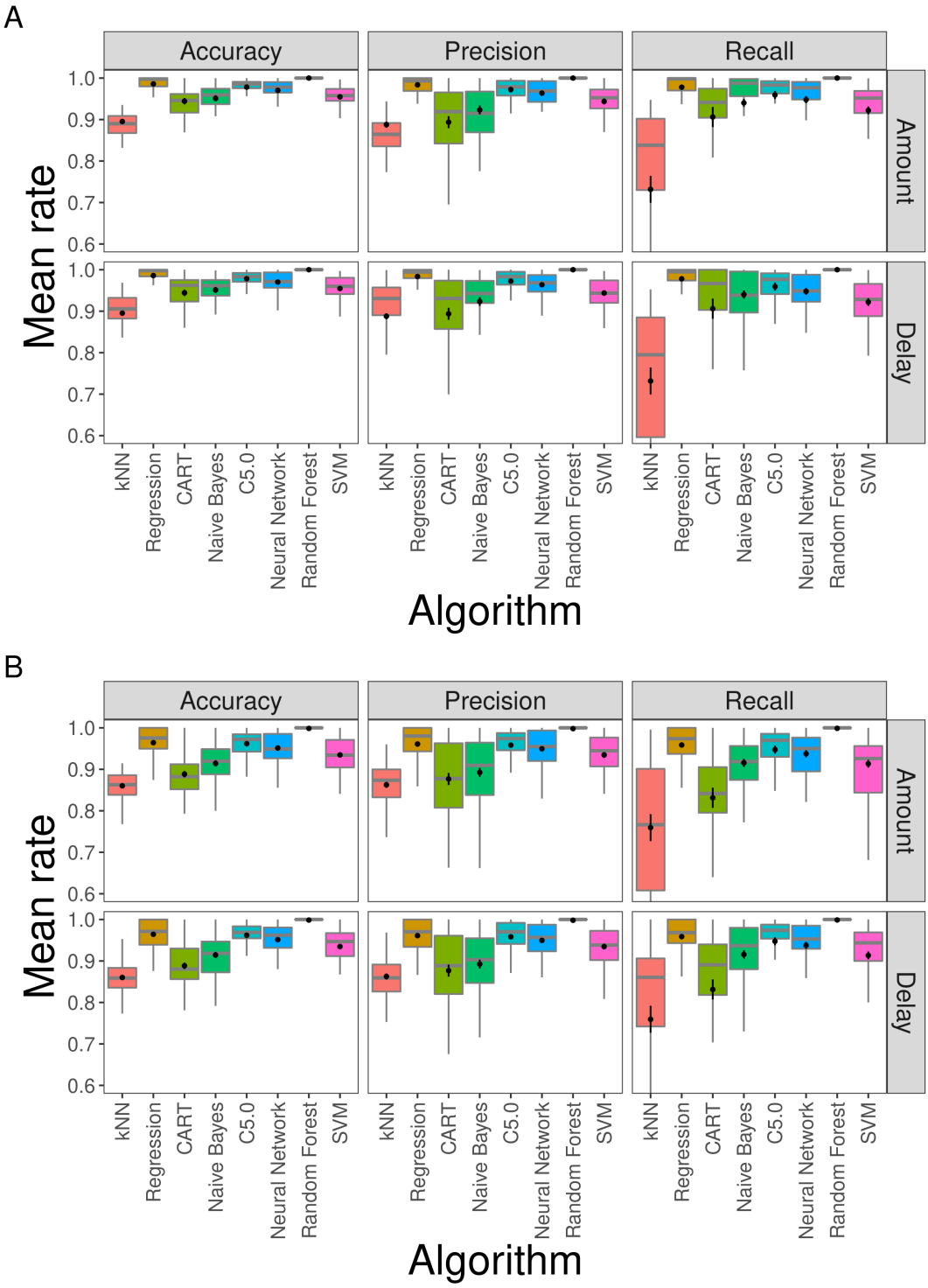

Figure S4. Training set accuracy, precision, and recall rates for each algorithm, judgment type, and data set (A = Data set 1, B = Data set 2). Algorithms are ordered by overall testing accuracy rates. Dots represent means, error bars represent within-subjects 95% confidence intervals, boxplot horizontal lines represent medians, boxes represent interquartile range, whiskers represent 1.5 × interquartile range. Outliers are not shown. Note the y-axis is truncated at 0.6 to enlarge the presentation of the means and confidence intervals.

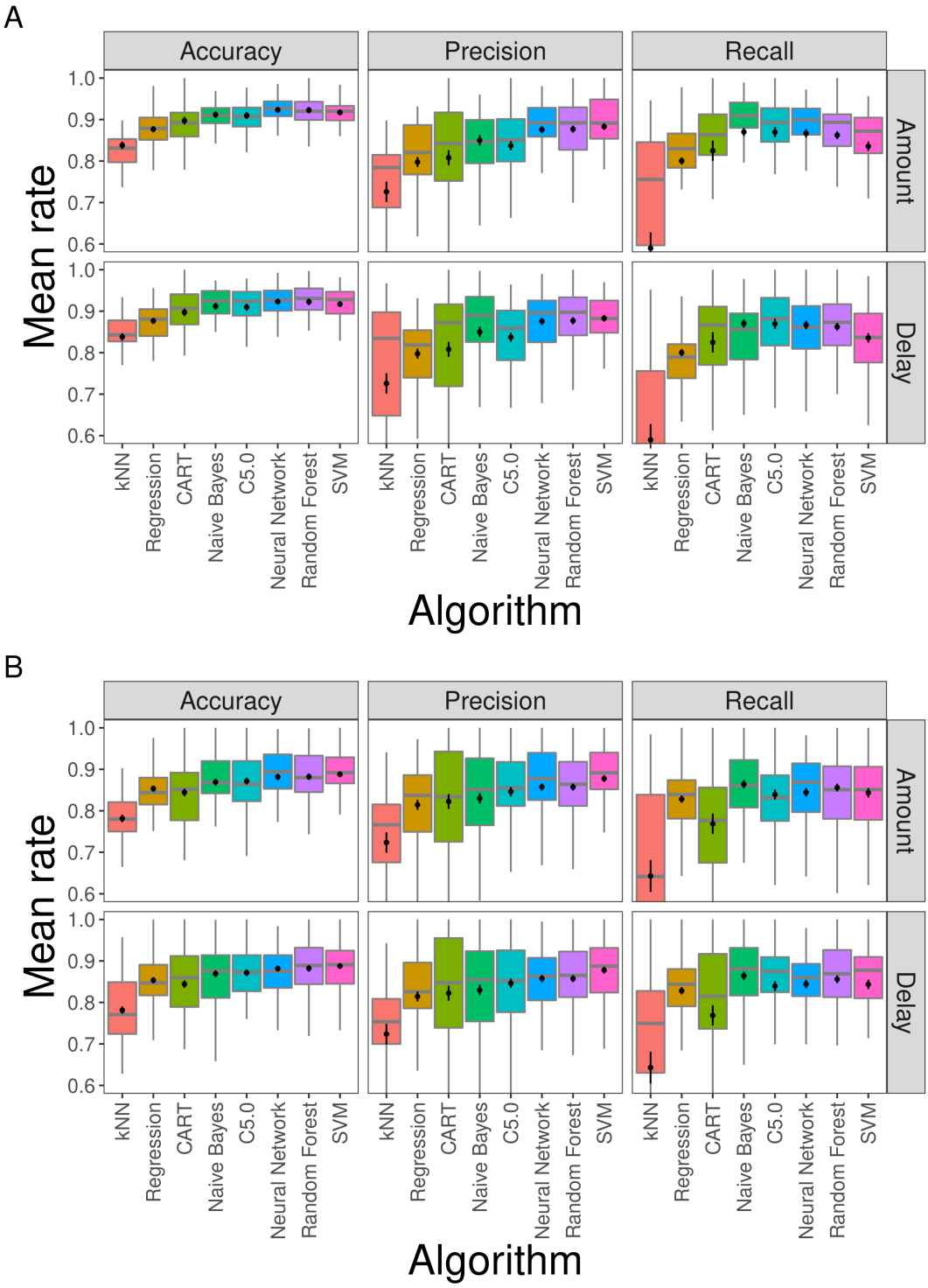

Figure S5. Out-of-sample accuracy, precision, and recall rates for each algorithm, judgment type, and data set (A = Data set 1, B = Data set 2). Algorithms are ordered by overall testing accuracy rates. Dots represent means, error bars represent within-subjects 95% confidence intervals, boxplot horizontal lines represent medians, boxes represent interquartile range, whiskers represent 1.5 × interquartile range. Outliers are not shown. Note the y-axis is truncated at 0.6 to enlarge the presentation of the means and confidence intervals.

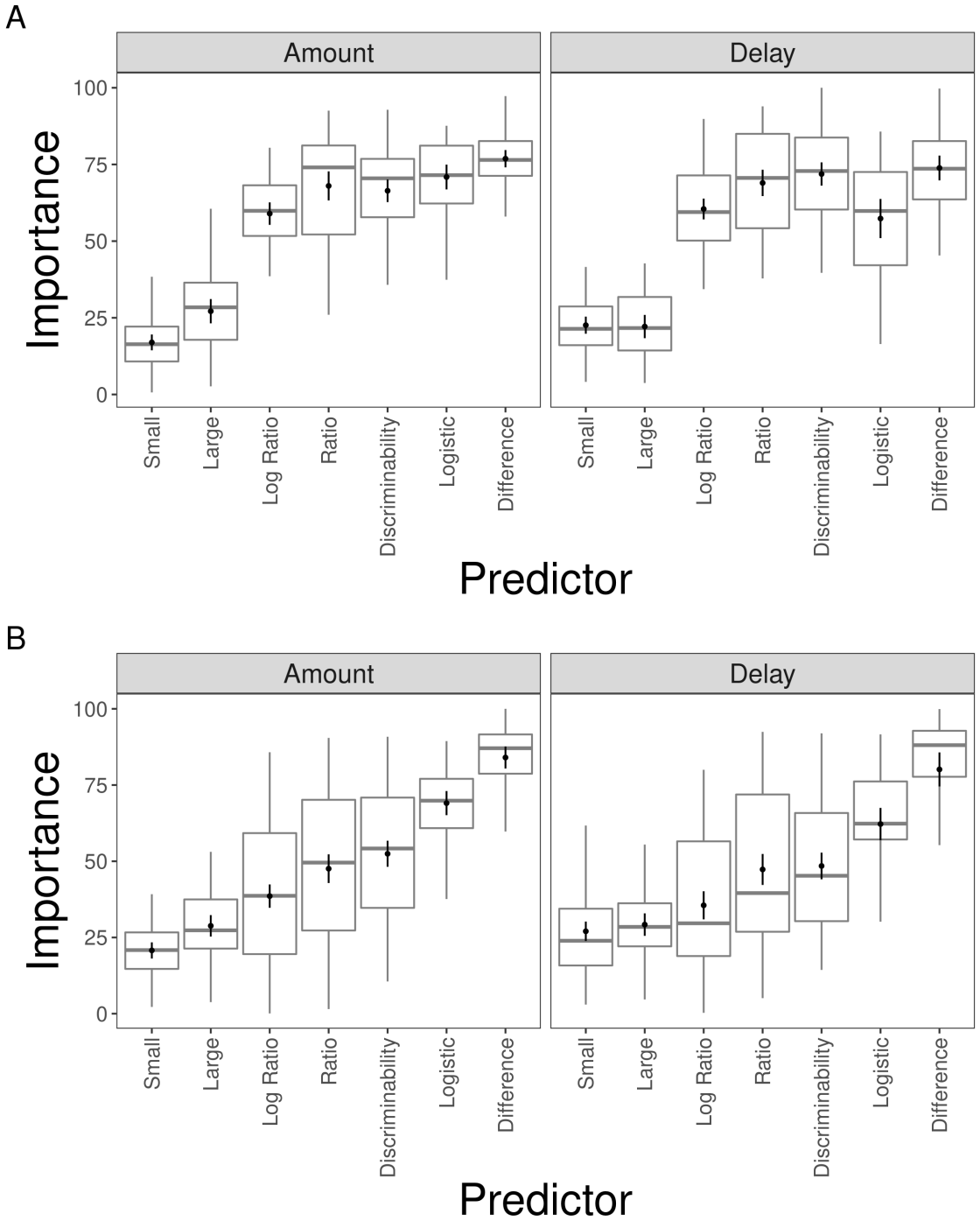

Figure S6. Predictor importance for each judgment type and data set (A = Data set 1, B = Data set 2). Predictor importance refers to the relative contribution of each predictor to the response. Predictors are ordered by overall mean importance. Dots represent means, error bars represent within-subjects 95% confidence intervals, boxplot horizontal lines represent medians, boxes represent interquartile range, whiskers represent 1.5 × interquartile range. Outliers are not shown.

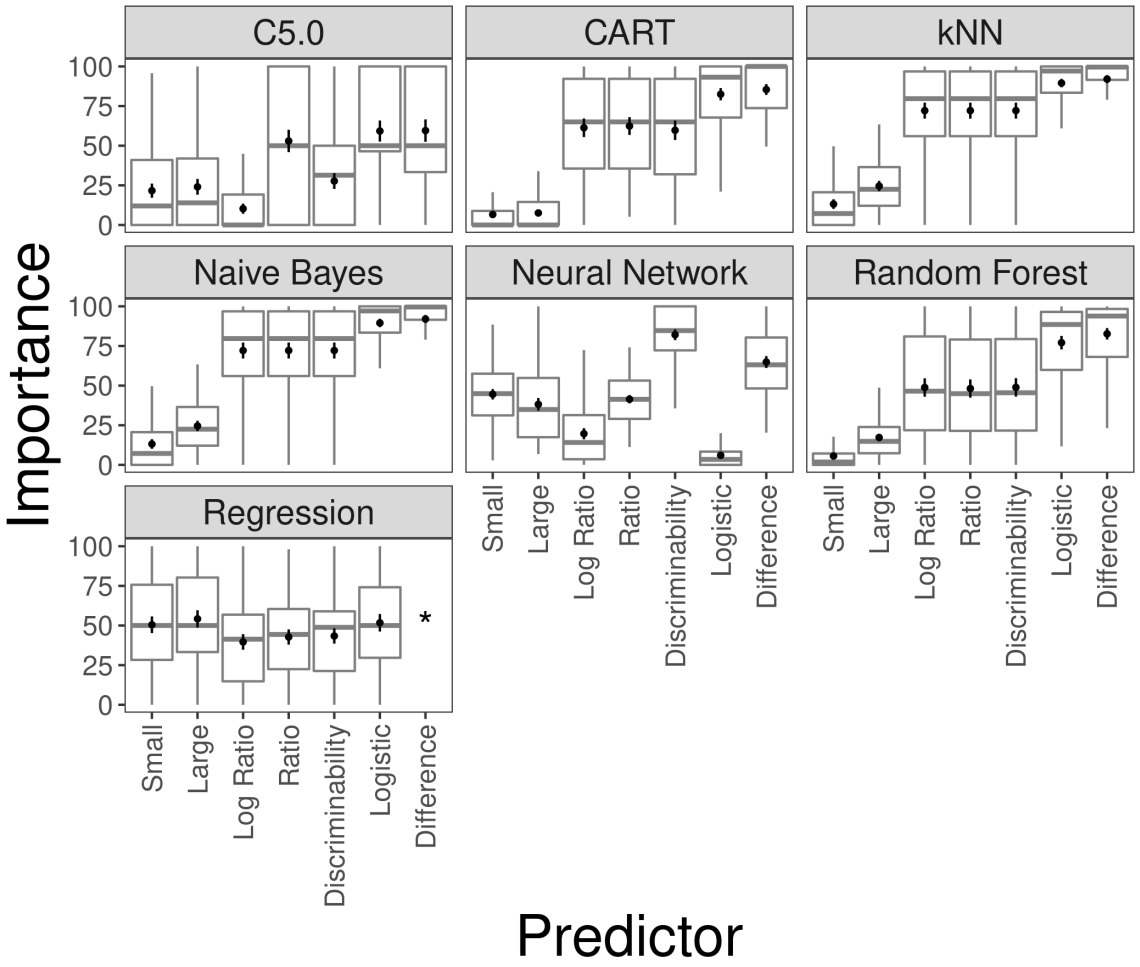

Figure S7. Predictor importance for each judgment type and algorithm. Predictor importance refers to the relative contribution of each predictor to the response. Predictors are ordered by overall mean importance. Dots represent means, error bars represent between-subjects 95% confidence intervals (the failure of regression models to calculate importance for the difference predictor prevents calculation of within-subject confidence intervals), boxplot horizontal lines represent medians, boxes represent interquartile range, whiskers represent 1.5 × interquartile range, and * represents the failure of regression models to calculate importance for the difference predictor. Outliers are not shown.

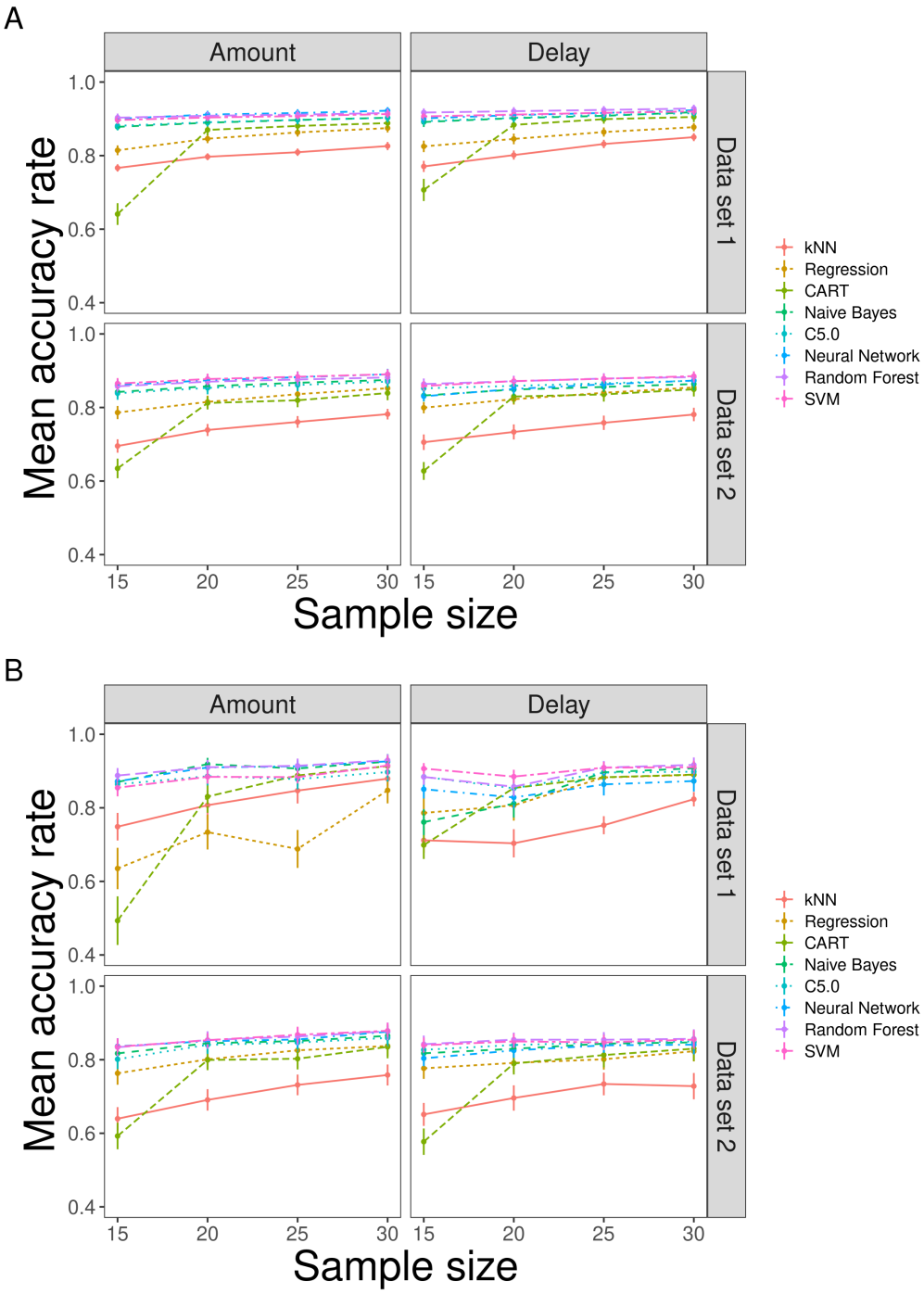

Figure S8. Out-of-sample accuracy for each sample size, judgment type, data set, and ordering (A = Random, B = Sequential). Sample size refers to number of questions per participant used to train the algorithms. Random refers to a random sample of training questions used to predict a random sample of 10 testing questions. Sequential refers to a sample of training questions drawn in order of presentation to each participant that was used to predict a random sample of 10 testing questions. Dots represent means, and error bars represent between-subjects 95% confidence intervals (within-subject confidence intervals were not used because excessive missing data for small sample sizes caused too many participants to be removed from the calculations).